UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

MARINA ALVES AMORIM

# Análise de sensibilidade de técnicas de amostragem em grafos aleatórios

Profa. Dra. Denise Duarte
Orientador

Prof. Dr. Gilvan Ramalho Guedes
Co-orientador

Belo Horizonte, Fevereiro de 2020

# Análise de sensibilidade de técnicas de amostragem em grafos aleatórios

**Marina Alves Amorim**

Dissertação apresentada, ao Programa de Pós-Graduação em Estatística, da Universidade Federal de Minas Gerais, como requisito para obtenção do título de Mestre em Estatística.

Apresentado por:

_____

Marina Alves Amorim

Aprovado por:

_____

Profa. Dra. Denise Duarte

_____

Prof. Dr. Gilvan Ramalho Guedes

_____

Prof. Dr. Luiz Henrique Duczmal

_____

Prof. Dr. Fabricio Murai Ferreira

BELO HORIZONTE, MG - BRASIL

Fevereiro de 2020

# Agradecimentos

Quero agradecer aos meus pais; Emir e Cleunice e toda a minha família e amigos, em especial minha tia Eucy e minha melhor amiga Marina pelo apoio incondicional ao longo desta jornada.

Agradeço aos meus orientadores, Profa. Dra. Denise Duarte e Prof. Dr. Gilvan Ramalho Guedes , por me auxiliarem na minha vida acadêmica, fazendo com que eu possa cooperar com a comunidade científica.

Agradeço também a FAPEMIG - Fundação de Amparo à Pesquisa do Estado de Minas Gerais, por ter me auxiliado nesta caminhada.

# RESUMO

Neste trabalho, propomos uma análise de sensibilidade dos métodos de amostragem para grafos aleatórios, buscamos encontrar a melhor estratégia de amostragem para cada modelo analisado. Quando nos referimos a uma boa estratégia de amostragem, estamos analisando à capacidade de um método de amostragem em preservar as características do grafo populacional observado. Os seguintes modelos de grafos aleatórios foram usados para capturar diferentes estruturas de dados relacionais: Erdös Rényi, Geométrico, Barabasi Albert e Watts Strogatz. Para cada um desses modelos de grafos, testamos os seguintes métodos de amostragem: amostragem aleatória de vértices, amostragem aleatória de arestas e amostragem por bolas de neve.

Amostragem em grafos é uma campo promissor, e existem estudos na área que utilizam medidas topológicas individuais para validar a estratégia de amostragem. Nosso trabalho difere dos demais ao propor o uso de uma informação sintética mais robusta − a densidade espectral do grafo. Além de ser uma medida sintética, ela preserva todas as informações contidas no grafos, incluindo as métricas topológicas usadas individualmente. Utilizamos a divergência de Kullback-Leibler entre a densidade espectral do grafo aleatório e suas versões amostradas para validar seu uso e, em seguida, usando densidades espectrais, construímos um teste a partir das diferenças de Jensen Shannon para verificar se a perda de vértices ou arestas afeta a identificabilidade do modelo original.

Nossa abordagem de amostragem produziu dois resultados principais. Primeiro, encontramos um limiar de 500 vértices para garantir a recuperação do modelo original, independentemente do método de amostragem ou modelo de grafo utilizado. Segundo, nossa abordagem nos permitiu informar qual método de amostragem é mais apropriado para cada modelo de grafo observado.

**Palavras-Chave:** Grafos, Amostragem em grafos, Densidade Espectral.

# ABSTRACT

*In this work, we propose a sensitivity analysis of sampling methods for random graphs in order to find the best sampling strategy for each model analyzed. For best sampling strategy we mean the ability of a sampling method to preserve the characteristics of the graph, even under increasing loss of information. The following random graph models were used to capture different relational data structures: Erdös Rényi, Geometric, Barabasi Albert and Watts Strogatz. For each of these graph models we tested the following sampling methods: random vertex sampling, random edge sampling, and snowball sampling.*

*Sampling graphs is a promising area and there are studies using individual topological measures to validate the sampling strategy. Our work differs from the others in proposing the use of a more robust synthetic information − the spectral density of the graph. In addition to being a synthetic measure, it preserves all the information contained in the graph, including the topological metrics individually used. We use the Kullback-Leibler Divergence between spectral density of the original graph and their sampled versions to validate its use and then, using spectral densities, we built a test from the Jensen Shannon test statistics to check if the loss of vertices or edges affects the identifiability of the original model.*

*Our sampling approach yielded two main results. First, we found a lower limit of 500 vertices to guarantee the recovery of the original model, regardless of the sampling method or graph model used. Second, our approach allowed us to inform which sampling method is most appropriate for each observed graph.*



***Keywords:*** *Graphs, Graph sampling, Spectral density.*

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| TEG | Teoria Espectral dos Grafos |
| G | Notação para um grafo qualquer |
| V | Vértices |
| E | Arestas |
| $A_G$ | Matriz de adjacências |
| $\lambda_i$ | Autovalor i da Matriz de adjacências |
| KLD | Divergência de Kullback Leibler |
| iid | Independente e identicamente distribuídos |
| GIC | *Graph Information Criterion* |
| ER | Erdös Rényi |
| GRG | Geométrico |
| WS | Watts Strogatz |
| BA | Barabasi Albert |
| JS | Jensen Shannon |
| MC | *Monte Carlo* |

# Sumário

# Capítulo 1

# Introdução

Nos últimos anos, as redes sociais online surgiram como uma plataforma de compartilhamento de várias informações sobre pessoas e seus interesses, atividades, eventos e notícias. Devido à grande escala de informações e as limitações de acesso (por exemplo, políticas de privacidade) de serviços de redes sociais online, como *Facebook* e *Twitter*, é difícil acessar o grafo todo em um período de tempo limitado. Por esta razão, os pesquisadores tentam estudar e caracterizar os grafos tomando amostras apropriadas e confiáveis (Rezvanian et al., 2014). Por meio de amostragem dos grafos, podemos processar informações em um tempo razoável e com menor esforço computacional.

A área de amostragem em grafos está em desenvolvimento e com um vasto campo para crescimento. Com isso, ainda existe um número limitado de pesquisas sobre, e sobre como as técnicas de amostragem impactam nas propriedades estruturais dos grafos. Atualmente, existem alguns métodos de amostragem para grafos propostos na literatura, neste trabalho, propomos um estudo de sensibilidade de métodos de amostragem para vários modelos de grafos aleatórios já consolidados na literatura no intuito de encontrar qual seria o melhor método amostragem para cada um dos modelos analisados.

Escolhemos os principais modelos de grafos aleatórios utilizados atualmente para captar as diferentes estruturas de redes relacionais. Aplicamos o teste proposto neste trabalho para alguns métodos de amostragem e em vários modelos de grafos. Os modelos de grafos utilizados são: Erdős Rényi (Erdös and Rényi, 1959), Geométrico (Penrose et al.,

2003), Barabasi Albert (Barabási and Albert, 1999) e Watts Strogatz (Watts and Strogatz, 1998).

Diversos trabalhos recentes propõem e avaliam diferentes processos de amostragem em grafos (Takahashi et al., 2012; Wagner et al., 2017; Leskovec and Faloutsos, 2006; Smith and Moody, 2013). Esses processos buscam entender e identificar propriedades nas estruturas dos grafos. Para cada um dos modelos descritos anteriormente os métodos de amostragem empregados foram: Amostragem aleatória por vértices (Mastrandrea et al., 2015), amostragem aleatória por arestas (Hidalgo and Rodríguez-Sickert, 2008) e Amostragem por Bola de neve (*Snowball Sampling*) (Goodman, 1961).

Existem trabalhos na literatura a respeito de amostragem em grafos, porém, todos utilizam medidas topológicas individuais como a distribuição dos graus, menor distância e o menor diâmetro (Smith and Moody, 2013; Smith et al., 2017; Wagner et al., 2017; Leskovec and Faloutsos, 2006). Nosso trabalho difere dos demais ao propor o uso de uma informação sintética que é mais robusta, essa medida é a densidade espectral do grafo. Além de ser uma medida sintética, ela preserva toda a informação contida no grafo, incluindo as métricas topológicas utilizadas de forma individual nos trabalhos supra citados sobre amostragem em grafos. O grande benefício de utilizar a densidade espectral é que ela capta vários atributos do grafo simultaneamente e não pontualmente como as outras medidas, com isso, ela consegue resumir as principais características do grafo em uma única informação.

Calculamos a diferença Jensen Shannon (JS) para comparar a distribuição das densidades espectrais. Nós construímos um teste a partir das diferenças de Jensen Shannon no intuito de observar o quanto que a perda de vértices ou arestas influencia na identificabilidade do modelo original. O nosso objetivo principal é responder qual técnica de amostragem é mais adequada para cada modelo ou tipo de grafo.

Antecipamos que a nossa proposta é robusta para diferentes tipos de amostragem e com ela conseguimos apontar um limiar de 500 vértices mínimos para garantir a recuperação do modelo original. Além disso, nosso trabalho consegue informar qual o método de amostragem mais adequado para cada modelo observado.

# Capítulo 2

# Conceitos e Definições

Neste capítulo introduzimos alguns conceitos básicos para entender a forma como abordamos e conduzimos este estudo.

## 2.1 Grafos

Um grafo $G = (V, E)$ é definido como um conjunto de objetos de duas naturezas: o conjunto de vértices $(V)$, que é a unidade fundamental sob a qual o grafo é construído e as arestas $(E)$, que representam a ligação entre pares de vértices. A essas ligações e/ou os seus vértices podem ser atribuídas muitas características, como direção, distância e peso.

Os grafos podem ser representados através de matrizes de adjacência ($A_G$), as linhas e colunas representam os vértices e as células representam a ligação entre cada par linha-coluna. Para o caso de grafos não direcionados, a matriz recebe 1 se existe ligação entre dois vértices $i$ e $j$ e 0 caso contrario. Caso o grafo seja ponderado, cada casela da matriz recebe o valor da ligação entre os dois vértices $i$ e $j$.

Para este trabalho, os grafos considerados são simples e não-direcionados. A relação de arestas estará alocada em uma matriz de adjacências que representa se existe ou não a conexão entre vértices.

## 2.2 Modelos de Grafos Aleatórios

### 2.2.1 Grafos Aleatórios

Esta seção apresenta alguns modelos clássicos de grafos aleatórios. Os primeiros modelos foram introduzidos por Erdős e Rényi em 1959 para grafos simples em que cada aresta é incluída no grafo com igual probabilidade (Erdös and Rényi, 1959).

Um grafo aleatório é um grafo que foi gerado por um processo aleatório. Os grafos são estruturas muito usadas para representar a existência ou não de relações entre elementos de um dado conjunto. Eles são usados em muitas aplicações práticas, e em áreas em que as redes complexas precisam ser modeladas. Estas estruturas relacionais podem ser representas por ligações entre os vértices, tais como amizade entre indivíduos, migrações, parceria entre empresas, etc. Existe uma dependência tanto nos vértices quanto entre as ligações (arestas). Outros exemplos de aplicações são: redes de comunicação (como em uma rede de computadores), fluxos em rede de transporte (um mapa de estradas, por exemplo), rotas de distribuição de produtos ou serviços, como dutos de gás ou água e relações binárias que em geral podem ser representadas por grafos (Gersting, 2001). Como podemos perceber pelos exemplos, várias questões de interesse podem ser investigadas utilizando grafos.

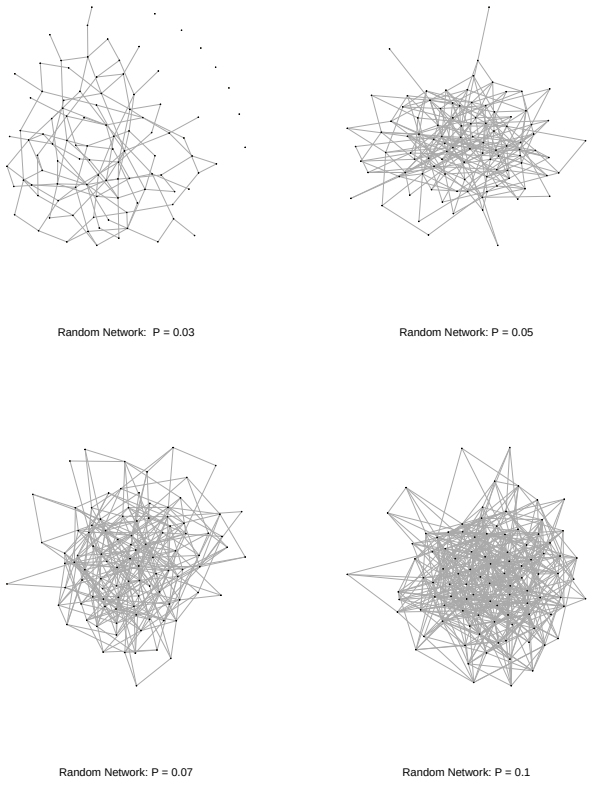


Figura 2.1: Grafos Aleatórios Erdős e Rényi com diferentes probabilidades de conexão e com 100 vértices

#### *2.2.1.1 Modelos*

A seguir, vamos descrever os modelos: Erdős Rényi, Geométrico, Barabási Albert e Watts Strogatz que são os principais modelos de grafos aleatórios utilizados atualmente para captar as diferentes estruturas relacionais.

- **Erdős Rényi (ER):** Os grafos aleatórios Erdős Rényi são os mais simples em termos de construção. Consistem em grafos aleatórios com $n$ vértices, nos quais cada par de vértices $(i, j)$ é conectado por uma aresta com uma dada probabilidade $p$ (Erdös and Rényi, 1959).

  Um grafo aleatório é obtido a partir de um conjunto de $n$ vértices em que adicionamos arestas entre eles de forma aleatória com uma certa probabilidade $p$. Percebemos

que, realizações de um grafo aleatório com $n$ e $p$ fixos darão origem a estruturas diferentes, mas a média do número de ligações destes diferentes grafos será igual a $p \times \frac{n(n-1)}{2}$. Não existe nenhum critério que privilegie uma ligação em relação à outra, e portanto o grafo fica caracterizado pelo número de vértices $n$ e pela probabilidade $p$ de que uma ligação qualquer das $\frac{n(n-1)}{2}$ possíveis ligações entre os diferentes vértices seja estabelecida. Na Figura 2.1 podemos ver algumas realizações de grafos aleatórios Erdős Rényi com 100 vértices para diferentes probabilidades de conexão das arestas ($p$).

Temos dois tipos de modelos classicados como os mais simples, o $G(n,p)$ e $G(n,m)$.

No modelo G(n, m), um grafo é construído de maneira uniforme e aleatória a partir da coleção de todos os grafos possíveis que possuem $n$ vértices e $m$ arestas. Por exemplo, em um modelo G(3, 2), ou seja, 3 vértices e 2 arestas, cada um dos três grafos possíveis com três vértices e duas arestas são incluídos com probabilidade $\frac{1}{3}$.

Já no modelo G(n, p), proposto por Edgar Gilbert em 1959 (Gilbert, 1959), um grafo é construído conectando os vértices aleatoriamente. Cada aresta é incluída no grafo com probabilidade $p$, independente de todas as outras arestas. Equivalentemente, todos os grafos com $n$ vértices e $N$ arestas têm igual probabilidade. Por essa razão, esses modelos, conhecidos como Bernoulli ou Gilbert, são equivalentes ao Erdős Rényi.

$$P^N(1-P)^{(\binom{n}{2}-N)}$$

O parâmetro $p$ neste modelo G(n, p) pode ser considerado uma função de ponderação. À medida que $p$ aumenta de 0 para 1, o modelo torna-se cada vez mais propenso a incluir grafos com mais arestas. Em particular, o caso $p = 0.5$ corresponde ao caso em que todos os $2^{\binom{n}{2}}$ grafos com $n$ vértices são escolhidos com igual probabilidade.

Na prática, o modelo G(n, p) é o mais comumente usado hoje, em parte devido à facilidade de análise permitida pela independência das arestas. Alguns exemplos de aplicação são: formação de estrelas na galáxia, modelos de construção de coalhadas no leite, etc.

- **Geométrico (GE):** É outro tipo de modelo aleatório. É simples em termos de construção, $n$ vértices são desenhados aleatoriamente e uniformemente em um quadrado unitário e um par de vértices é conectado por uma aresta se a distância entre eles for no máximo um dado parâmetro $r$ (Penrose et al., 2003). Os modelos geométricos são geralmente usados em problemas no qual conhecemos a distribuição espacial dos indivíduos, como por exemplo, dutos de gás ou água, pontos de parada de uma rota de ônibus, etc.

