UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Estimação do número de comunidades em redes ponderadas

**Luana Ayumi Tamura**

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

**Luana Ayumi Tamura**

# Estimação do número de comunidades em redes ponderadas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Andressa Cerqueira

**USP – São Carlos**
**Março de 2025**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Tamura, Luana Ayumi
T159e Estimação do número de comunidades em redes ponderadas / Luana Ayumi Tamura; orientador Andressa Cerqueira. -- São Carlos, 2025.
73 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Redes. 2. Teste de Hipóteses para redes binárias. 3. Teste de Hipóteses para redes ponderadas. 4. Simulações. 5. Aplicações. I. Cerqueira, Andressa, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**Luana Ayumi Tamura**

# Estimating the number of communities in weighted networks

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Andressa Cerqueira

**USP – São Carlos**
**March 2025**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

## Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Luana Ayumi Tamura, realizada em 24/02/2025.

### Comissão Julgadora:

Profa. Dra. Andressa Cerqueira (UFSCar)

Prof. Dr. Aline Duarte de Oliveira (IME-USP)

Prof. Dr. Jodavid de Araújo Ferreira (UFPE)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

# AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer aos meus pais Sergio e Helena pelo apoio dado durante todos os meus anos de graduação e mestrado.

Gostaria de agradecer à minha orientadora e professora Andressa Cerqueira por ser mentora e meu suporte ao longo dessa jornada acadêmica. Você é minha inspiração como professora e pesquisadora.

Agradeço a todos os meus amigos do ICMC e da UFSCar, aos estatísticos e matemáticos. Em especial, gostaria de agradecer aos meus amigos Catharina, Gustavo Menani, Leonardo Gustavo, Luben, Rodrigo e Paulo, por sempre estarem ao meu lado durante esta etapa.

Quero agradecer ao Gustavo Kodama pelo companheirismo, apoio e paciência durante todos os anos do mestrado.

Por fim, gostaria de agradecer a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) pelo apoio financeiro (Código de Financiamento 001) que foi importante e necessária para a realização do projeto.

# RESUMO

TAMURA, A. L. **Estimação do número de comunidades em redes ponderadas**. 2025. 73 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

O Modelo Estocástico de Blocos é um modelo comumente utilizado em redes reais que modela a estrutura da comunidade, isto é, os vértices da rede são divididos em grupos. No entanto, o número de comunidades fornecidas pelo modelo pode não ser conhecido, então é necessário usar um método inferencial para estimar esse número. O objetivo deste trabalho é estudar um método baseado em teste de hipóteses para a estimação do número de comunidades em redes binárias, adaptá-lo para redes com pesos e realizar um estudo via simulação para compreender a performance do teste nesse caso. Além disso, aplicamos testes de hipóteses a rede com pesos do sistema BART (Bay Area Rapid Transit) que é um sistema de transporte rápido que atende à área da baía de São Francisco, para estimar o número de comunidades e analisar a atribuição das estações de trens às comunidades.

**Palavras-chave:** teste de hipóteses, grafos aleatórios, detecção de comunidades, redes ponderadas.

# ABSTRACT

TAMURA, A. L. **Estimating the number of communities in weighted networks**. 2025. 73 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

The Stochastic Block Model is a commonly used model in real networks that exhibits community structure, that is, the vertices of the network are divided into groups. However, the number of communities related to the underlying model is not specified in real data, so it is necessary to use inferential methods to estimate this number. The objective is to study a method based on hypothesis tests that estimates the number of communities in binary networks, adapt it to weighted networks and perform a simulated analysis to comprehend the performance in this case. Furthermore, we applied hypothesis tests to estimate the number of communities and analyze the attribution of stations to communities in BART's (Bay Area Rapid Transit) weighted network system, which is a rapid transit system that covers San Francisco's bay.

**Keywords:** test of hypotheses, random graphs, communities detection, weighted networks.

# LISTA DE ILUSTRAÇÕES

# SUMÁRIO

CAPÍTULO

1

# INTRODUÇÃO

As redes descrevem interações entre objetos ou indivíduos, de forma individual ou coletiva, em que existe a possibilidade de conexões entre objetos e quando analisamos esse comportamento é esperado que identifiquemos um padrão nessas conexões entre os objetos. Newman (2010) aborda a existência de muitos sistemas de interesse pela comunidade científica, em que o foco está relacionado às aplicações da teoria estatística em redes reais, sendo essas ciências físicas, sociais, biológicas ou de informação.