- **Barabási Albert (BA):** O modelo Barabási Albert é utilizado para gerar redes aleatórias livres de escala, é baseado na ideia de *preferential attachment*. *Preferential attachment* é ideia de que novos vértices que entram na rede tendem a se relacionar com os vértices mais populares, ou seja, vértices populares tendem a ser cada vez mais populares com o passar do tempo. O modelo proposto por Barabási and Albert (1999), tem uma distribuição de graus como lei de potência, devido ao acoplamento preferencial do vértice. Dizemos que $p(k)$ é o grau do k-ésimo vértice e $p(k)$ segue distribuição de lei de potência com parâmetro $\gamma$ se:

$$p(k) \sim k^{-\gamma}$$

Barabási e Albert propuseram a seguinte construção de uma rede livre de escala: começar com um pequeno número de vértices *(n0)* e, a cada passo de tempo adicionar um novo vértice com $m1(\leq n0)$ arestas que ligam o novo vértice para $m1$ vértices diferentes já presentes no sistema. Ao escolher os vértices aos quais o novo vértice se conecta, assuma que a probabilidade de que um novo vértice seja conectado ao vértice $i$ é proporcional ao grau do vértice $i$, como *preferential attachment*, e ao expoente de escala $p$ que indica a ordem da proporcionalidade ($p = 1$ linear , $p = 2$ quadrático) (Barabási and Albert, 1999).

Os grafos que seguem o modelo de Barabási e Albert geralmente apresentam muitos vértices com graus baixos e poucos vértices com graus muito altos, os chamados *hubs*. Esse modelo possui diversas aplicações em redes reais, como por exemplo páginas na web, onde novas páginas tendem a criar links para páginas mais populares. Outro exemplo é a rede de citações no qual novos artigos tendem a citar artigos mais

populares.

- **Watts Strogatz (WS):** É comumente conhecido como redes de mundo pequeno pois, tem a ideia de que uma pessoa está a apenas algumas conexões de qualquer outra pessoa no universo observado. O modelo WS, proposto por Watts and Strogatz (1998), possui um parâmetro que interpola entre uma rede regular e um grafo aleatório de Erdős Rényi (Takahashi et al., 2012). Primeiro, um grafo anelar é criado com os argumentos fornecidos (tamanho e vizinhança). Em seguida, as arestas deste grafo são religados de forma aleatória e uniforme com probabilidade $p$. Em um primeiro momento temos um grafo com simetria de conexões. Um grafo simétrico não é um modelo apropriado pra ver relações de pessoas, afinal, não temos a mesma quantidade de amigos.

  Um grafo totalmente aleatório também não é ideal pois, a propensão de formar triângulos é maior ao analisar relações entre pessoas, pois, se dois indivíduos conhecem uma mesma pessoa, existe uma probabilidade maior desses dois indivíduos se conhecerem por ter um amigo em comum. É irreal pensar que as conexões são formadas de forma totalmente aleatória. O algoritmo faz uma mistura entre o grafo totalmente simétrico e o grafo aleatório. Cria-se um grafo regular e, para cada vértice, é escolhida uma aresta aleatória e ela é apontada aleatoriamente para outro vértice, ligando o vértice em questão a outro vértice. O parâmetro $p$ controla a variação entre ordem e aleatoriedade, quanto mais próximo de 1, mais aleatório e próximo de um modelo ER. Um exemplo popular para aplicação do modelo é uma rede formada por atores que aparecem em um mesmo filme. O modelo WS não se limita à somente redes de pessoas, outras aplicações variam de redes elétricas a redes neurais.

  A Figura 2.2 apresenta a representação gráfica dos quatro grafos aleatórios apresentados acima (Erdős-Renyi, Geométrico, Barabási Albert e Watts Strogatz). Os grafos representados possuem 500 vértices. Os parâmetros utilizados para gerar cada grafo está descrito na figura.

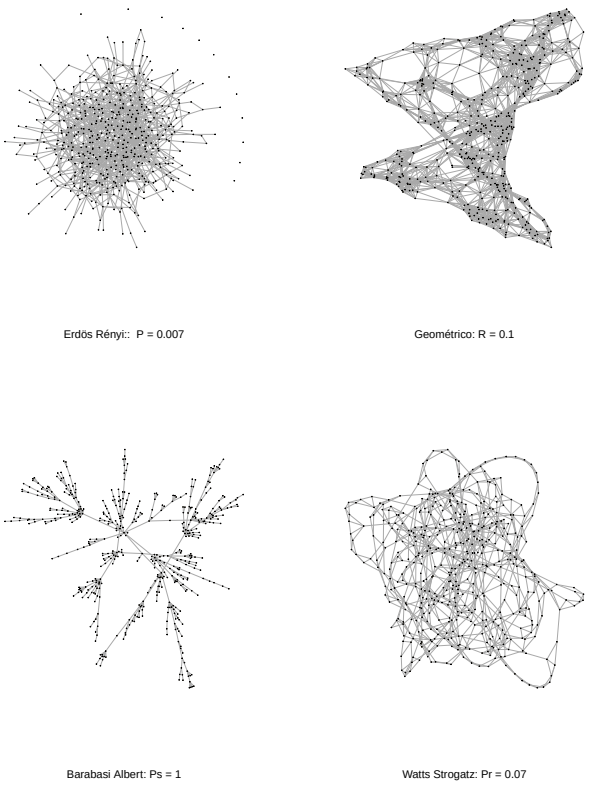


Figura 2.2: Modelos de Grafos Aleatórios com 500 vértices para os modelos : Erdős Rényi, Geométrico, Barabási Albert, Watts Strogatz

## 2.3 Amostragem em Grafos

Amostragem de grafos aleatórios representa um desafio fundamental para a pesquisa em grafos. Para conseguirmos tirar conclusões válidas das amostras, é essencial entendermos como estas refletem a posição dos vértices na rede original, ou seja, a sua topologia (Wagner et al., 2017).

Embora existam alguns estudos que proponham técnicas de amostragem capazes de preservar medidas topológicas isoladas (Smith and Moody, 2013; Smith et al., 2017; Wagner et al., 2017; Leskovec and Faloutsos, 2006), até onde sabemos não há um estudo que teste a sensibilidade de diversas técnicas de amostragem combinadas a uma medida topológica sintética capaz de garantir que se recupere o modelo de grafo original.

Para o nosso estudo, geramos diversos grafos que variam em tamanho de 500 a 3.000 vértices. Esses grafos foram gerados a partir dos modelos descritos na Seção 2.2.1.1.

Para cada grafo gerado, selecionamos aleatoriamente uma proporção de atributos a serem preservados (esses atributos podem ser vértices ou arestas, segundo o critério definido na técnica de amostragem em questão). Posteriormente, reconstruímos o grafo com os atributos que foram selecionados com quaisquer uma de suas arestas ou vértices contribuintes; assim, temos um novo grafo que é um sub conjunto do grafo original. Com o grafo amostral formado, conseguimos calcular qualquer escores e/ou métrica de interesse.

Como no estudo feito por Smith and Moody (2013); Smith et al. (2017), vamos retirar atributos de acordo com o tamanho do grafo, ou seja, vamos retirar do grafo uma proporção de informações (vértices ou arestas). Vamos repetir este processo 200 vezes para cada nível de perda/retirada de informações avaliado o comportamento do grafo amostrado a 1, 10, 20, 50, 80 por cento de perda. Desta forma, teremos ferramentas para medir a sensibilidade das técnicas de amostragem de acordo com a perda de informação. Independente do método de amostragem e do percentual de perda, as informações são perdidas de forma aleatória e aplicadas a todos os modelos de grafos que o estudo contempla.

A seguir, apresentamos as técnicas de amostragem adotadas neste trabalho.

### 2.3.1 Técnicas de Amostragem

- **Amostragem aleatória de Vértices/nós:** Esta é a técnica de amostragem mais básica. Um subconjunto aleatório de $K$ vértices é selecionado do grafo. A amostragem, em seguida, contém esses $K$ vértices e todas as arestas entre eles. A amostragem de vértices aleatórios é usada quando uma amostra de indivíduos é primeiramente selecionada e então o seu comportamento de contato é observado. Numerosas pesquisas e coletas de dados usam esse método, por exemplo, medindo o padrão de contato entre estudantes do ensino médio de uma escola (Mastrandrea et al., 2015).

  Na Figura 2.3 podemos ver um exemplo do processo de amostragem por vértices. Seja um grafo qualquer $G$ com 10 vértices. Para a amostragem, sorteamos aleatoriamente $K$ vértices. Neste exemplo, usaremos 5 vértices como amostra; então, a partir desses 5 vértices sorteados, formamos um subgrafo $G'$ do grafo $G$. No caso da

Figura 2.3 os vértices sorteados foram : $1, 2, 5, 6, 8$. E então, nosso grafo amostrado ou subgrafo $G'$ será formado pelos vértices sorteados e todas as arestas entre eles, assim como estão estruturados em $G$.

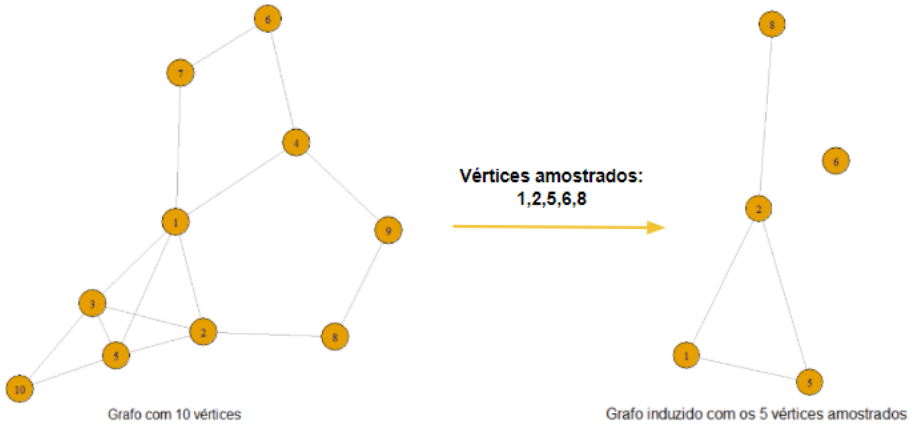


Figura 2.3: Exemplo de grafo induzido a partir de vértices amostrados

- **Amostragem aleatória de arestas:** Nesta técnica de amostragem, sorteiam-se aleatoriamente um número qualquer, pré-definido, de arestas de um grafo $G$. O grafo amostrado, contém $K$ as arestas amostradas e seus vértices, ou seja, os vértices que pertencem às arestas amostradas formam um subconjunto da rede. A amostragem aleatória de arestas é comumente usada para construir um grafo social usando informações sobre contatos, por exemplo, chamadas telefônicas, que são amostradas de um grafo de chamadores e receptores (Hidalgo and Rodríguez-Sickert, 2008).

Na Figura 2.4 podemos ver um exemplo de amostragem por arestas. Seja um grafo qualquer $G$ com 10 vértices. Para a amostragem, sorteamos aleatoriamente $K$ arestas, neste exemplo usaremos 5 arestas, então, a partir dessas 5 arestas sorteadas, formamos um subgrafo $G'$ do grafo $G$. Como mostrado na Figura 2.4 as arestas sorteadas foram: $1, 2, 5, 6, 8$. E então, nosso grafo amostrado ou subgrafo $G'$ será formado pelas arestas sorteadas e todos os vértices entre elas assim como estão estruturados em $G$.

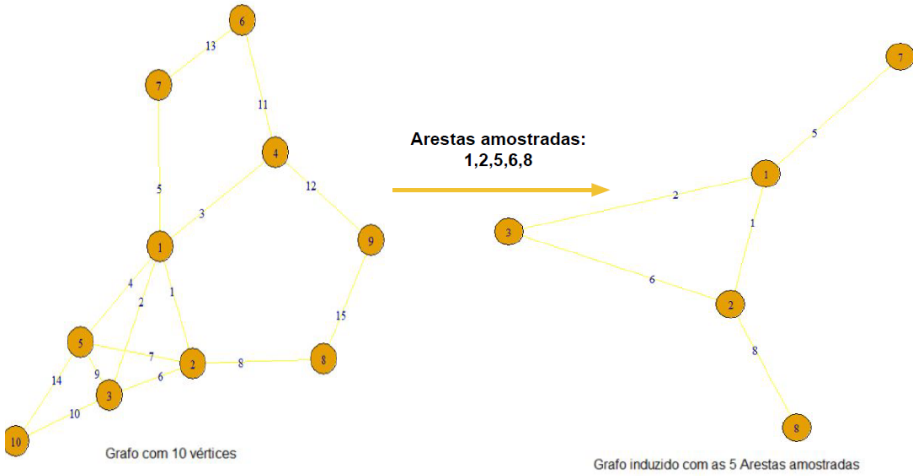


Figura 2.4: Exemplo de grafo induzido a partir de arestas amostradas

- **Amostragem por Bola de Neve ou *Snowball*:** Na amostragem por bola de neve, escolhemos um vértice $u$ como o vértice inicial e adicionamos todos os seus $n$ vizinhos, bem como os vizinhos dos vizinhos. Repetimos isso até termos os $K$ vértices para a amostra. A amostra contém esses $K$ vértices e todas as arestas que os conectam. Tradicionalmente, a amostragem de bola de neve é usada quando a população sob estudo não é facilmente acessível. De fato, a promessa da amostragem por bola de neve é acessar uma população difícil de alcançar (Goodman, 1961).

  O algoritmo funciona da seguinte forma: temos um grafo populacional $G$ e devemos escolher os $K$ vértices que o nosso subgrafo deve conter. Posteriormente, devemos informar o vértice inicial $u$ no qual, o algoritmo deve considerar como ponto de partida. No exemplo da Figura 2.5, o vértice inicial é o vértice *1* e o tamanho do grafo amostrado é de *5* vértices. Então o algoritmo pega os 4 vértices mais próximos do vértice 1 e inclui todas as arestas que estão contidas entre esses vértices de acordo com o grafo populacional $G$.

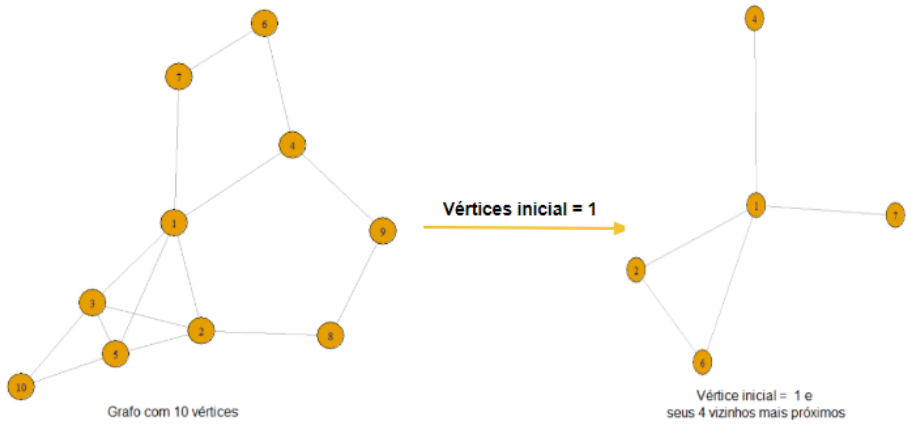


Figura 2.5: Exemplo de grafo induzido a partir da técnica de Bola de neve

Para cada um dos cenários de amostragem apresentados vamos variar o percentual de perda de 1 a 80% das características pretendidas a partir da rede populacional.

Na Tabela 2.1, apresentamos uma lista das técnicas de amostragem e dos percentuais de perda que vamos aplicar em cada um dos modelos de grafos aleatórios.

Tabela 2.1: Métodos de amostragem e os porcentuais de perda que vão ser aplicado para cada modelo

| **Modelo** | **Técnicas de Amostragem Aplicadas** | **Percentuais de perda considerados** |
|---|---|---|
| Erdős Rényi (ER) | Amostragem aleatória por vértices, Amostragem aleatória por arestas e Amostragem por Bola de neve | 1%, 10%, 20%, 50%, 80% |
| Geométrico (GE) | Amostragem aleatória por vértices, Amostragem aleatória por arestas e Amostragem por Bola de neve | 1%, 10%, 20%, 50%, 80% |
| Barabási Albert (BA) | Amostragem aleatória por vértices, Amostragem aleatória por arestas e Amostragem por Bola de neve | 1%, 10%, 20%, 50%, 80% |
| Watts Strogatz (WS) | Amostragem aleatória por vértices, amostragem aleatória por arestas e Amostragem por Bola de neve | 1%, 10%, 20%, 50%, 80% |

## 2.4 Espectro de um Grafo

O espectro de um grafo $G$ é o conjunto de autovalores da matriz de adjacência que o representa. Os autovalores resumem as propriedades essenciais da matriz. São como uma impressão digital do grafo, pois, se dois grafos são gerados de um mesmo processo aleatório, eles têm o mesmo espectro (Dehmer et al., 2017).

A Teoria Espectral dos Grafos (TEG), teve origem na Química Quântica quando, em 1931, Huckel produziu um modelo teórico para um problema a partir de moléculas de hidrocarbonetos não saturadas em que os níveis de energia de certos elétrons eram representados por autovalores de um grafo. Mas, somente em 1957, com um artigo de Collatz e Sigogowitz é que se iniciou a fundamentação teórica da TEG (Abreu, 2005). A TEG estuda propriedades de um grafo por meio de suas representações matriciais e de seus respectivos espectros. Em geral, estudam-se as propriedades estruturais decorrentes das matrizes de adjacências que representam os grafos.

Seja $G = (V,E)$ um grafo não direcionado e $n$ o número de vértices. O espectro de $G$ é o conjunto de autovalores de sua matriz de adjacência, denotado por $A_G$. Temos que $\lambda$ é um autovalor, se existir um vetor $x$ diferente de zero tal que $A\mathbf{x} = \lambda\mathbf{x}$.

Takahashi et al. (2012), mostra que o espectro consegue conservar melhor as características topológicas de um grafo do que os métodos usuais (por exemplo a distribuição dos graus dos vértices). Segundo os autores, o espectro apresenta uma série de propriedades que provam isto. Vamos apresenta-lás a seguir.

Seja $\lambda_1, \lambda_2, ..., \lambda_n$ espectro de $G$ tal que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. A teoria espectral dos grafos estuda as propriedades do espectro e sua associação com a estrutura do grafo. Evidenciamos algumas propriedades (Takahashi et al., 2012) :

1. Seja $d(i)$ o número de arestas conectadas a $i$ (o grau do vértice i). O autovalor $\lambda_i$ é pelo menos:

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

e no máximo $max_{i \in V}$ d(i) (Dorogovtsev and Mendes, 2013)

2. O grafo $G$ é bipartido apenas se $\lambda_n = (-\lambda_1)$ (Dorogovtsev and Mendes, 2013) .

3. Se $G$ é conexo, então o autovalor $\lambda_1$ é estritamente maior que $\lambda_2$ e existe um autovetor positivo de $\lambda_1$ (Mowshowitz and Dehmer, 2012) .

4. Cada vértice em $V$ é conectado a exatamente $\lambda_1$ vértices (ou seja, $G$ é $\lambda_1$-regular) apenas se o vetor de 1s é um autovetor de $\lambda_1$ (Mowshowitz and Dehmer, 2012) .

5. Seja C $\subseteq$ V tal que cada par de vértices em C esteja conectado em $G$ (isto é, C é um clique em G). Então, o tamanho de C é no máximo $\lambda_1 + 1$ (Barabási and Albert, 1999) .

6. Seja $k$ o diâmetro de $G$. Se $G$ é conexo, então $A_G$ tem pelo menos $k+1$ autovalores distintos (Kleinberg, 2000) .

Para mais informações sobre as propriedades do espectro do grafo e aplicações, consultar Abreu (2005) e Cvetkovic et al. (1997).

## 2.5 Densidade Espectral

Dado um conjunto de $n$ vértices rotulados V = 1, 2,. . . , n, e seja $G$ um grafo aleatório. Definimos o espectro de $G$ como um vetor aleatório contendo $n$ variáveis aleatórias $\lambda_1, \lambda_2, ..., \lambda_n$. Seja $\delta$, delta de Dirac, que satisfaz:

(1) $\delta(x) = 0, x \varepsilon \mathbb{R}^*$

(2) $\delta(0) = \infty$

(3) $\int_{-\infty}^{\infty} \delta(x)dx = 1$

A densidade espectral empírica em um grafo $G$ é definida como:

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

É usual tomar o limite da esperança da densidade espectral empírica (denotada por $\langle\cdot\rangle$) de acordo com a lei de probabilidade de G:

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

Nos referimos à densidade espectral empírica definida acima como a densidade espectral de $G$. Para a obtenção empírica da densidade espectral utilizaremos núcleo estimador.

Como mencionado na Seção 2.4, o espectro consegue sintetizar várias propriedades de um grafo. Por este motivo a densidade espectral é tão rica, e ao utiliza-lá estamos levando em consideração várias características do grafo conjuntamente e não apenas um dos atributos possíveis, como é comum na literatura em que os trabalhos utilizam apenas medidas topológicas individuais.

Wagner et al. (2017), por exemplo, utilizam apenas a distribuição dos graus para testar a aplicação das técnicas de amostragem. Já Smith and Moody (2013) e Smith et al. (2017) utilizam medidas de centralidade como: distribuição dos graus, intermediação e proximidade para observar a sensibilidade da amostra quando se introduz a perda de informação. Takahashi et al. (2012), ao amostrar grafos Erdős Rényi, Geométrico, Barabási Albert, Watts Strogatz e K-regular, baseia-se apenas no uso da densidade espectral como ferramenta para a recuperação do modelo gerador, sem considerar perda de informação e o método de amostragem utilizado. Nos trabalhos citados acima, são utilizadas apenas uma medida topológica ou várias medidas individualmente. Usar apenas uma dessas informação pode ser falho ao passo que usar a densidade espectral é uma forma mais robusta pelas propriedades descritas acima, garantindo também a recuperação simultânea das características topológicas do grafo.

## 2.6 Distância de Jensen Shannon

A Distância de Jensen-Shannon é dada entre dois grafos aleatórios $g1$ e $g2$, com o objetivo de determinar uma noção de distância entre eles com base na entropia e divergência de KL. Em outras palavras, estamos interessados em identificar grafos que são gerados

pelo mesmo processo aleatório.

A divergência de KL é adequada para fins de estimativa de parâmetros e seleção de modelos, conforme explicado na seção seguinte. No entanto, não é uma medida simétrica, isto é, em geral $KLD(\rho_1||\rho_2) \neq KLD(\rho_2||\rho_1)$. Por esse motivo, a divergência de KL não é adequada quando não está claro qual é a distribuição de referência. Introduzimos a distância de Jensen-Shannon (JS) entre duas densidades espectrais $\rho_1$ e $\rho_2$ definidas como:

$$JS(\rho_{g1}, \rho_{g2}) = \frac{1}{2}KL(\rho_{g1}|\rho_m) + \frac{1}{2}KL(\rho_{g2}|\rho_m)$$

onde $\rho_m = \frac{1}{2}(\rho_{g1} + \rho_{g2})$.

Essa distância é simétrica e não negativa. Também é zero se e somente se $\rho_1$ e $\rho_2$ forem iguais. Iremos definir um teste estatístico para divergência de JS entre dois conjuntos de espectros de grafos $\rho_1$ e $\rho_2$ como ( $H_0 : JS(\rho_1, \rho_2) \sim 0$ versus $H_1 : JS(\rho_1, \rho_2) > 0$). Detalhes do respectivo teste são apresentados no Capítulo 3.

## 2.7 Entropia

Quando falamos de informação de uma variável aleatória, temos um conceito amplo demais para ser capturado completamente por uma única definição. No entanto, para qualquer distribuição de probabilidade, definimos uma quantidade chamada entropia, que tem muitas propriedades que concordam com a noção intuitiva do que uma medida de informação deve ser (Cover and Thomas, 2012).

Primeiro introduzimos o conceito de entropia, que representa uma medida da incerteza de uma variável aleatória. Seja $X$ uma variável aleatória discreta, com função de probabilidade $p(x) = P(X = x), x \in X$.

A entropia $H(X)$ de uma variável aleatória discreta $X$ é definida por:

$$H(X) = -\sum_{x \in X} p(x) log(p(x))$$

Generalizando para o caso discreto e contínuo, podemos escrever a entropia como a

esperança de uma variável aleatória:

$$H(X) = E_p log(\frac{1}{p(x)})$$

No qual $X$ segue distribuição $p(x)$. Neste caso, para achar a entropia basta calcular a esperança da variável aleatória.

O log é na base 2 e a entropia é expressa em bits. Usaremos a convenção de que $0log0 = 0$, o que é facilmente justificado pela continuidade, uma vez que $xlogx \to 0$ quando $x \to 0$. A adição de termos de probabilidade zero não altera a entropia (Cover and Thomas, 2012).

Note que a entropia é uma função da distribuição de $X$. Não depende dos valores reais tomados pela variável aleatória $X$, apenas das probabilidades.

## 2.8 Entropia Relativa

A entropia é estendida para definir informação mútua, que é uma medida da quantidade de informação que uma variável aleatória contém sobre outra. A informação mútua, é uma medida da distância entre duas distribuições de probabilidade. Todas estas quantidades estão intimamente relacionadas e compartilham um número de propriedades simples, essas quantidades de informações surgem como respostas naturais a várias questões de comunicação, estatística, complexidade e jogos de azar (Cover and Thomas, 2012).

A entropia relativa é uma medida da distância, ou disparidade, entre duas distribuições (Cover and Thomas, 2012). A medida de entropia apresentada na Subseção 2.7 pode ser generalizada pelo conceito de entropia relativa ou divergência de Kullback Leibler entre duas funções de massa de probabilidade $p(x)$ e $q(x)$.

$$KLD(p||q) = \sum_{x \in X} p(x) log(\frac{p(x)}{q(x)})$$

Na definição acima, usamos novamente a convenção de que que $0log\frac{0}{q} = 0$ e $plog\frac{p}{0} = \infty$. Portanto, para qualquer $x \in X$ tal que $p(x) > 0$ e $q(x) = 0$, $KLD(p||q) = \infty$.

A entropia relativa é sempre não negativa e é zero se e somente se $p = q$. No entanto, não é uma distância verdadeira entre distribuições, uma vez que não é simétrica, muitas vezes é útil pensar em entropia relativa como uma distância entre distribuições.

Utilizaremos a KLD como um método de validação para provar que a densidade espectral realmente consegue conservar as características topológicas do grafo e caracterizar os modelos, assim como mostrado por Takahashi et al. (2012), que utiliza a KLD com fins de classificação dos modelos. Isto é, vamos calcular a divergência de Kullback Leibler entre um grafo amostrado e um modelo aleatório de grafos utilizando a densidade espectral como parâmetro, como apresentado na Seção 2.9, e a partir da KLD vamos classificar o modelo gerador do grafo amostrado.

## 2.9 Validação da Densidade Espectral como medida sintética da estrutura topológica do grafo

Takahashi et al. (2012) apresentam uma metodologia que permite, a partir da densidade espectral de um grafo amostrado, conseguir recuperar o modelo gerador utilizando a divergência de Kullback-Leibler. Vamos apresentar e descrever algumas ferramentas utilizadas por ele e como elas vão ser aplicadas neste trabalho.

Vamos utilizar o trabalho de Takahashi et al. (2012) para determinar qual é o tamanho mínimo de vértices que um grafo deve conter para que o mesmo seja considerado como uma população do modelo gerado, ou seja, quantos vértices são necessários para que o grafo convirja para a população e possamos assim fazer a amostragem.

Seja $G$ um grafo aleatório e $\rho$ sua densidade espectral. A entropia espectral de $G$ é definida como:

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

onde 0 log 0 = 0.

Propomos que a entropia espectral descreve características importantes do grafo. Mais

especificamente, propomos que a entropia espectral mede a incerteza associada ao grafo aleatório.

Podemos calcular a entropia espectral aproximada para um grafo aleatório Erdos-Renyi $G$ com o parâmetro $p$ da seguinte maneira. Para $n$ grande, temos:

$$\rho_g(\lambda) = \frac{\sqrt{4p(1-p)-\lambda^2}}{2\pi p(1-p)}$$

No caso específico do grafo aleatório de Gilbert (1959), a densidade espectral pode ser aproximada por:

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

onde $p$ é a probabilidade de conectar um par de vértices. Então, a entropia espectral máxima do grafo aleatório de ER é alcançada quando $p = 0,50$. Isso é consistente com a ideia intuitiva de que quando todos os resultados possíveis têm a mesma probabilidade de ocorrer, a capacidade de prever o sistema é fraca. Por contraste, quando $p \to 0$ ou $p \to 1$, a construção do grafo torna-se determinística, e a quantidade de incerteza associada à estrutura do grafo alcança seu valor mínimo.

A divergência de Kullback-Leibler (KLD) mede a quantidade de informação perdida quando uma distribuição de probabilidade é usada para se aproximar de outra distribuição.

Claramente, se duas densidades espectrais são diferentes, então os grafos aleatórios correspondentes são diferentes. No entanto, diferentes grafos aleatórios podem ter a mesma densidade espectral. É extremamente importante identificar quando dois grafos advém de processos diferentes, pois o modo como amostramos o grafo é influenciado pelo modelo gerador.

Sejam dois grafos aleatórios com densidades espectrais $\rho_1$ e $\rho_2$, respectivamente. A divergência de Kullback-Leiber é definida da seguinte forma. Se o suporte de $\rho_2$ contiver o suporte de $\rho_1$, então a divergência entre $\rho_1$ e $\rho_2$ é:

$$KL(\rho_1 \mid \rho_2) = \int_{-\infty}^{\infty} \rho_1(\lambda) log \frac{\rho_1(\lambda)}{\rho_2(\lambda)} d\lambda$$

onde 0 log 0 = 0 e $\rho_2$ é a medida de referência. Se o suporte de $\rho_2$ não contiver o suporte de $\rho_1$, então $KL(\rho_1 \mid \rho_2) = +\infty$.

A divergência de KL é não-negativa, e é zero apenas se $\rho_1$ e $\rho_2$ forem iguais. Note que em muitos casos, $KL(\rho_1 \mid \rho_2)$ e $KL(\rho_2 \mid \rho_1)$ são diferentes quando $\rho_1 \neq \rho_2$, isto é, a medida de KLD é assimétrica. A propriedade assimétrica da divergência de KL é particularmente útil quando queremos encontrar a medida de referência que melhor descreve o espectro observado.

Nosso trabalho propõe o uso da divergência de Kullback-Leibler como um instrumento para validar o uso da densidade espectral, ou seja, utilizaremos a KLD entre duas densidades espectrais apenas para comprovar que conseguimos recuperar o modelo do grafo.

Usaremos as definições apresentadas por Takahashi et al. (2012) para definir um tamanho de $N$ vértices, tal que o nosso grafo seja próximo o suficiente do modelo gerador para ser considerado uma população, e assim possamos fazer a amostragem. Na Seção 4, vamos apresentar um estudo de simulação que mostra que, com $N$ a partir de 500 a diferença entre o grafo gerado e o modelo é aproximadamente zero, o que nos leva a acreditar que para os modelos testados, com $N \geq 500$ podemos considerar o grafo como sendo uma população, vamos considerar este número como um resultado assintótico.

# Capítulo 3

# Metodologia

## 3.1 Estimação das densidades espectrais

Existem várias formas de ajuste não paramétricos para encontrar a distribuição de uma amostra, dentre elas se destacam os estimadores de Kernel, que será o método utilizado neste trabalho.

Para estimar a densidade espectral vamos utilizar núcleo estimador. Este é um método não paramétrico que resulta em uma curva suavizada da função densidade de probabilidade que segue o comportamento da amostra, no nosso caso, autovalores, já que não conhecemos a distribuição real dos autovalores do modelo.

Formalmente, o estimador da função densidade em um ponto $x$ é:

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

sendo $h$ o parâmetro de suavização, no qual, a escolha de $h$ afeta a forma da densidade estimada, pois, $h$ determina a largura de vizinhos que o algoritmo irá considerar, $x_i$ $i = 1, 2, 3, ..., n$ é a amostra dos dados e $K$ é uma função kernel escolhida. Neste trabalho iremos usar a janela ótima do kernel (Sheather and Jones, 1991).