As redes aleatórias estão relacionadas as redes reais, pelo fato das interações entre os objetos serem imprevísiveis. Assim, Newman (2010) considera que as conexões possuem um comportamento aleatório, ou ainda, seguem alguma distribuição de probabilidade. Para compreender os efeitos das redes aleatórias é necessário utilizar modelos matemáticos ou estatísticos para analisar as características das redes. Uma possível análise de uma rede é através da teoria do espectro de um grafo, que aborda propriedades estruturais decorrentes de matrizes de adjacência do grafo.

Os objetos da rede são denominados como vértices e as conexões entre eles são chamadas de arestas. Quando os vértices possuem muitas conexões em relação a outros ou possuem características em comum, podemos associá-los à grupos ou comunidades, de acordo com Lee e Wilkinson (2019). O estudo de redes com estrutura de comunidades, tem como objetivo a análise de padrões e o comportamento entre as comunidades, por exemplo: se estamos interessados em estudar dois cursos de graduação, Estatística e Matemática, a probabilidade das pessoas dentro de seus cursos se conhecerem é diferente da probabilidade das pessoas entre os cursos de conhecerem. Nesse caso, temos uma rede com dois grupos de indivíduos, gerando assim uma rede com estrutura de grupos. Lee e Wilkinson (2019) descreve uma rede como sendo conjuntos de comunidades de vértices, onde os vértices da mesma comunidade tem maior probabilidade de estarem conectados entre si do que com vértices de outras comunidades, uma maneira de estudar esse comportamento é através do modelo estocástico de blocos.

Em redes reais, as conexões entre os vértices não são simples, mas variam em intensidade. Por exemplo, em uma rede social, o grau de amizade entre dois indivíduos pode ser mais forte em alguns casos do que em outros. Essa intensidade atribui valores numéricos (discretos ou contínuos) às conexões entre seus vértices, chamados de pesos, refletindo a intensidade ou a força das interações. Desse modo, estudar redes com pesos oferece uma representação mais precisa das conexões entre os elementos, permitindo uma modelagem mais realista e detalhada de fenômenos complexos. As redes binárias, por sua vez, simplificam as conexões para a existência ou a ausência de ligação, sendo úteis em contextos simples ou em modelos iniciais.

O interesse em estudar comunidades em redes aleatórias é analisar como são os padrões dentro e entre os vértices das comunidades, além de identificar se as conexões dos vértices podem ser descritas através de uma distribuição de probabilidade. Entretanto, em redes reais o número de comunidades não é conhecido, então é necessário utilizar métodos de inferência para determinar este número.

### 1.0.1 Contribuições

O objetivo deste trabalho é estudar o teste de hipóteses sequencial proposto na literatura por Lei (2016) para estimar o número de comunidades em redes binárias e adaptá-lo para redes com pesos. Como no trabalho de Lei (2016) não foi abordado a aplicação do teste de hipóteses para redes com pesos, não há resultados teóricos para o estimador do número de comunidades. O objetivo desta dissertação é explorar via estudos de simulação a eficiência do estimador $\hat{K}$ para estimar o número de comunidades em redes com pesos discretos, descritos pela distribuição de Poisson.

## 1.1 Organização do trabalho

Inicialmente, no Capítulo 2 são apresentados conceitos relacionados à teoria de grafos. Assim como, os modelos de grafos aleatórios utilizados nos estudos para a compreensão do método.

No Capítulo 3 é apresentado o teste de hipóteses sequencial para estimar o número de comunidades em redes binárias proposto por Lei (2016). Ainda, verificamos algumas construções realizadas no teste como o Estimador de Máxima Verossimilhança e resultados de matrizes aleatórias.

No Capítulo 4 é proposto a adaptação do teste de hipóteses sequencial para estimar o número de comunidades em redes ponderadas. Calculamos para redes ponderadas, o Estimador de Máxima Verossimilhança utilizado no teste de hipóteses.

No Capítulo 5 é realizado estudos via simulações para verificar a eficiência do teste de hipóteses para redes ponderadas em diferentes tipos de redes, quanto ao balanceamento, o

número de vértices e o peso médio entre as conexões dos vértices. Além disso, realizamos a comparação dos métodos do teste de hipóteses com outros métodos como *Fast Greedy*, *Louvain* e *Walktrap*.

No Capítulo 6 é realizada uma aplicação em dados reais dos testes de hipóteses para estimar o número de comunidades em redes ponderadas. Comparamos os métodos "Spectral Clustering" e "Spectral Clustering Spherical" na atribuição dos vértices às comunidades e sua interpretação no contexto.

Por fim, no Capítulo 7 são feitas as considerações finais a respeito dos resultados obtidos até o momento e os próximos passos a serem desenvolvidos.