A fim de conservar as propriedades de uma função de densidade, a função kernel deve satisfazer a seguinte condição:

$$\int_{-\infty}^{\infty} K(u)du = 1$$

Embora a escolha do kernel afete diretamente a estimativa da densidade, a literatura sugere que esse efeito é bastante pequeno, com resultados empíricos muito semelhantes para diferentes escolhas de $K$. Segundo Scott (2015), a escolha da função kernel tem um papel menor na qualidade final da estimativa. Como o kernel não impacta muito na estimativa da densidade, optamos por utilizar o kernel gaussiano, que é muito utilizado na literatura.

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

Desta forma, vamos utilizar o núcleo estimador para estimar a densidade empírica dos autovalores.

## 3.2 Densidade Espectral para caracterização dos modelos utilizados

Em nosso estudo, constatamos que a densidade espectral segue um padrão de acordo com cada modelo de grafo aleatório, ou seja, a forma da densidade de cada modelo pode ser usada para caracterizar o mesmo. Veremos a seguir as representações gráficas da densidade para cada um dos modelos descritos na Seção 2.2.1.1.

Primeiramente, geramos grafos com 300, 500, 1.000 e 1.500 vértices. Posteriormente, calculamos a densidade espectral dos mesmos e observamos as características de cada densidade. A partir de uma análise visual, percebemos que a mesma apresenta um padrão de acordo com o modelo gerador do grafo. Este é mais um indício de que a densidade espectral é adequada para caracterizar o grafo, além das propriedades descritas na Seção 2.4.

Na Figura 3.1 observamos a representação gráfica das densidades espectrais do modelo Erdős Rényi. Tanto para grafos menores, com 300 vértices, como para grafos maiores com 1.500 vértices, observamos um padrão na estrutura da densidade, ou seja, o modelo Erdős Rényi sempre apresenta uma curva assimétrica a esquerda com uma cauda longa a

direita e poucas observações extremas. Observamos também que a medida que o número de vértices cresce a variância também cresce.

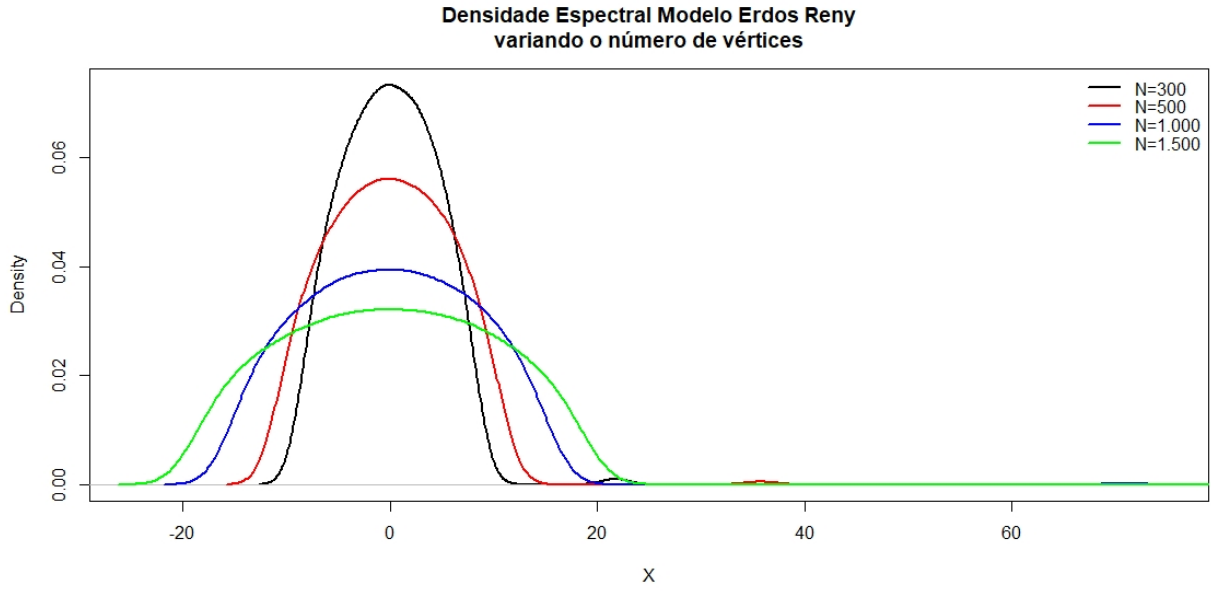


Figura 3.1: Densidade Espectral para o Modelo Erdős Rényi com probabilidade 0.07 e o número de vértices variando (300 , 500, 1.000 e 1.500)

Percebemos que a densidade estimada coincide com expressão analítica aproximada que é apresentada na Seção 2.9 para o caso do modelo ER, ou seja, que converge para a lei do semicírculo.

Para o modelo Geométrico, como mostra a Figura 3.2, o padrão é consistente para todos os tamanhos de grafos testados, em todos percebemos uma curva assimétrica a direita com o desvio padrão também muito parecido independente da quantidade de vértices, e uma cauda longa a direita com poucas observações.

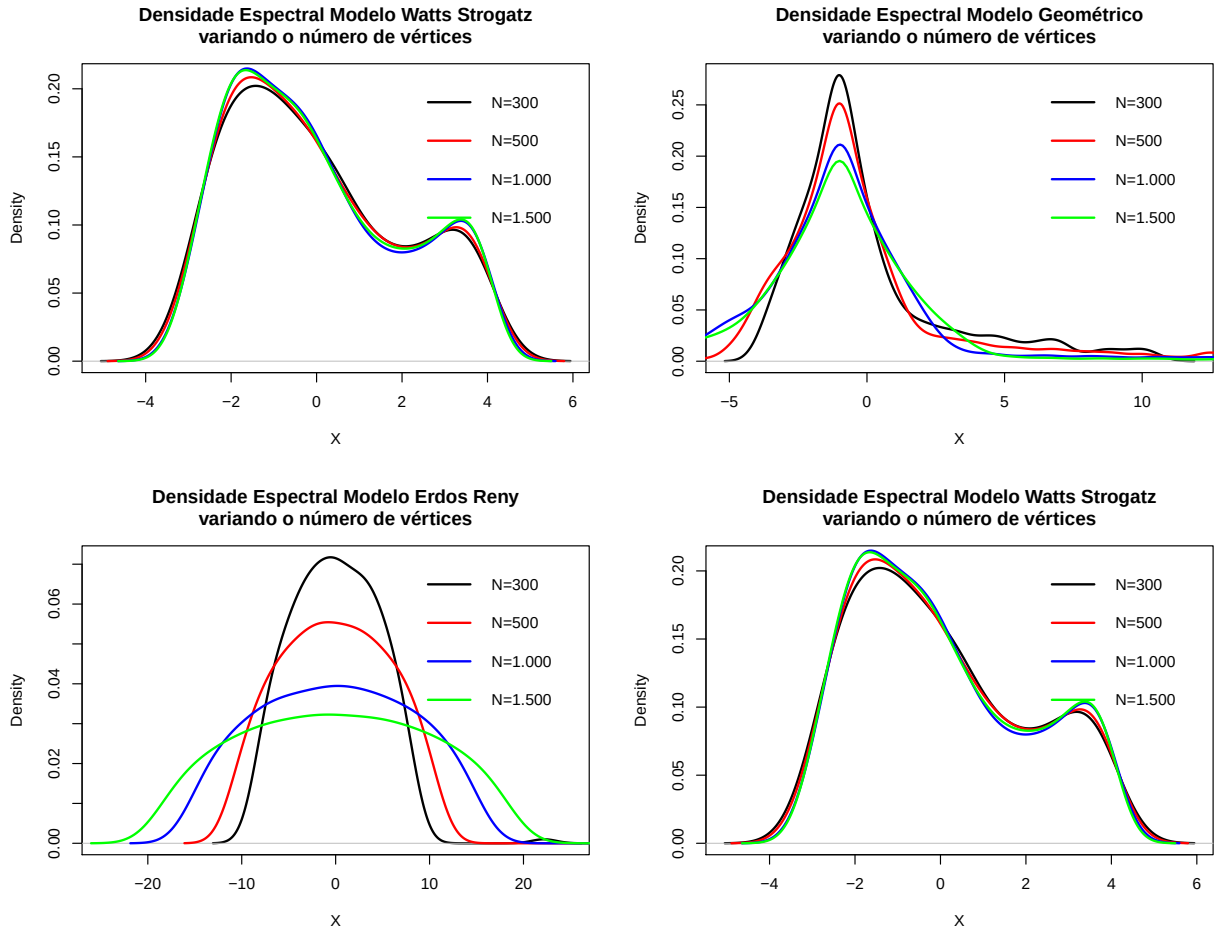


Figura 3.2: Densidade Espectral para os Modelos observados e o número de vértices variando (300 , 500, 1.000 e 1.500)

Analisando agora a densidade do modelo Barabasi Albert, também na Figura 3.2, observamos uma curva quase simétrica, com uma cauda longa e com poucas observações. A curva tem a característica de ter um desvio padrão maior a medida que as observações se afastam da média. Independente do número de vértices do modelo a densidade é bem similar.

Na densidade do modelo Watts Strogatz, representado também na Figura 3.2, observamos uma curva bimodal com grande desvio padrão. E como nos resultados anteriores, temos coerência entre as densidades, independentemente do número de vértices do grafo.

Por fim, podemos observar que qualquer que seja o modelo, o número de vértices não altera muito a forma da densidade, e sim sua variância e curtose, isso é um forte indício de

que a densidade é uma boa ferramenta para caracterizar o modelo. Na Seção 4, provamos a partir de simulações que a densidade espectral de fato consegue caracterizar e recuperar o modelo do grafo.

## 3.3 Procedimento de Análise de sensibilidade das técnicas amostrais utilizadas

Vamos gerar vários grafos com tamanho a partir de 500 vértices pois, temos evidências de que em grafos com mais de 500 vértices conseguimos recuperar a densidade do modelo original. Este resultado será apresentado na Seção 4. Para avaliar o comportamento da amostra, geramos vários grafos aleatórios com diferentes tamanhos e parâmetros.

Foram construídos grafos com vértices variando de 500 a 3.000 vértices (500, 1.000, 3.000), de modelos como: Erdős Rényi, Geométrico, Barabasi Albert e Watts Strogatz. Utilizamos o ambiente de programação do *R Core* Team (2015) como ferramenta para gerar e tratar os dados. Para facilitar a manipulação de medidas e modelos de grafos, contamos também com as ferramentas desenvolvidas pelo pacote *Igraph* (Csardi et al., 2006) e *StathGraph* (Santos et al., 2019).

Para que o texto não se torne muito extenso, vamos apresentar os resultados obtidos via simulação para os grafos de 3.000 vértices. Para encontrar os testes feitos com os grafos com $N = 500$ e $N = 1.000$ e com os parâmetros variando de acordo com cada modelo, consultar Apêndice A.

O procedimento realizado para a obtenção dos resultados foi gerar uma amostra grande o suficiente para que um grafo seja considerado uma população, com isso, vamos aplicar neste grafo a diversos níveis de perda, como descrito na Seção 2.3 e aplicar o método de amostragem.

Esse processo foi realizado para todos os modelos mencionados na Seção 2.2.1.1. Afim de exemplificar e descrever o processo: seja um grafo $G \sim ER(n, p)$, um grafo gerado de um modelo Erdős Rényi com $n$ vértices e probabilidade $p$ de conexão entre dois vértices, no nosso exemplo geramos um grafo com $n = 3.000$ e $p = 0.07$. Construímos um grafo com

essas especificações e em seguida, vamos aplicar uma perda de informação de 50% nos vértices (isto para amostragem por vértices), ou seja, o grafo que antes tinha 3.000 vértices agora vai ter apenas 1.500 vértices ( 50% de 3.000) e temos agora um grafo $G'$ que será um subgrafo de $G$ (amostra). Logo após aplicar a perda de informação, vamos comparar se a densidade espectral do grafo populacional $G$ ($G \sim ER(n = 3.000, p = 0.07)$) e a densidade espectral do grafo amostrado $G'$ vem de uma mesma distribuição de probabilidade. Isto é possível utilizando a distância de Jensen Shannon para comparar as densidades espectrais dos dois grafos, como descrito na Seção 2.6.

Com a distância de Jensen Shannon, vamos construir uma ferramenta para demonstrar se a porcentagem de perda no qual o grafo foi exposto foi suficiente para desconfigurar as características do grafo a ponto de não conseguirmos identificar uma semelhança estatisticamente significativa entre a densidade espectral do grafo populacional $G$ com o grafo amostrado $G'$. De modo contrário, se mesmo com a perda, a densidade espectral do grafo amostrado e do populacional seguem a mesma distribuição de probabilidade. Com essa informação podemos construir um teste e indicar um limiar, mostrando a partir de qual nível de perda o modelo começa a perder suas características estruturais e identificáveis.

Nosso teste terá como hipótese nula que $H_0 : JS(\rho_1, \rho_2) \sim 0$ versus $H_1 : JS(\rho_1, \rho_2) > 0$ já que a distância de JS é zero se e somente se $\rho_1$ e $\rho_2$ forem iguais. E a nossa estatística de teste será a própria distância de Jensen Shannon.

Para comparar se determinado nível de perda está influenciando no valor da distância de Jensen Shannon, vamos utilizar 200 réplicas para cada nível de perda, que varia de 1 a 80%. Para cada nível de perda, guardamos o valor encontrado na distância de JS e posteriormente registramos a distribuição destas distâncias. Também registramos a distribuição das distâncias sob $H_0$ (de que as duas distribuições são iguais), isto para cada um dos modelos. Deste modo temos duas densidades, uma é a densidade das distâncias de JS sob $H_0$, e a outra é a densidade das distâncias de JS entre a densidade espectral do grafo populacional e o grafo amostrado.

Com isso, podemos comparar a distribuição da distância sob $H_0$ com a distribuição da distância encontrada utilizando a amostra do grafo populacional para cada nível de perda. Optamos por apresentar o teste desta forma pois, consideramos mais informativo do

que informar um único valor, e também porque, se informarmos apenas um valor de teste não teríamos como mensurar o erro tipo II do teste, ou seja, não temos como mensurar se estamos perdendo poder no teste. A seguir, vamos explicar como foi feito esta analogia.

Para exemplificar como será feita a nossa interpretação do teste de hipótese que construímos, vamos utilizar como exemplo os testes do modelo Geométrico com 3.000 vértices e raio de conexão de 0.1. Na figura a seguir, iremos observar a distribuição das estatísticas de teste do modelo GE para os diversos níveis de perda utilizando amostragem por vértices.

Na Figura 3.3, a distância de JS foi realizada utilizando a distribuição espectral do grafo populacional $G = GE \sim (n = 3.000, r = 0.1)$ e os grafos amostrados $G'_1 = GE \sim (n = 600, p = 0.1)$ referente a 80% de perda e $G'_2 = GE \sim (n = 1.500, p = 0.e)$ referente a 50% de perda e assim por diante para os outros níveis de perda. Registramos a distribuição das distâncias de JS sob $H_0$ (linha preta) e a distribuição das distâncias de JS para o grafo com as perdas, os grafos amostrados (linha vermelha). A linha azul é referente ao percentil 95% sob a distribuição de $H_0$.

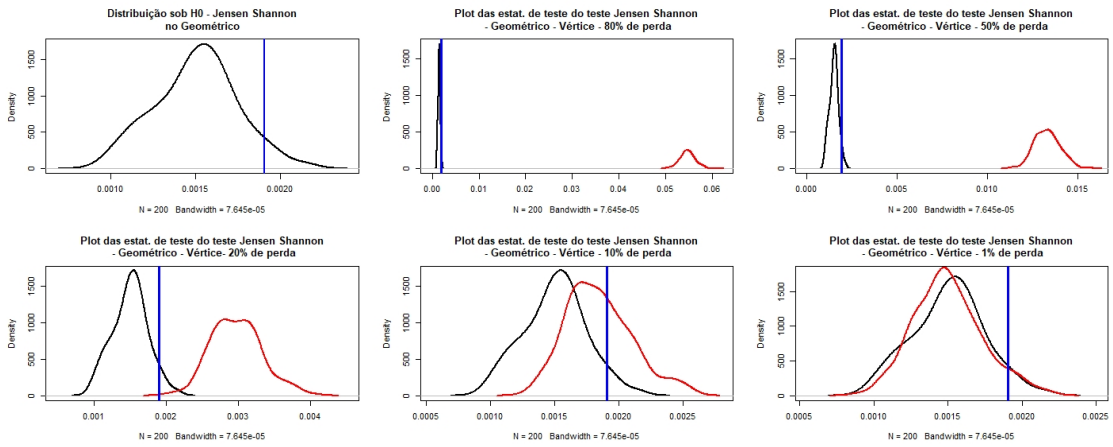


Figura 3.3: Classificação do teste de hipóteses utilizando Jensen Shannon com a Amostragem por Vértices no modelo Geométrico

Espera-se que um grafo que perca 80% ou 50% dos seus vértices tenha uma mudança em relação à sua estrutura e consequentemente haja uma descaracterização entre a distribuição espectral do grafo populacional e o amostral. Isto pode ser visto na Figura 3.3, pois, com 80% ou 50% no modelo GE, rejeitamos todos os testes de que os grafos

venham de uma mesma população ao nível de 5% de significância, isto pode ser facilmente identificado pois, a curva das estatísticas de teste dos grafos amostrados está a direita do percentil 95% da distribuição sob $H_0$, isto é, mais de 95% dos testes realizados são rejeitados quando temos 80% ou 50% para o modelo GE.

Na medida em que diminuímos o nível de perda, temos mais chances de não rejeitar $H_0$, pois, estamos perdendo menos informações sobre a população, observamos esse comportamento nos testes para 20% e 10% de perda de informação.

Podemos ver claramente que as duas curvas vão se aproximando à medida que a perda diminui pois, estamos perdendo menos características do grafo. Para o caso em que temos 1% de perda, notamos que a distribuição da estatística de teste da amostra está a esquerda do percentil 95%, isto é um indício de que não rejeitamos $H_0$ em quase nenhum dos 200 testes realizados, isto a 5% de significância.

Agora que já sabemos como foi construído o nosso teste e como interpretar os nossos resultados, vamos analisar como cada um dos métodos de amostragem e como cada modelo é sensível ao nível de perda de informação.

## 3.4 StatGraph

Takahashi et al. (2012), desenvolvem um pacote para o *software* R denominado *statGraph*. O pacote que foi desenvolvido na Universidade de São Paulo, vinculado ao grupo de pesquisa de Biologia de Sistemas e Sistemas de Neurociências (Santos et al., 2019).

O *"statGraph"* é um pacote capaz de compilar vários métodos estatísticos para grafos. Por exemplo, métodos para estimação de parâmetros, seleção de modelos *(GIC - Graph Information Criterion)*, testes estatísticos para discriminar duas ou mais populações de grafos, correlação entre grafos, etc. Esses recursos serão utilizados no decorrer deste trabalho para atingir os objetivos propostos.

**Função utilizada**

- **GIC:** A função retorna a divergência de Kullback-Leibler entre um grafo não

direcionado e um modelo dado (Santos et al., 2019).

A função recebe como argumento $A$ que representa a matriz de adjacências do grafo que estamos testando, já o parâmetro *"model"* é o modelo no qual queremos comparar com a matriz $A$, $p$ é a probabilidade de formar conexões do modelo ou o parâmetro específico do modelo. Temos a opção de entrar com os autovalores do grafo no lugar da matriz de adjacências, caso seja necessário é só mudar o argumento $eigenvalues = TRUE$.

$$GIC(A, model, p = NULL, eigenvalues = NULL)$$

A função GIC computa os autovalores do grafo, calcula a densidade espectral e em seguida compara com a densidade de um modelo aleatório com um grafo fornecido e retorna a divergência de Kullback-Leibler entre eles.

Utilizaremos a KLD apenas para validar o uso da densidade espectral como instrumento de caracterização do grafo, para essa validação vamos utilizar a função GIC para recuperar o modelo de um grafo através da densidade espectral do mesmo.

# Capítulo 4

# Resultados Preliminares

## 4.1 Tamanho de um grafo populacional

Como resultado inicial, vamos mostrar como o algoritmo proposto por Takahashi et al. (2012) consegue captar o modelo/processo gerador de um grafo utilizando a divergência de Kullback-Leibler. O objetivo deste capítulo é mostrar que com a densidade espectral conseguimos recuperar o modelo gerador do grafo, ou seja, que a densidade espectral de fato consegue manter as características estruturais de um modelo permitindo que consigamos identificar o processo através da mesma.

Geramos grafos com 50, 100, 500 e 1.000 vértices a partir dos modelos apresentados na Seção 2.2.1.1. Refizemos esse processo de gerar grafos 1.000 vezes para cada modelo. Posteriormente, extraímos a matriz de adjacências e a calculamos a densidade espectral destes grafos; a partir disso, podemos classificar qual o modelo do gerador grafo de acordo com o menor KLD, calculada pela função GIC (*Graph Information Criterion*).

Utilizamos a função GIC do pacote *statGraph*, definida na Seção 3.4. Esse algoritmo retorna a divergência de Kullback Leibler entre um grafo observado e um grafo aleatório qualquer de um modelo que você deseja testar. Esperamos que, por exemplo, quando testarmos um grafo *Erdős Rényi* qualquer e calcularmos o GIC para este grafo comparado com todos os outros modelos observados, o modelo que fornecer a menor divergência de Kullback Leibler, seja o modelo responsável por gerar do grafo, neste exemplo, o

modelo de menor KLD deve ser o modelo *Erdős Rényi*, que é o processo gerador do grafo em questão, como definimos no exemplo. Quanto menor é a KLD mais próximo é a distribuição dos dois grafos, ou seja, mais parecida é a densidade espectral.

Com a função *GIC* nós conseguimos calcular a divergência entre dois grafos utilizando a densidade espectral, e como vamos observar a seguir, essa técnica nos permite recuperar o modelo gerador do grafo. Com isso, temos evidências de que a densidade espectral é uma boa ferramenta e que ela de fato é uma medida sintética para as características estruturais do grafo.

### 4.1.1 Sensibilidade da Densidade Espectral como identificadora do modelo do grafo

Como primeiro caso, geramos um único grafo com 100 vértices para cada um dos modelos descritos e observamos o valor da divergência de Kullback Leibler entre o grafo gerado e outros grafos de modelo aleatório. Como descrito anteriormente, esperamos que a menor KLD seja referente ao modelo comparado com ele mesmo, neste caso, isso indica que não existe diferença entre a distribuição espectral do grafo gerado e a de um grafo aleatório qualquer do mesmo modelo. Como primeiro resultado, podemos observar que a menor KLD sempre indica o processo gerador real, ou seja, o algoritmo consegue acertar o modelo, como pode ser visto na Tabela 4.1.

Tabela 4.1: KLD de Grafos aleatórios com 100 vértices para os modelos observados

| Modelo Comparado / Modelo Gerador | ER | BA | WS | GRG |
|---|---|---|---|---|
| ER | 0.00 | 0.58 | 0.13 | 0.77 |
| GRG | 0.24 | 0.16 | 0.55 | 0.01 |
| BA | 0.33 | 0.00 | 0.14 | 0.12 |
| WS | 0.12 | 0.22 | 0.00 | 0.08 |

Na Tabela 4.1 temos em linhas os modelos dos grafos gerados e nas colunas o grafo aleatório que cada modelo foi comparado. Na Tabela 4.1 podemos observar que, quando comparamos a densidade espectral de um grafo *Erdős Rényi* com ele mesmo a KLD

encontrada é 0, mas, quando comparamos esse mesmo modelo *Erdős Rényi* com o modelo Geométrico, *Watts Strogatz ou Barabasi Albert* ela é diferente de 0, indicando que o grafo comparado de fato foi gerado de um modelo *Erdős Rényi* de acordo com os conceitos vistos na seção anterior.

Afim de provar a consistência da técnica, vamos replicar os resultados encontrados acima. Geramos 1.000 grafos de 50, 100, 500 e 1.000 vértices e calculamos a divergência entre o grafo gerado e um grafo aleatório qualquer gerado por outros modelos para compararmos.

Na Tabela 4.2, apresentamos o número de classificações dos 1.000 grafos gerados a partir do modelo Erdős Rényi. Contamos quantos destes grafos gerados foram classificados como Erdős Rényi para cada um dos casos (50, 100, 500 e 1.000 vértices). Observamos que para grafos grandes, com mais de 500 vértices o algoritmo acerta o modelo em todas as vezes. Já para grafos menores, com 50 e 100 vértices, temos alguns poucos casos em que o algoritmo não consegue identificar bem o modelo gerador, porém, isso acontece em poucos casos, menos de 15% dos casos.

Tabela 4.2: Classificação dos modelos gerados a partir do modelo Erdős Rényi (p= 0.007)

| **nº de vértices** | ER | GRG | BA | WS |
|---:|---:|---:|---:|---:|
| N=50 | 848 | 152 | 0 | 0 |
| N=100 | 797 | 116 | 87 | 0 |
| N=500 | 1000 | 0 | 0 | 0 |
| N=1.000 | 1000 | 0 | 0 | 0 |

Vamos agora, verificar o que ocorre quando geramos grafos de outros modelos, como o Geométrico, Barabasi Albert e Watts Strogatz. Assim como no caso anterior, geramos 1.000 grafos com 50, 100, 500 e 1.000 vértices. O resultados estão disponíveis na tabela a seguir.

Tabela 4.3: Classificação dos modelos gerados a partir do modelo Geométrico (r=0.1)

| nº de vértices | ER | GRG | BA | WS |
|---|---|---|---|---|
| N=50 | 32 | 955 | 12 | 1 |
| N=100 | 1 | 999 | 0 | 0 |
| N=500 | 0 | 1000 | 0 | 0 |
| N=1.000 | 0 | 1000 | 0 | 0 |

Na Tabela 4.3 observamos que para o modelo geométrico, o algoritmo apresenta uma quantidade de acerto excelente, quase todos os casos são classificados corretamente até para grafos pequenos $(N = 100)$. Já na Tabela 4.4, referente ao modelo Barabasi Albert, percebermos que para redes pequenas temos um erro de classificação de mais ou menos 26%, o que não ocorre para redes grandes (N=500 e N=1.000) do modelo.

Tabela 4.4: Classificação dos modelos gerados a partir do modelo Barabasi Albert (pl=1)

| nº de vértices | ER | GRG | BA | WS |
|---|---|---|---|---|
| N=50 | 0 | 0 | 741 | 259 |
| N=100 | 0 | 0 | 869 | 131 |
| N=500 | 0 | 0 | 1000 | 0 |
| N=1.000 | 0 | 0 | 1000 | 0 |

Por fim, para o modelo Watts Strogatz, temos um acerto de 100% das classificações, mesmo em redes pequenas (N=50), como pode ser visto na Tabela 4.5.

Tabela 4.5: Classificação dos modelos gerados a partir do modelo Watts Strogatz (p=0.07)

| nº de vértices | ER | GRG | BA | WS |
|---|---|---|---|---|
| N=50 | 0 | 0 | 0 | 1000 |
| N=100 | 0 | 0 | 0 | 1000 |
| N=500 | 0 | 0 | 0 | 1000 |
| N=1.000 | 0 | 0 | 0 | 1000 |

Quando comparamos as classificações obtidas neste trabalho com as realizadas por

Takahashi et al. (2012), os resultados são bem parecidos, indicando mais uma vez que a densidade espectral é uma técnica adequada para sintetizar as estruturas de um grafo. Através dos resultados encontrados, temos evidências de que a densidade espectral consegue com êxito recuperar o modelo gerador de um grafo. Com esse resultados também constatamos que um grafo deve ter um tamanho de no mínimo 500 vértices para que essa metodologia seja aplicada com sucesso e não apresente erros de classificação.

Com este estudo de simulação, analisando conjuntamente todos os modelos testados, temos evidências de que, para um grafo com $N \geq 500$ vértices, o algoritmo proposto não erra nenhuma das classificações do modelo gerador, indicando que a diferença entre os grafos gerados e o modelo real é desprezível a medida que N cresce ($N \geq 500$), isto para os modelos testados.

Portanto, com $N \geq 500$ vértices, já podemos considerar o grafo como sendo uma aproximação da população do modelo gerado. Vamos assumir este resultado como verdade de agora em diante.

# Capítulo 5

# Resultados e Discussão

## 5.1 Teste de hipóteses

A seguir vamos analisar o desempenho de cada um dos métodos de amostragem e como cada modelo é sensível ao nível de perda de informação através do teste proposto neste trabalho e descrito na Seção 3.3.

### 5.1.1 Amostragem por Vértices

Para cada modelo apresentamos os resultados das amostragens por vértices em 6 gráficos. O primeiro é referente a distribuição da estatística de teste de JS sob $H_0$, já os outros 5 são referentes as comparações entre as distribuições das estatísticas de teste de $H_0$ e da amostra com determinado nível de perda ( 80%, 50%, 20%, 10% e 1%).

Na Figura 5.1, temos os resultados para o modelo Erdős Rényi utilizando a amostragem por vértices. Observamos que a proporção de não rejeição do modelo é de 100% em todos os casos. Esse é um forte indício de que o método de amostragem por vértices é bom para realizar amostragem para este modelo pois, quase todos os testes ficam a esquerda do percentil 95% mostrando que em pelo menos 95% dos casos não rejeitamos a hipótese de que o grafo populacional e o amostrado venham da mesma população.

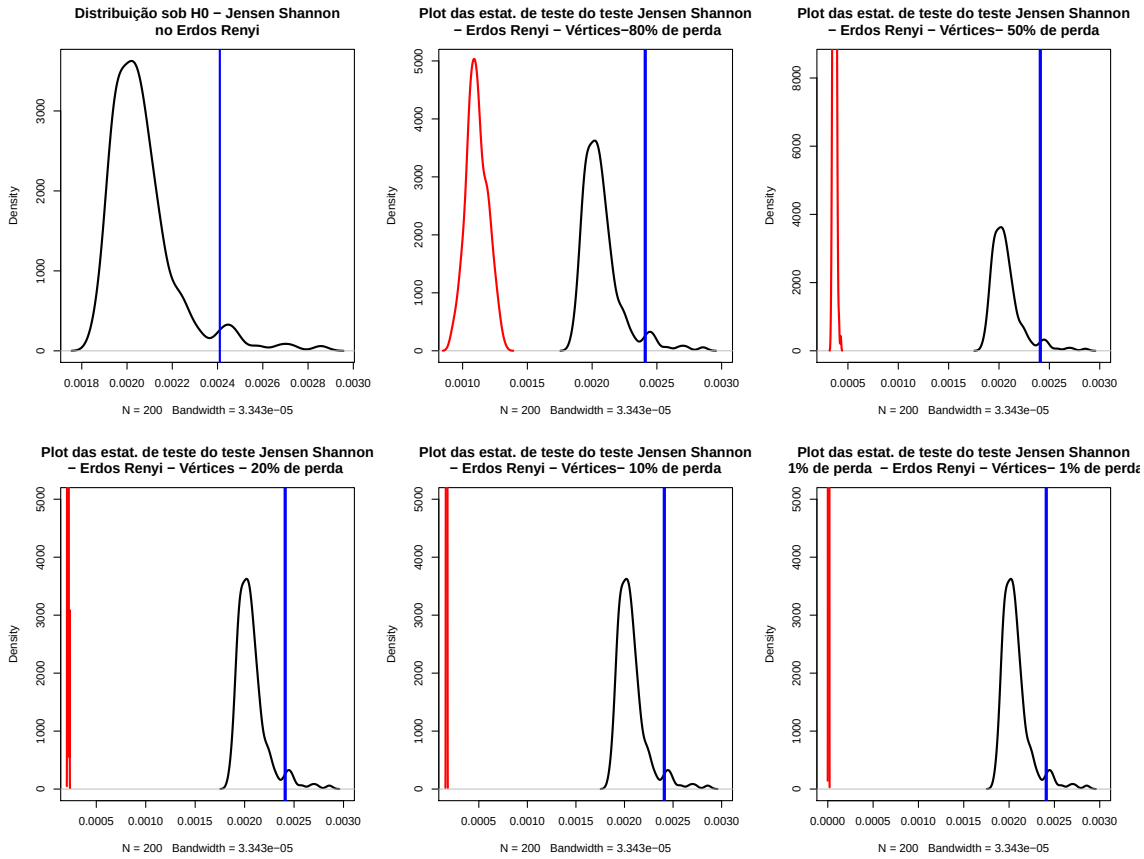


Figura 5.1: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Erdős Rényi por Vértices

Já na Figura 5.2 que contém os resultados para o modelo Geométrico, observamos que com 80% e 50% de perda rejeitamos a hipótese nula do teste de JS em aproximadamente 100% dos casos. Para 20% de perda as curvas já começam a se aproximar e ter interseção, com 10% temos um pouco mais da metade dos casos não rejeitando a hipótese nula e com 1% de perda a taxa de rejeição é aproximadamente 0 pois, a curva esta a esquerda do percentil 95%, mostrando que em pelo menos 95% dos casos não rejeitamos a hipótese de que o grafo populacional e o amostrado venham da mesma população. Isto indica que o método de amostragem por vértices consegue captar bem a estrutura do grafo geométrico desde que a perda máxima seja de 10% dos vértices, ou seja, é um candidato a método de amostragem para modelos com essas características.