CAPÍTULO

# 2

# REDES

Este capítulo introduz a definição de grafos, discutindo os elementos essenciais que compõem essa estrutura. Na literatura, existem dois modelos conhecidos que abordaremos na Seção 2.3 o Modelo de Erdös-Rényi para estudar grafos aleatórios e na Seção 2.4 o Modelo Estocástico de Blocos para estudar estruturas de comunidades em grafos aleatórios.

## 2.1 Redes

O estudo de redes no ramo da matemática é conhecido como teoria dos grafos, que estuda as relações entre elementos de um conjunto, por meio de estruturas que consistem em vértices e arestas. A ideia básica de rede é uma coleção de objetos que podem estar conectados entre si. Esses objetos são representados por vértices e as relações entre eles são representadas graficamente por linhas conectando pares de vértices, como mostrado na Figura 1.

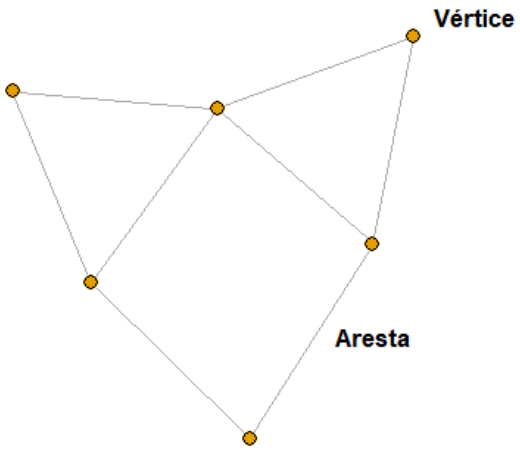


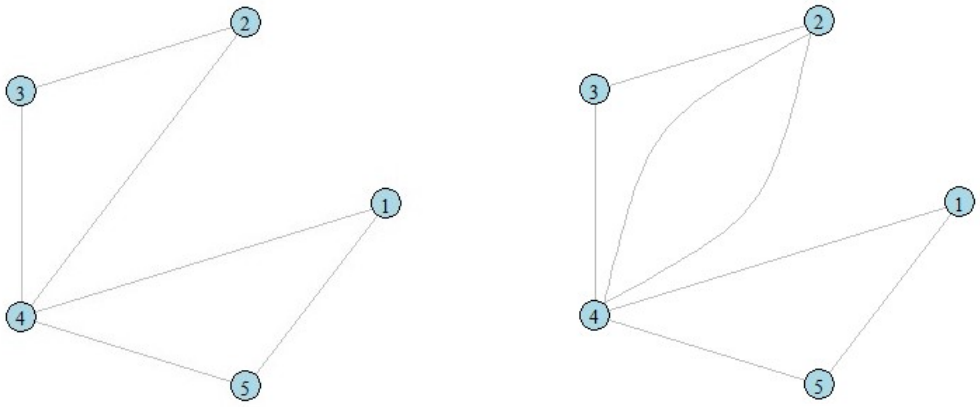

Figura 1 – Uma rede composta de 6 vértices e 8 arestas

Desta forma, a seguir apresentamos alguns conceitos teóricos básicos usadas para descre-

ver e analisar redes, as quais provêm principalmente da teoria dos grafos.

Existem várias maneiras para rotular e representar uma rede matematicamente, nesse estudo os vértices de um grafo serão rotulados com números inteiros positivos: $1, 2, \ldots, n$, para evitar ambiguidade ao nos referirmos a qualquer vértice.

Um grafo simples tem somente uma única aresta entre pares de vértices representado na Figura 2a e o multigrafo tem pelo menos um par de vértices com duas ou mais arestas representado na Figura 2b.




(a) Grafo simples composto de 5 vértices e 6 arestas

(b) Multigrafo composto de 5 vértices e 7 arestas

Figura 2 – Representação dos grafos

Um grafo simples pode ser representado por sua matriz de adjacência $\mathbf{A}$, cuja cada entrada indica a existência de arestas entre pares de vértices. Cada entrada a matriz de adjacência $\mathbf{A}$ é definida:

$$\mathrm{A}_{ij} = \begin{cases} 1, & \text{se existe uma aresta entre os vértices } i \text{ e } j \\ 0, & \text{caso contrário,} \end{cases}$$