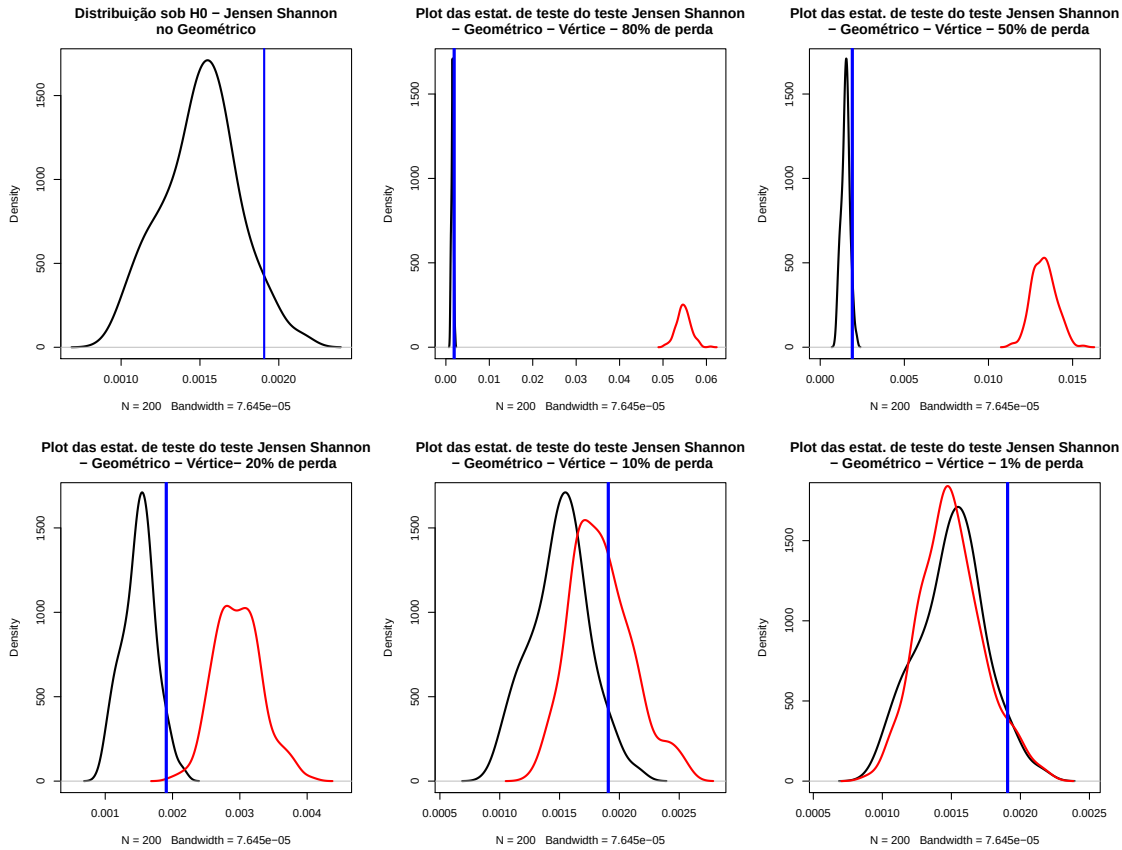


Figura 5.2: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Geométrico por Vértices

Na Figura 5.3 temos os resultados para o modelo Barabasi Albert. Claramente, podemos observar que o método de amostragem por vértices capta bem as caraterísticas da rede apenas quando temos 10% ou menos de perda pois, com 80% ou 50% de perda temos uma alta taxa de rejeição de $H_0$ e conforme vamos diminuindo a perda (20%) o número de rejeições do teste vai ficando cada vez menor e as curvas vão se aproximando.

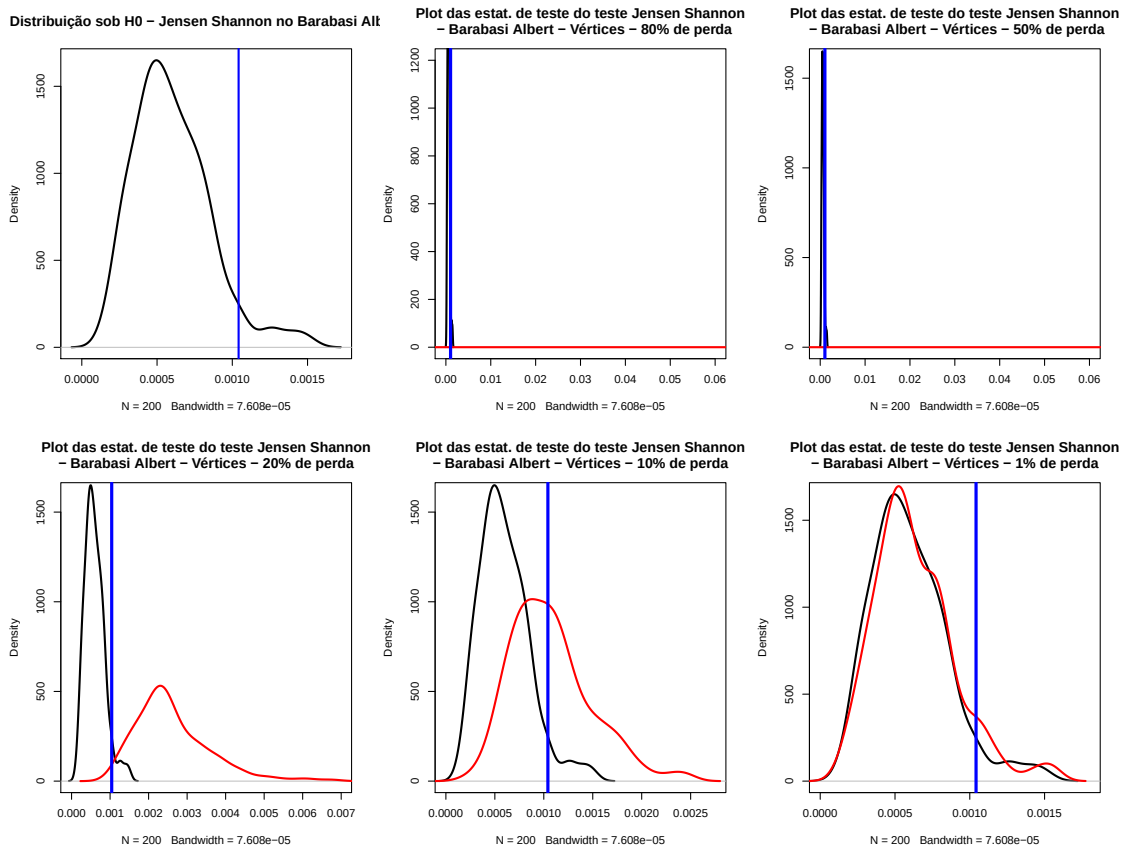


Figura 5.3: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Barabasi Albert por Vértices

A Figura 5.4 apresenta os resultados para o modelo SW. Podemos observar que o método de amostragem por vértices capta bem as caraterísticas da rede pois, mesmo com 80% de perda ainda temos uma boa taxa de não rejeição de $H_0$ e conforme vamos diminuindo a perda o número de rejeições do teste vai ficando cada vez mais rara, ou seja, mesmo com altos níveis de perda o teste ainda consegue associar a estrutura da amostra com o modelo gerador.

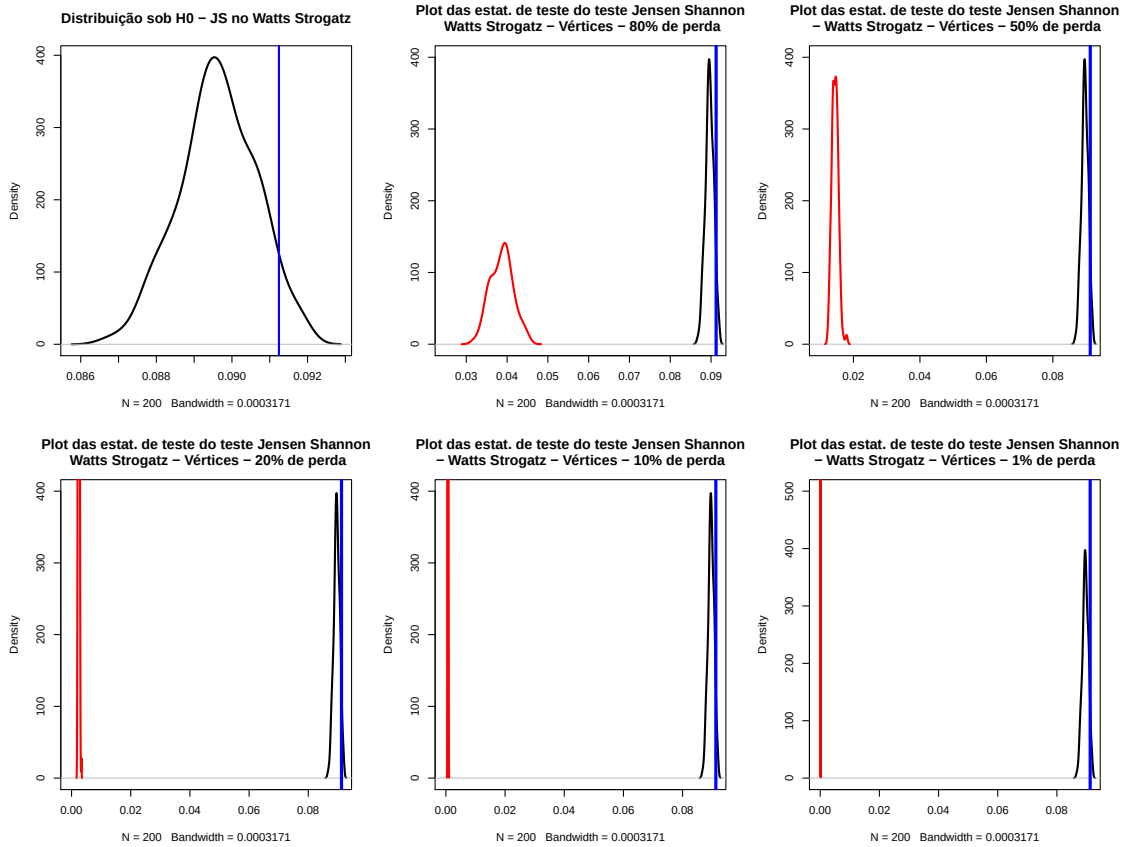


Figura 5.4: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Watts Strogatz por Vértices

### 5.1.2 Amostragem por Arestas

Da mesma forma que no método de amostragem por vértices, apresentamos os resultados das amostragens por arestas em 6 gráficos. O primeiro é referente a distribuição da estatística de teste de JS sob $H_0$, já os outros 5 são referentes as comparações entre as distribuições das estatísticas de teste de $H_0$ e da amostra com determinado nível de perda ( 80%, 50%, 20%, 10% e 1%).

Podemos observar na Figura 5.5, Figura 5.6 e Figura 5.7 que os resultados só são satisfatórios quando temos 10% ou 1% de perda de informação e claramente, o método de amostragem por arestas também não é uma boa técnica para amostrar os modelos Erdős Rényi, Geométrico e Barabasi Albert. Quase todos os testes ficam a direita do percentil 95%, mostrando que em pelo menos 95% dos casos rejeitamos a hipótese de que o grafo populacional e o amostrado venham da mesma população, apenas a partir de 10% de perda

que não vemos esse padrão.

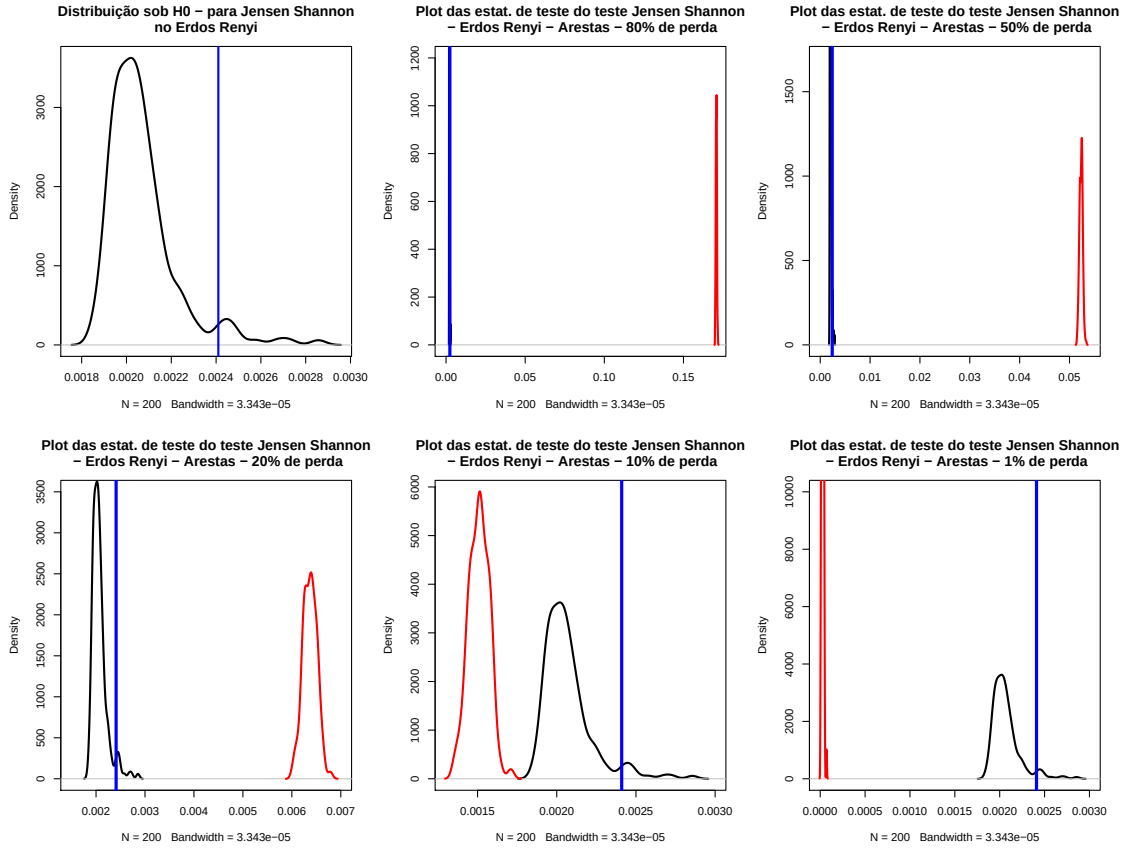


Figura 5.5: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Erdős Rényi por Aresta

Para o modelo Geométrico, os resultados são apresentados na Figura 5.6. O método de amostragem por arestas não é melhor que a amostragem por vértices para este modelo. Observamos que apenas com 1% de perda que as curvas começam a ter interseção e mesmo assim ainda temos uma alta taxa de rejeição de $H_0$, enquanto que, para o mesmo modelo no método de amostragem por vértices, com 10% de perda tínhamos aproximadamente mais da metade dos testes não rejeitados. Logo, concluímos que o método de amostragem por arestas não é bom recurso para amostrar grafos com características estruturais compatíveis com o modelo geométrico pois, a amostragem por aresta não consegue manter as estruturas do grafo mesmo com baixos níveis de perda.

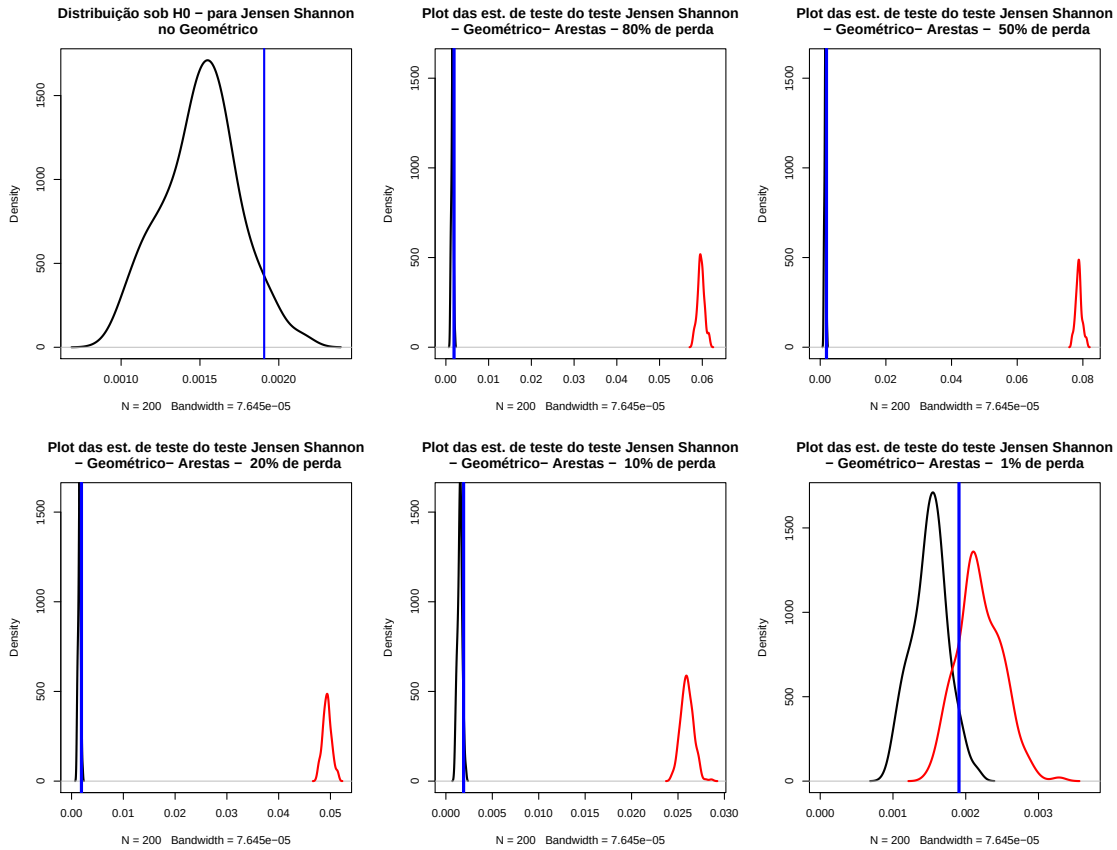


Figura 5.6: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Geométrico por Aresta

O modelo Barabasi Albert apresenta resultados insatisfatórios para a amostragem por arestas. Na Figura 5.7 podemos notar que apenas com 10% de perda de informação que conseguimos manter as estruturas do grafo e então temos quase todos os testes de JS não rejeitando $H_0$, isto é, apenas com 10% de perda de informação, pelo método de amostragem por arestas ainda é possível classificar o grafo amostrado como sendo da mesma distribuição do grafo populacional. Os resultados obtidos para a amostragem por arestas não é mais consistente que os obtidos para o método de amostragem por vértices, indicando que o método não é o mais indicado para o modelo BA.

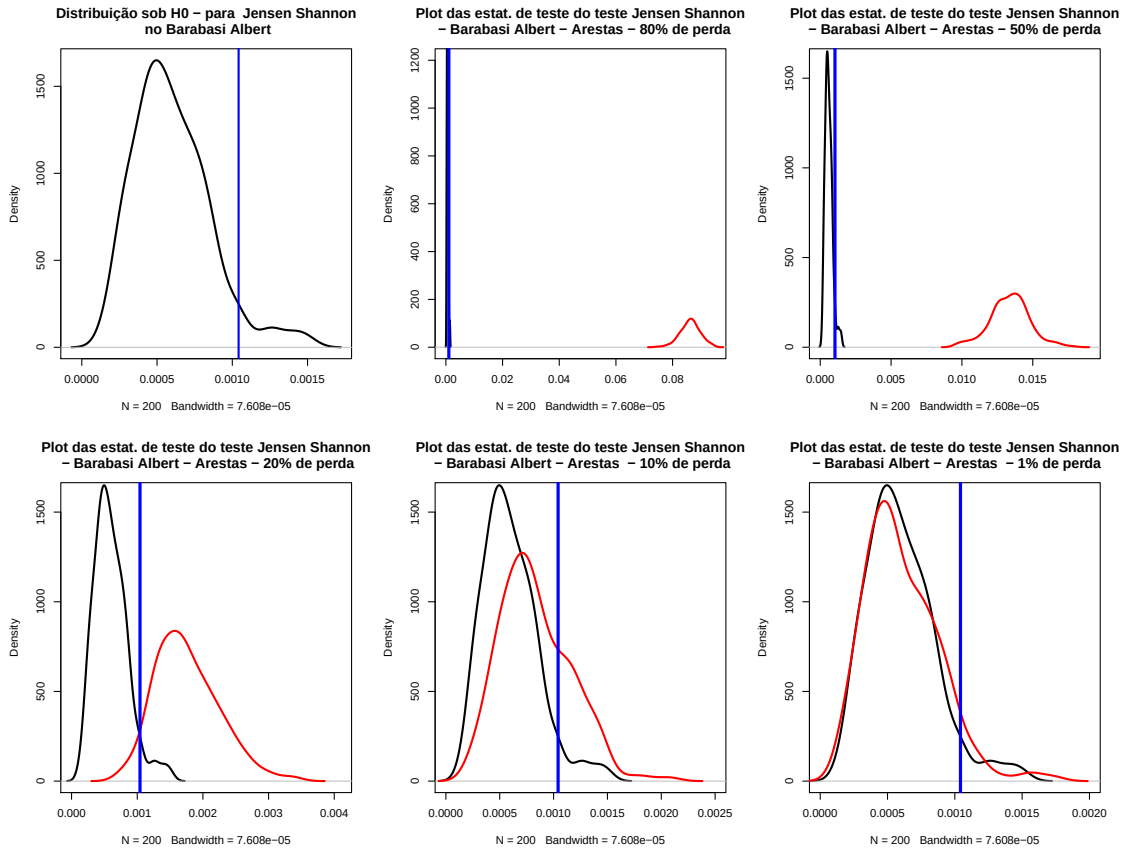


Figura 5.7: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Barabasi Albert por Aresta

Claramente na Figura 5.8, podemos observar que o método de amostragem por vértices capta bem as caraterísticas da rede pois, mesmo com 80% de perda ainda temos uma boa taxa de não rejeição de $H_0$ e conforme vamos diminuindo a perda o número de rejeições do teste vai ficando cada vez mais rara, ou seja, mesmo com altos níveis de perda o teste ainda consegue associar a estrutura da amostra com o modelo gerador.

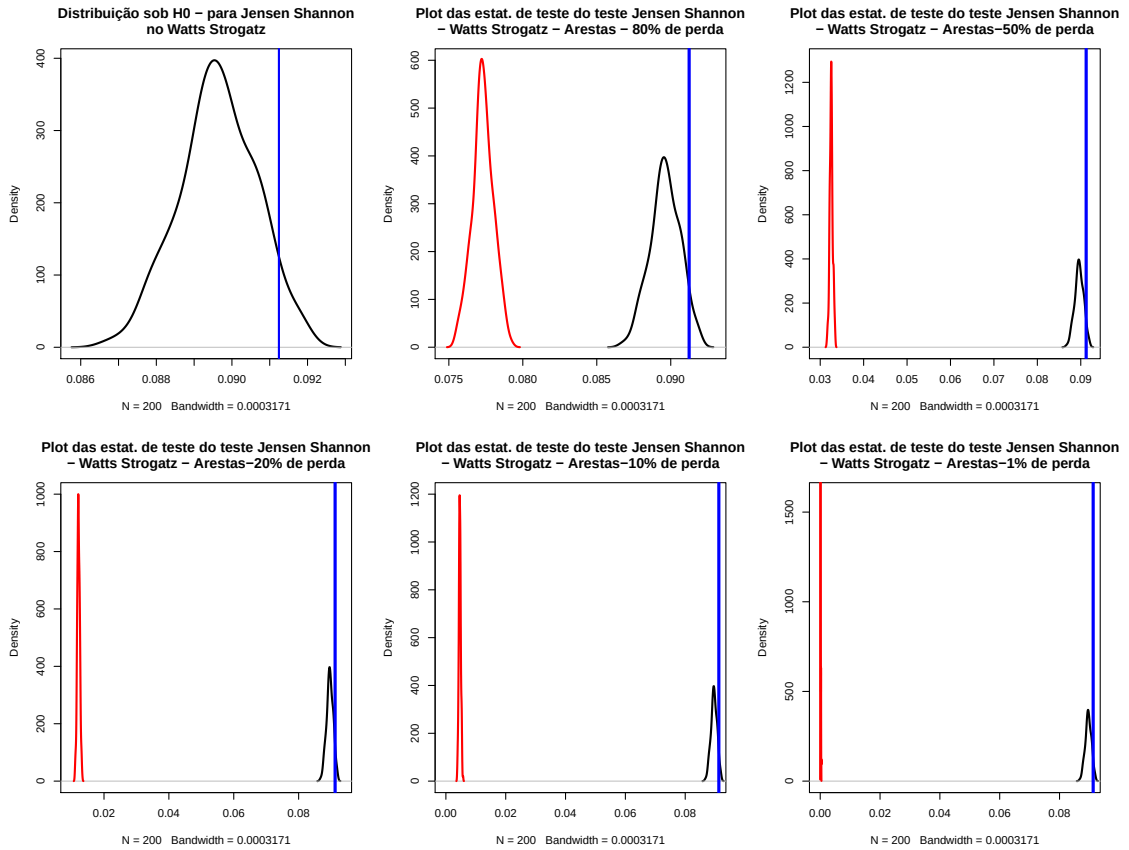


Figura 5.8: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Watts Strogatz por Aresta

### 5.1.3 Amostragem por Bola de neve

#### *5.1.3.1 Amostragem por Bola de neve com nó inicial aleatório*

Pelos gráficos apresentados a seguir, podemos notar que os modelos que não obtiveram resultados satisfatórios para as amostragens por arestas, apresentam melhorias quando realizamos amostragem por Bola de Neve. Os modelos em questão são o Erdős Rényi, Geométrico e Barabasi Albert que antes só tinham resultados satisfatórios com 1% de perda, quando utilizamos amostragem por bola de neve com o vértice inicial sendo aleatório podemos perceber pela Figura 5.9 que mesmo com 80% de perda já temos todos os testes a esquerda do percentil 95%, mostrando que em pelo menos 95% dos casos não rejeitamos a hipótese de que o grafo populacional e o amostrado venham da mesma população, isto para o modelo ER.

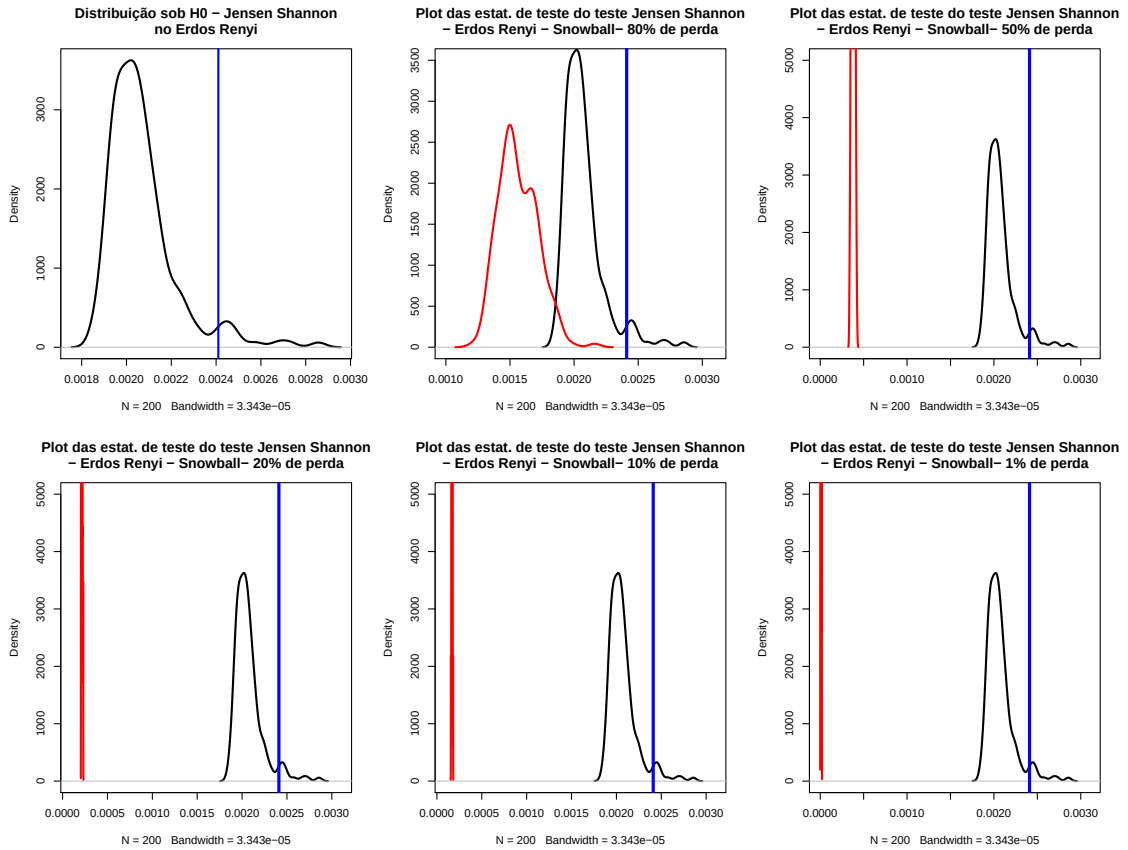


Figura 5.9: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Erdős Rényi por Bola de Neve (Grau aleatório como inicial)

Na Figura 5.11 e Figura 5.10 constatamos que o método de amostragem por bola de nele com o vértice inicial aleatório também é um bom método para capturar as estruturas dos grafos Geométrico e Barabasi Albert.

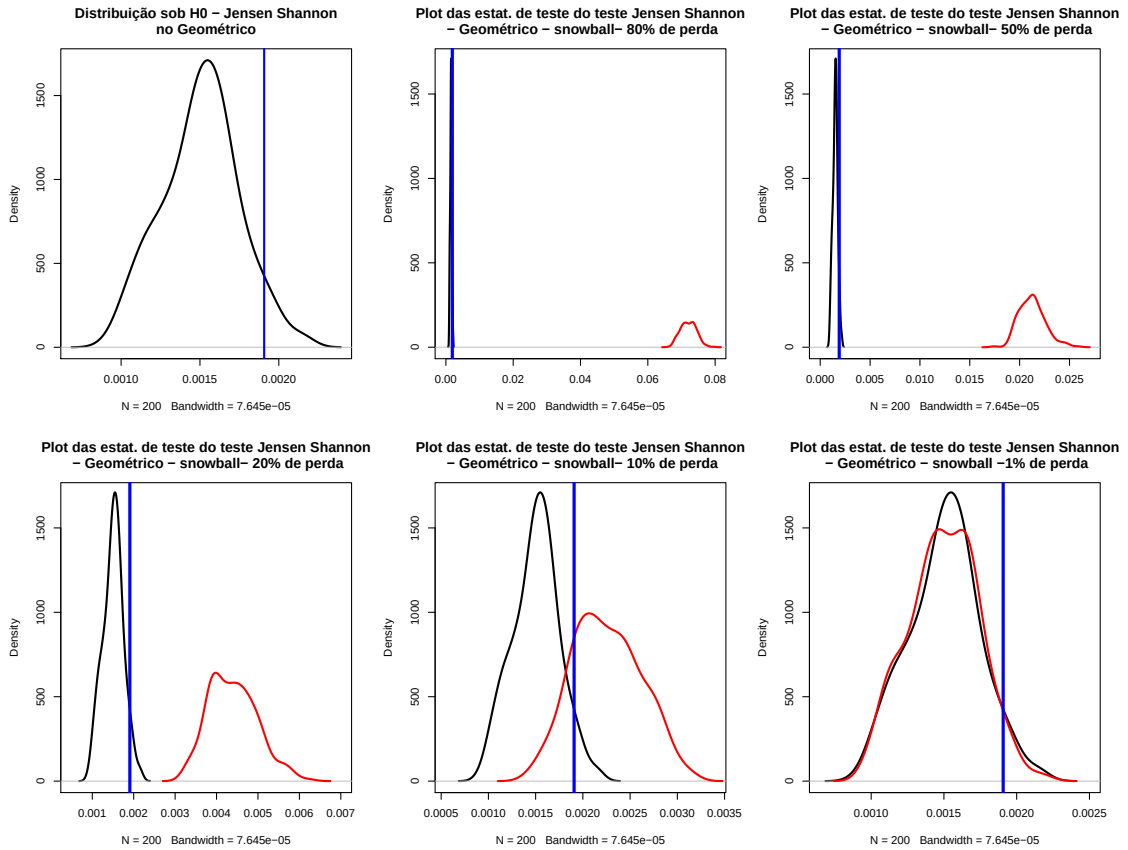


Figura 5.10: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Geométrico por Bola de Neve (Grau aleatório como inicial)

Para o modelo BA, os resultados obtidos pela amostragem por bola de neve não são tão bons quanto os encontrados para o método de amostragem por vértices e arestas, temos mais da metade dos casos rejeitando $H_0$ mesmo com 10% de perda.