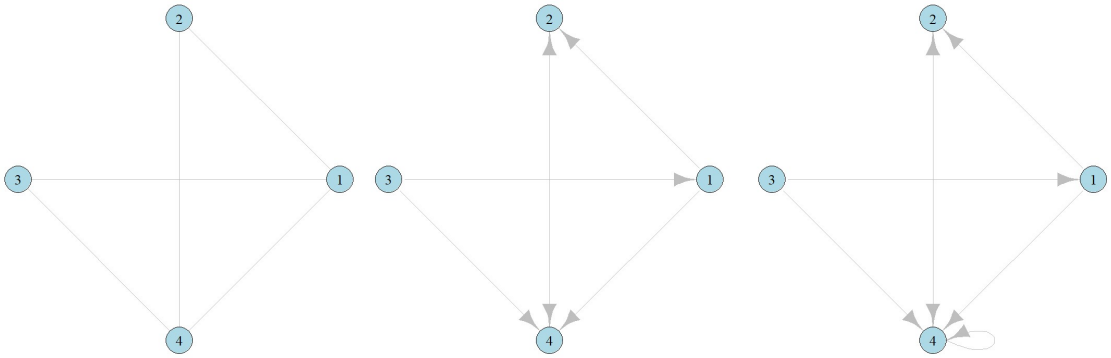
em que $1 \leq i < j \leq n$.

Por exemplo, o grafo representado na Figura 2a, tem a matriz de adjacência dada por:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Um grafo não direcionado tem arestas sem direção, ou seja, se existe uma aresta entre $i$ e $j$, então existe uma aresta entre $j$ e $i$, representado na Figura 3a. Um grafo direcionado tem arestas com direção, isto significa dizer que entre dois vértices $i$ e $j$, se existe uma aresta de $i$ para $j$ não necessariamente há uma aresta de $j$ para $i$, representado na Figura 3b. Um grafo com

laços tem pelo menos um vértice no qual está conectado com si mesmo, representado na Figura 3c.




(a) Grafo não direcionado (b) Grafo direcionado (c) Grafo com laço

Figura 3 – Representação dos grafos

As matrizes de adjacência para os respectivos grafos na Figura 3 são representados na Figura 4. Na Figura 4a temos um grafo não direcionado com uma matriz de adjacência simétrica e com elementos zero na diagonal principal. Na Figura 4b temos um grafo direcionado com uma matriz de adjacência não simétrica e com elementos zero na diagonal principal. E por fim, na Figura 4c temos um grafo direcionado com laço com sua matriz de adjacência não simétrica e com um elemento na diagonal principal igual a 1, que diz respeito ao laço.

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

(a) Grafo não direcionado (b) Grafo direcionado (c) Grafo direcionado com laço

Figura 4 – Representação dos grafos através da matriz de adjacência

Existem dois resultados a serem observados sobre a matriz de adjacência de um grafo simples, não direcionado e sem laços: os elementos da diagonal principal da matriz são todos zero e ela é simétrica, pois se houver uma aresta entre $i$ e $j$ então existe uma aresta entre $j$ e $i$.

## 2.2 Redes Aleatórias

Uma rede com $n$ vértices pode ser considerada aleatória quando a existência de arestas entre dois vértices segue alguma distribuição de probabilidade. É possível estudar as entradas da matriz de adjacência $A_{ij}$ do grafo como variáveis aleatórias, sendo a existência das arestas ou os pesos das arestas.

Quando estudamos as entradas sendo a existência das arestas como variáveis aleatórias, o grafo é binário (tem ou não arestas). Quando estudamos as entradas sendo os pesos das arestas como variáveis aleatórias, o grafo é ponderado, pois se houver conexão entre dois vértices então o elemento da matriz seguirá uma distribuição discreta ou contínua, caso contrário a distribuição assumirá valor 0.

## 2.3 Modelo de Erdös-Rényi

Considere **A** a matriz de adjacência de um grafo aleatório simples, não direcionado e sem laços com $n$ sendo o número de vértices. Seja $p \in (0,1)$, $p$ a probabilidade de existir conexão entre os vértices, ou seja, $\mathbb{P}(A_{ij} = 1) = p$, e $1 - p$ a probabilidade de não existir conexão entre os vértices, ou seja, $\mathbb{P}(A_{ij} = 0) = 1 - p$.

Assim, assumindo que as conexões entre os vértices são independentes entre si, a matriz de adjacência do grafo **A**, tem as suas entradas seguindo uma distribuição de Bernoulli: $\mathrm{A}_{ij} \sim \text{Bernoulli}(p)$.

A probabilidade de observar o grafo **A** sendo uma rede **a**, é dada pela probabilidade:

$$\begin{aligned}
\mathbb{P}(\mathbf{A} = \mathbf{a}) &= \prod_{i=1}^{n} \prod_{j>i} \mathbb{P}(\mathrm{A}_{ij} = a_{ij}) \\
&= \prod_{i=1}^{n} \prod_{j>i} p^{a_{ij}} (1-p)^{1-a_{ij}} \\
&= p^{\sum_{i=1}^{n} \sum_{j>i} a_{ij}} (1-p)^{\binom{n}{2} - \sum_{i=1}^{n} \sum_{j>i} a_{ij}}.
\end{aligned}$$

Note que $N = \sum_{i=1}^{n} \sum_{j>i} a_{ij}$ é o número de arestas da rede. Então,

$$\mathbb{P}(\mathbf{A} = \mathbf{a}) = p^{N} (1-p)^{\binom{n}{2} - N}. \tag{2.1}$$

Quando temos uma rede social de pessoas em que as conexões são as relações de amizade, sabendo que o indivíduo A é amigo de B e B é amigo de C, espera-se que a probabilidade de o indíviduo A ser amigo de C seja maior quando sabemos que B é amigo em comum dos dois. Espera-se que a probabilidade de A ser amigo de C seja maior que a probabilidade de A ser amigo de D, quando D não é amigo em comum de A, por exemplo.

Nesse caso, trata-se de uma rede em que as conexões podem ter dependência entre si, e o fato de B é amigo de C pode influenciar o cálculo da probabilidade entre A e C. No modelo de Erdös-Rényi as conexões entre os vértices devem ser independentes entre si, portanto, não seria adequado aplicar este modelo a essa situação.

Da mesma forma, se estamos interessados em estudar dois cursos de graduação: Estatística e Matemática, a probabilidade de as pessoas dentro de seus cursos se conhecerem é diferente da probabilidade de as pessoas entre os cursos de conhecerem. Nesse caso, temos uma rede com dois grupos de indivíduos, gerando assim, uma rede com estrutura de grupos. Uma extensão do modelo de Erdös-Rényi é o modelo estocástico de blocos, que modela como os vértices em uma rede interagem e podem ser organizados em grupos com base em seus padrões de conexões.

## 2.4 Modelo Estocástico de Blocos

Os modelos estocástico de blocos (SBM) são uma classe cada vez mais popular de modelos na análise estatística de grafos. Lee e Wilkinson 2019 propõem descrever a estrutura de comunidades tanto de redes simples quanto em redes complexas. Devido à sua versatilidade, podem ser usados para modelar a estrutura de comunidade (latente) de uma rede, bem como para fins de agrupamento.

Para o modelo estocástico de blocos, consideramos um grafo aleatório simples, não direcionado, sem laços, com $n$ vértices, que neste estudo será aplicado para redes binárias e redes com pesos.

### 2.4.1 Redes Binárias

Este modelo para redes binárias, modela a existência conexão entre os vértices das comunidades, sendo elas dentro da própria comunidade ou entre comunidades.

Para exemplificar esse modelo, considere um grafo de seis vértices, divididos em três comunidades, representados na Figura 5.

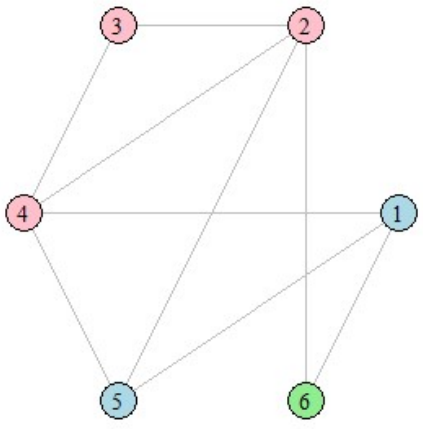



Figura 5 – Grafo de 6 vértices e três comunidades

Considerando a comunidade composta pelos vértices 1 e 5 como comunidade $q$, a comunidade composta pelos vértices 2, 3 e 4 como comunidade $r$ e a comunidade composta pelo

vértice 6 como comunidade $s$, temos três comunidades $q$, $r$ e $s$.

Ainda, é possível descrever de forma vetorial a atribuição de comunidades dos vértices na rede, ou seja, descrever a informação sobre cada vértice estar associado a uma comunidade. Desse modo, temos um vetor $Z_i$, de dimensão igual ao número de comunidades, associado a cada vértice $i$ do grafo, em que cada entrada do vetor representará uma comunidade, atribuindo valores 0 caso o vértice não pertença à comunidade e 1 caso pertença à comunidade:

- $Z_1 = (1,0,0)$, o vértice 1 está na comunidade $q$;
- $Z_2 = (0,1,0)$, o vértice 2 está na comunidade $r$;
- $Z_3 = (0,1,0)$, o vértice 3 está na comunidade $r$;
- $Z_4 = (0,1,0)$, o vértice 4 está na comunidade $r$;
- $Z_5 = (1,0,0)$, o vértice 5 está na comunidade $q$;
- $Z_6 = (0,0,1)$, o vértice 6 está na comunidade $s$.