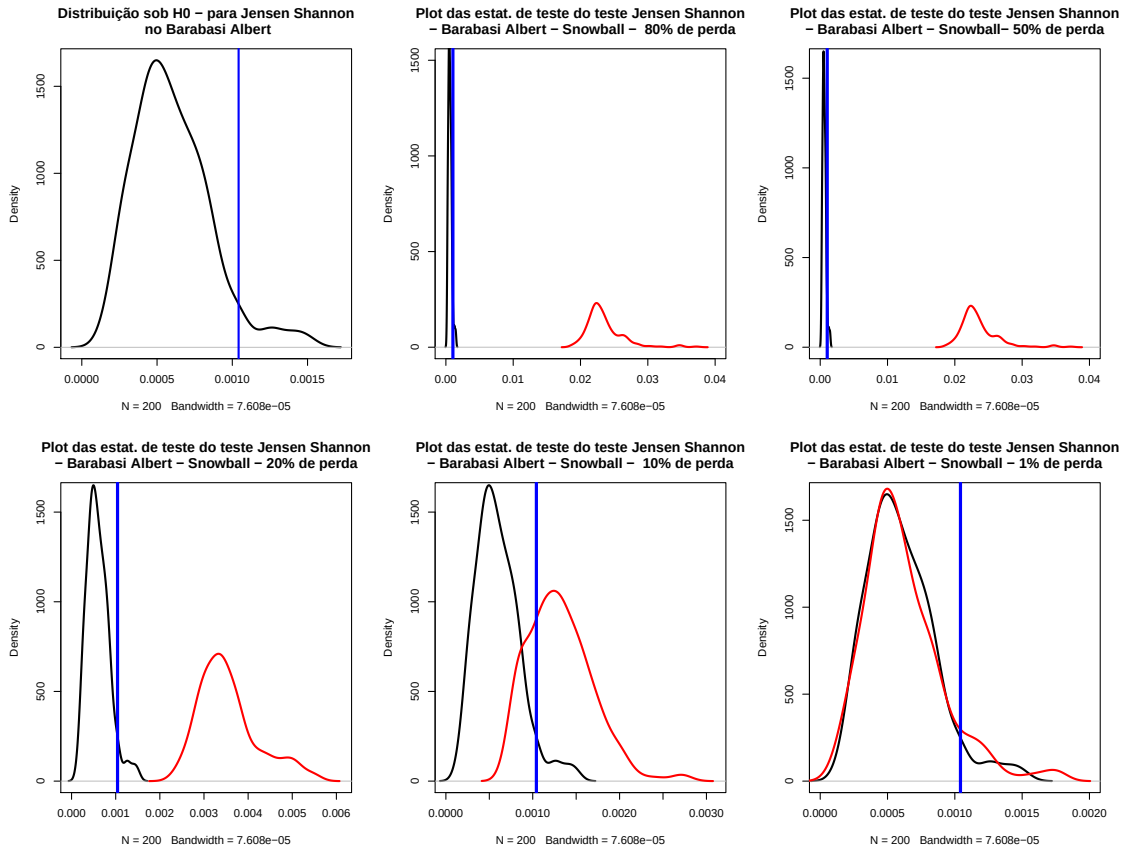


Figura 5.11: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Barabasi Albert por Bola de Neve (Grau aleatório como inicial)

O modelo Watts Strogatz apresentado na Figura 5.12 apresenta bons resultados mesmo com 80% de perda, porque ainda temos uma boa taxa de não rejeição de $H_0$ e conforme vamos diminuindo a perda o número de rejeições do teste vai ficando cada vez mais rara, ou seja, mesmo com altos níveis de perda o teste ainda consegue associar a estrutura da amostra com o modelo gerador.

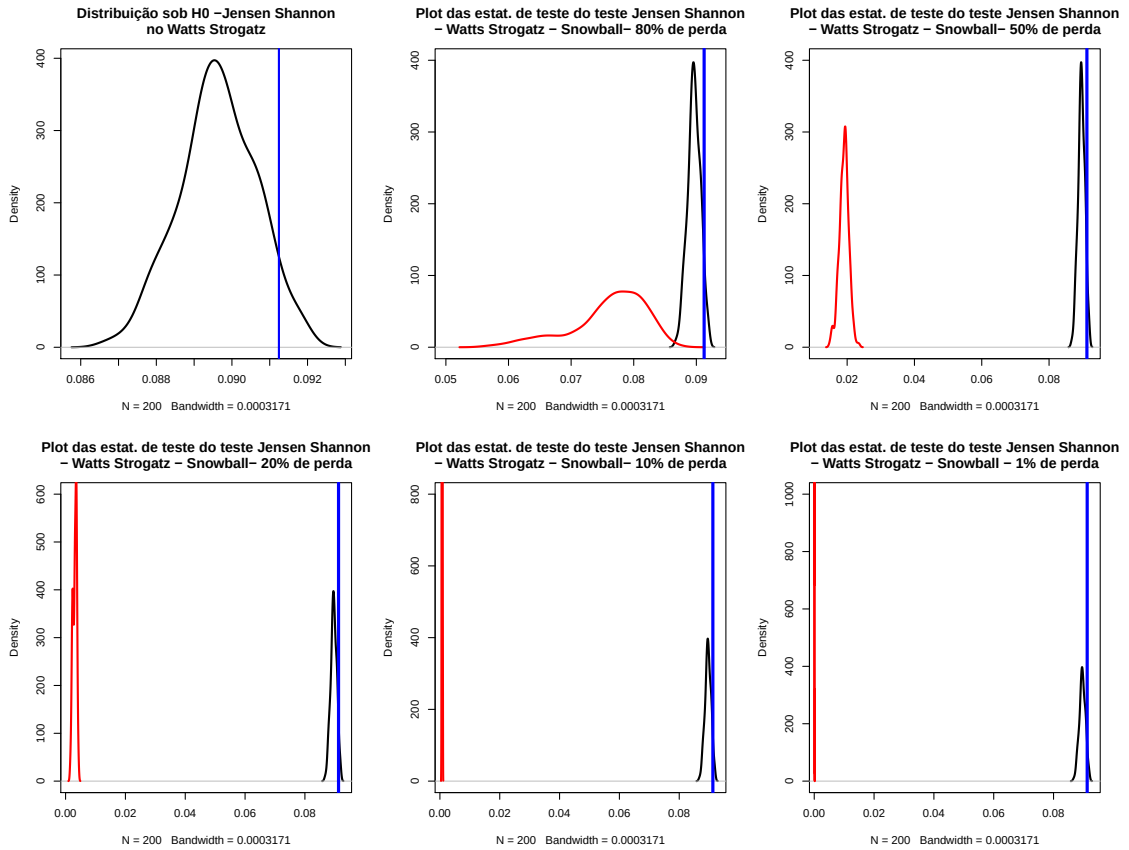


Figura 5.12: Classificação do teste de hipóteses utilizando Jensen Shannon - Amostragem para o modelo Watts Strogatz por Bola de Neve (Grau aleatório como inicial)

# Capítulo 6

# Conclusão

Esta dissertação apresentou uma proposta metodológica para identificar e quantificar a sensibilidade de métodos de amostragem para vários modelos de grafos aleatórios já consolidados na literatura no intuito de encontrar qual seria o melhor método amostragem para cada um dos modelos analisados. Para conseguir resultados satisfatórios, nós introduzimos o uso da densidade espectral como uma informação sintética a respeito da estrutura do grafo. Por essa razão, ao aplicarmos as técnicas de amostragem utilizando a densidade espectral conseguimos analisar várias características do grafo simultaneamente e oferecer resultados mais robustos. Nossa proposta metodológica é capaz de identificar qual o método de amostragem é mais indicado para cada tipo de modelo de grafo aleatório e também qual é a porcentagem máxima de perda de informação que o método suporta para continuar recuperando e identificando o modelo gerador.

A Tabela 6.1 resume os resultados obtidos neste trabalho, identificamos para cada modelo testado, quais são os métodos de amostragem mais adequados e qual é o nível máximo de perda de informação que cada método sustenta sem perder as propriedades estruturais e identificáveis do grafo.

Tabela 6.1: Método de amostragem mais indicado de acordo com cada modelo de grafo aleatório e o percentual de perda de informação que o método suporta

| **Modelo** | **Técnicas de Amostragem indicada** | **Percentuais de perda aceitáveis** |
|---|---|---|
| Erdős Rényi (ER) | Amostragem por Bola de neve e Vértices | 1%, 10%, 20%,50%, 80% |
| Geométrico (GE) | Amostragem por Vértice | 1%, 10% |
| Barabasi Albert (BA) | Amostragem por Vértices | 1%, 10% |
| Watts Strogatz (WS) | Amostragem por Bola de neve,Vértices,Arestas | 1%, 10%, 20%,50%, 80% |

Segundo os resultados obtidos, a amostragem por bola de neve e a amostragem por vértices são os métodos mais apropriados para quase todos os modelos testados, os resultados não são satisfatórios apenas para o modelo Geométrico, que apresenta resultados inferiores para a amostragem por bola de neve. Os níveis de perda que cada modelo suporta depende da estrutura de cada um deles. Os modelos Erdős Rényi e Watts Strogatz conseguiram manter as estruturas do grafo com uma perda de informação de até 80%.

Nos modelos Geométrico e Barabási Albert, os métodos de amostragem conseguem manter bons resultados com no máximo 10% de perda. Para perdas maiores que esse percentual a metodologia não garante que a estrutura do grafo será mantida e, consequentemente, que o modelo gerador correto será associado à amostra.

Vale ressaltar que os resultados são consistentes mesmo quando diversificamos os parâmetros de cada modelo e o tamanho dos grafos. Os resultados desta diversificação podem ser observados no Apêndice A deste trabalho.

O trabalho tem como conclusão que, existem diferenças evidentes na forma como cada método de amostragem preserva a estrutura de um grafo e como cada uma dessas técnicas é sensível à perda de informação. Com a nossa metodologia, conseguimos identificar qual é o método de amostragem mais indicado para amostrar os modelos de grafos aleatórios observados e ainda, qual é a quantidade máxima de perda de informação que cada técnica consegue suportar e continuar preservando as estruturas topológicas do modelo.

# Referências


Abreu, N. (2005). Teoria espectral dos grafos: um híbrido entre a álgebra linear e a matemática discreta e combinatória com origens na química quântica. *TEMA-Tendências em Matemática Aplicada e Computacional*, 6(1):1–10.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Csardi, G., Nepusz, T., et al. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Cvetkovic, D., Cvetković, D. M., Rowlinson, P., and Simic, S. (1997). *Eigenspaces of graphs*. Number 66. Cambridge University Press.

Dehmer, M., Emmert-Streib, F., Chen, Z., Li, X., and Shi, Y. (2017). *Mathematical foundations and applications of graph entropy*, volume 6. John Wiley & Sons.

Dorogovtsev, S. N. and Mendes, J. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford.

Erdös, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Gersting, J. L. (2001). *Fundamentos matemáticos para a ciência da computação*. Livros Tecnicos e Cientificos.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Goodman, L. A. (1961). Amostra de bolas de neve. *Os anais da estatística matemática*, pages 148–170.

Hidalgo, C. A. and Rodríguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024.

Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798):845.

Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM.

Mastrandrea, R., Fournet, J., and Barrat, A. (2015). Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, 10(9):e0136497.

Mowshowitz, A. and Dehmer, M. (2012). Entropy and the complexity of graphs revisited. *Entropy*, 14(3):559–570.

Penrose, M. et al. (2003). *Random geometric graphs*, volume 5. Oxford university press.

Rezvanian, A., Rahmati, M., and Meybodi, M. R. (2014). Sampling from complex networks using distributed learning automata. *Physica A: Statistical Mechanics and its Applications*, 396:224–234.

Santos, S. S., Lira, E. S., Fujita, A., and Fujita, M. A. (2019). Package 'statgraph'.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690.

Smith, J. A. and Moody, J. (2013). Structural effects of network sampling coverage i: Nodes missing at random. *Social networks*, 35(4):652–668.

Smith, J. A., Moody, J., and Morgan, J. H. (2017). Network sampling coverage ii: The effect of non-random missing data on network measurement. *Social networks*, 48:78–99.

Takahashi, D. Y., Sato, J. R., Ferreira, C. E., and Fujita, A. (2012). Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS One*, 7(12):e49949.

Team, R. (2015). Rstudio: integrated development environment for r. rstudio. *Inc., Boston, MA*, 14.

Wagner, C., Singer, P., Karimi, F., Pfeffer, J., and Strohmaier, M. (2017). Sampling from social networks with attributes. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1181–1190. International World Wide Web Conferences Steering Committee.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440.

# Apêndice A

## A.1 Especificações do servidor utilizado para gerar os resultados

Para gerar os resultados apresentados deste trabalho utilizamos o servidor do CEDEPLAR - UFMG.

Os principais recursos deste servidor são:

- Poder de processamento – dois de núcleos de 10 cores, totalizando 20 núcleos;
- Capacidade de armazenamento convencional – doze discos SAS de 1,8 TB, totalizando cerca de 21 TB;
- Capacidade de armazenamento de alta performance – quatro discos SSD de 0,8 GB, totalizando cerca de 3TB.

## A.2 Registro de tempo computacional para os algoritmos de simulação

Na tabela A.1, temos os registros do tempo computacional (em minutos) gasto para realizar o método proposto utilizando grafos com 3.000 vértices e 200 réplicas Monte carlo.

Tabela A.1: Tempo computacional em minutos da aplicação do método em grafos com 3.000 vértices e 200 réplicas MC

| Modelos | Tempo para Amostragem por vértices | Tempo para Amostragem por Arestas | Tempo para Amostragem por bola de neve |
|---|---|---|---|
| **Erdös Rényi (ER)** | 42.4 | 77.62 | 49.8 |
| **Geométrico (GE)** | 42.6 | 76.4 | 141.1 |
| **Barabasi Albert (BA)** | 42.6 | 44.1 | 147.4 |
| **Watts Strogatz (WS)** | 39.5 | 75.1 | 11.9 |

## A.3 Diversificação dos parâmetros dos modelos de grafos aleatórios

Nesta seção vamos apresentar alguns resultados da aplicação da metodologia proposta neste trabalho aos modelos descritos na Seção 2.2.1.1 diversificando os parâmetros. Para mostrar todos os resultados encontrados no texto principal seria inviável pois, o mesmo se tornaria muito extenso. Como os resultados são consistentes independente do parâmetro do modelo, apresentamos os resultados da diversificação no Apêndice A.

Vamos apresentar na Tabela A.2 a lista de parâmetros que foram testados para cada um dos modelos de grafos aleatórios.

Tabela A.2: Lista de parâmetros testados para cada um dos modelos

| Modelos | Parâmetros Testados |
|---|---|
| Erdös Rényi (ER) | p= 0,07 ; p= 0,3 ; p= 0,8 |
| Geométrico (GE) | r = 0,1 ; r =0,5; r = 2 |
| Barabasi Albert (BA) | pl = 1 ; pl = 2; pl = 3 |
| Watts Strogatz (WS) | p= 0,07 ; p= 0,3 ; p= 0,8 |

A seguir, temos os resultados da metodologia proposta neste trabalho aplicada aos métodos de amostragem por vértices e amostragem por arestas nos modelos de grafos aleatórios conforme descritos na Tabela A.2. Os resultados são consistentes com os

apresentados no corpo do texto principal no Capítulo 5. Todas as diversificações foram feitas considerando grafos com 3.000 vértices. As densidades espectrais foram estimadas considerando 1.000 pontos.