Escrevendo $\mathbf{Z}$, como a matriz de combinação de $Z_1$, $Z_2$, $Z_3$, $Z_4$, $Z_5$ e $Z_6$:

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

A quantidade de vértices em de cada comunidade, será o somatório de cada coluna da matriz $\mathbf{Z}$. Assim, a quantidade de vértices de cada comunidade é $N_q = 2$, $N_r = 3$ e $N_s = 1$. O número de vértices em cada comunidade pode ser definido como o vetor $\mathbf{N} = (N_q, N_r, N_s)$.

Formalizando e generalizando os conceitos acima descritos para o exemplo em questão, considere $K$ o número de comunidades em uma rede $\mathbf{A}$ com $n$ vértices, temos que $\mathbf{Z}_i$ é o vetor de atribuição de comunidades com entradas iguais a zero e apenas a entrada associada a respectiva comunidade é igual a 1, para $i = 1,2,\ldots,n$. A matriz de combinação dos vetores que descrevem a informação sobre cada vértice pertencer a cada comunidade é dada por $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)^T$ com dimensão $n \times K$.

Com base nessa construção, temos que $Z_{il} = 1$ quando o vértice $i$ pertence a comunidade $l$ e $Z_{il} = 0$ quando o vértice $i$ não pertence a comunidade $l$.

Para encontrar a quantidade de vértices de cada comunidade de acordo com a matriz $\mathbf{Z}$, basta somar, para cada coluna, as entradas das linhas de $\mathbf{Z}$. Desse modo, o número de vértices na comunidade $l$, será denotado da seguinte maneira:

$$N_l = \sum_{i=1}^{n} \mathbb{1}\{Z_{il} = 1\}. \tag{2.2}$$

Seja $E_{lm}$ o número de arestas entre vértices da comunidade $l$ e $m$, onde $1 \leq l, m \leq K$, que é dado por:

$$E_{lm} = \sum_{1 \leq i < j \leq n} a_{ij}\, \mathbb{1}\{Z_{il} = 1, Z_{jm} = 1\} \tag{2.3}$$

Seja $\mathbf{C}$ uma matriz de probabilidade de conexão entre comunidades de dimensão $K \times K$, em que $C_{lm}$ representa a probabilidade de existir uma aresta entre vértices da comunidade $l$ e $m$. Por exemplo, para $K = 2$ temos:

$$\mathbf{C} = \begin{pmatrix} C_{ll} & C_{lm} \\ C_{ml} & C_{mm} \end{pmatrix}.$$

Note que $\mathbf{C}$ é simétrica, pois a probabilidade de ter arestas entre a comunidade $l$ e $m$ é a mesma probabilidade de ter arestas entre a comunidade $m$ e $l$. Ainda, as linhas de $\mathbf{C}$, não necessariamente somam um, pois as probabilidades de conexões entre comunidades e dentro das comunidades são independentes entre si.

No modelo SBM, as comunidades são atribuídas a cada vértice de maneira independente. Assim, quando $K$ é conhecido, é possível calcular a probabilidade de um certo $\mathbf{Z}_i$ estar em uma comunidade $K$, isto é, $\mathbb{P}(Z_{i1} = 1) = \pi_1$, $\mathbb{P}(Z_{i2} = 1) = \pi_2$, ..., $\mathbb{P}(Z_{iK} = 1) = \pi_K$.

Desse modo, como $\mathbf{Z}_i$ é um vetor de zeros com um único valor um, e com as probabilidades $(\pi_i)_{1 \leq i \leq K}$ descritas acima, temos que $\mathbf{Z}_i$ tem distribuição Multinomial com parâmetros 1 e $(\pi_1, ..., \pi_k)$. Portanto, temos o modelo:

$$\mathbf{Z}_i \sim \mathscr{M}(1; (\pi_1, ..., \pi_k)),$$

em que $\pi_1 + ... + \pi_K = 1$ e $(\mathbf{Z}_i)_{1 \leq i \leq n}$ são independentes.

Condicionado a $\mathbf{Z}$, as arestas $(A_{ij})_{1 \leq i \leq j \leq n}$ são variáveis aleatórias independentes, como $\mathbf{A}$ é um grafo binário, segue que:

$$A_{ij} \mid Z_{il} Z_{jm} = 1 \sim \text{Bernoulli}(C_{lm}).$$

### 2.4.2 Modelagem e verossimilhança

Como temos a seguinte condição $A_{ij} \mid Z_{il} Z_{jm} = 1 \sim \text{Bernoulli}(C_{lm})$, a probabilidade é dada por:

$$\mathbb{P}(A_{ij} = a_{ij} \mid Z_{il} Z_{jm} = 1) = C_{lm}^{a_{ij}} (1 - C_{lm})^{1 - a_{ij}}, \quad a_{ij} \in \{0, 1\}.$$

Deste modo, calculando a probabilidade condicional de se observar o grafo $\mathbf{a}$ dada as comunidades $\mathbf{z}$:

$$\begin{aligned}\mathbb{P}(\mathbf{A}=\mathbf{a}\mid\mathbf{Z}=\mathbf{z}) &= \prod_{1\le i<j\le n}\mathbb{P}(A_{ij}=a_{ij}\mid\mathbf{Z}=\mathbf{z})\\ &= \prod_{1\le i<j\le n}\prod_{l,m=1}^{K} C_{lm}^{a_{ij}\mathbb{1}\{z_{il}=1,z_{jm}=1\}}(1-C_{ab})^{(1-a_{ij})\mathbb{1}\{z_{il}=1,z_{jm}=1\}}\end{aligned}$$

$$\begin{aligned}&= \prod_{l,m=1}^{K} C_{lm}^{\sum_{1\le i<j\le n} a_{ij}\,\mathbb{1}\{z_{il}=1,z_{jm}=1\}}\\ &\quad\cdot(1-C_{lm})^{\sum_{1\le i<j\le n}(1-a_{ij})\,\mathbb{1}\{z_{il}=1,z_{jm}=1\}}.\end{aligned}$$

Note que o número de vértices na comunidade $l$ e na comunidade $m$ é dado por:

$$N_{lm} = \sum_{1\le i<j\le n}\mathbb{1}\{z_{il}=1,z_{jm}=1\}. \tag{2.4}$$

Por (2.4) e (2.3), temos:

$$\begin{aligned}\sum_{1\le i<j\le n}(1-a_{ij})\,\mathbb{1}\{z_{il}=1,z_{jm}=1\} &= \sum_{1\le i<j\le n}\mathbb{1}\{z_{il}=1,z_{jm}=1\}\\ &\quad-\sum_{1\le i<j\le n}a_{ij}\,\mathbb{1}\{z_{il}=1,z_{jm}=1\}\\ &= N_{lm}-E_{lm}.\end{aligned}$$

Veja que $N_{lm}-E_{lm}$ representa o número de não arestas entre $l$ e $m$. Voltando a $\mathbb{P}(\mathbf{A}=\mathbf{a}\mid\mathbf{Z}=\mathbf{z})$:

$$\begin{aligned}\mathbb{P}(\mathbf{A}=\mathbf{a}\mid\mathbf{Z}=\mathbf{z}) &= \prod_{l,m=1}^{K} C_{lm}^{\sum_{1\le i<j\le n} a_{ij}\,\mathbb{1}\{z_{il}=1,z_{jm}=1\}}\\ &\quad\cdot(1-C_{lm})^{\sum_{1\le i<j\le n}(1-a_{ij})\,\mathbb{1}\{z_{il}=1,z_{jm}=1\}}\\ &= \prod_{l,m=1}^{K} C_{lm}^{E_{lm}}(1-C_{lm})^{N_{lm}-E_{lm}}.\end{aligned}$$

Assim, a distribuição do grafo $\mathbf{A}$ dada as comunidades $\mathbf{Z}$:

$$\mathbb{P}(\mathbf{A}=\mathbf{a}\mid\mathbf{Z}=\mathbf{z}) = \prod_{l,m=1}^{K} C_{lm}^{E_{lm}}(1-C_{lm})^{N_{lm}-E_{lm}}. \tag{2.5}$$

Por $\mathbf{Z}_i$ seguir uma distribuição multinomial, ou seja, $\mathbf{Z}_i \sim \mathscr{M}(1;(\pi_1,...,\pi_k))$, é possível calcular a probabilidade das comunidades:

$$\begin{aligned}\mathbb{P}(\mathbf{Z}=\mathbf{z}) &= \prod_{i=1}^{n}\mathbb{P}(Z_i=z_i)\\ &= \prod_{i=1}^{n}\prod_{l=1}^{K}\pi_l^{\mathbb{1}\{z_{il}=1\}}\\ &= \prod_{l=1}^{K}\pi_l^{\sum_{i=1}^{n}\mathbb{1}\{z_{il}=1\}}\\ &= \prod_{l=1}^{K}\pi_l^{N_l},\end{aligned}$$

pela definição (2.2). Assim,

$$\mathbb{P}(\mathbf{Z}=\mathbf{z}) = \prod_{l=1}^{K}\pi_l^{N_l}. \tag{2.6}$$

A probabilidade de se observar o grafo $\mathbf{a}$ é dada pelas equações (2.5) e (2.6):

$$\begin{aligned}\mathbb{P}(\mathbf{A}=\mathbf{a}) &= \sum_{z\in Z}\mathbb{P}(\mathbf{A}=\mathbf{a}\mid\mathbf{Z}=\mathbf{z})\mathbb{P}(\mathbf{Z}=\mathbf{z})\\ &= \sum_{z\in Z}\left[\prod_{l,m=1}^{K}C_{lm}^{E_{lm}}(1-C_{lm})^{N_{lm}-E_{lm}}\prod_{l=1}^{K}\pi_l^{N_l}\right].\end{aligned}$$

No entanto, em aplicações para dados reais não são fornecidas algumas informações, como as associações dos grupos $\mathbf{Z}$, a matriz de probabilidade $\mathbf{C}$ e o vetor $(\pi_1,...,\pi_k)$. Portanto, o objetivo de ajustar um modelo estocástico de blocos a um grafo é inferir esses dois parâmetros simultaneamente. Desta maneira, surgem os desafios estatísticos usuais:

1. Inferência: Uma vez que a probabilidade pode ser calculada, como devemos inferir os membros do grupo e a matriz de blocos? Existem algoritmos de inferência eficientes e escaláveis?

2. Devemos incorporar $K$ como parâmetro do modelo e utilizar na inferência? Ou devemos utilizar um SBM com diferentes $K's$ e encontrar o melhor $K$ como um problema de seleção de modelo?

É possível determinar o valor de $K$ através do teste de hipóteses de Lei (2016) para redes sem pesos (binárias). Mais detalhes desse teste será visto na Seção 3.3.

### 2.4.3 Exemplos de redes reais

A seguir temos duas redes reais com comunidades nas quais pode ser aplicado o Modelo Estocástico de Blocos, em que são fornecidas as associações dos vértices às comunidades e, principalmente, a quantidade de comunidades das respectivas redes.

A primeira é uma rede social do artigo Zachary (1977) entre membros do clube de karate de uma universidade, em que John e Mr. Hi eram sócios e se separaram após um desentendimento. Por este motivo, cada um dos sócios abriu seu clube de karate e um dos estudantes decidiu analisar se ainda existiam relações entre as pessoas dos dois clubes.

Deste modo, temos 34 vértices e 78 arestas, assim na Figura 6, as pessoas são separadas por cores representando a participação, respectivamente, em um dos dois clubes de karate e as conexões entre os vértices representam as atividades acadêmicas em comum.

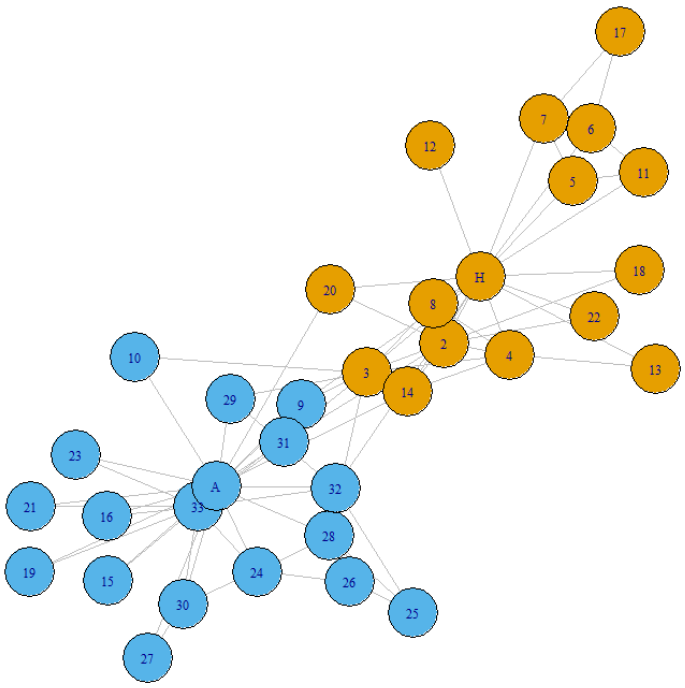



Figura 6 – Grafo com comunidades do clube de karate

A segunda é uma rede de amizade entre membros docentes de uma universidade no Reino Unido, do artigo Nepusz *et al.* (2008), que possui 81 vértices e 577 arestas. A afiliação escolar de cada indivíduo é armazenada como um atributo do vértice, para designar em qual comunidade está, como mostra na Figura 7:

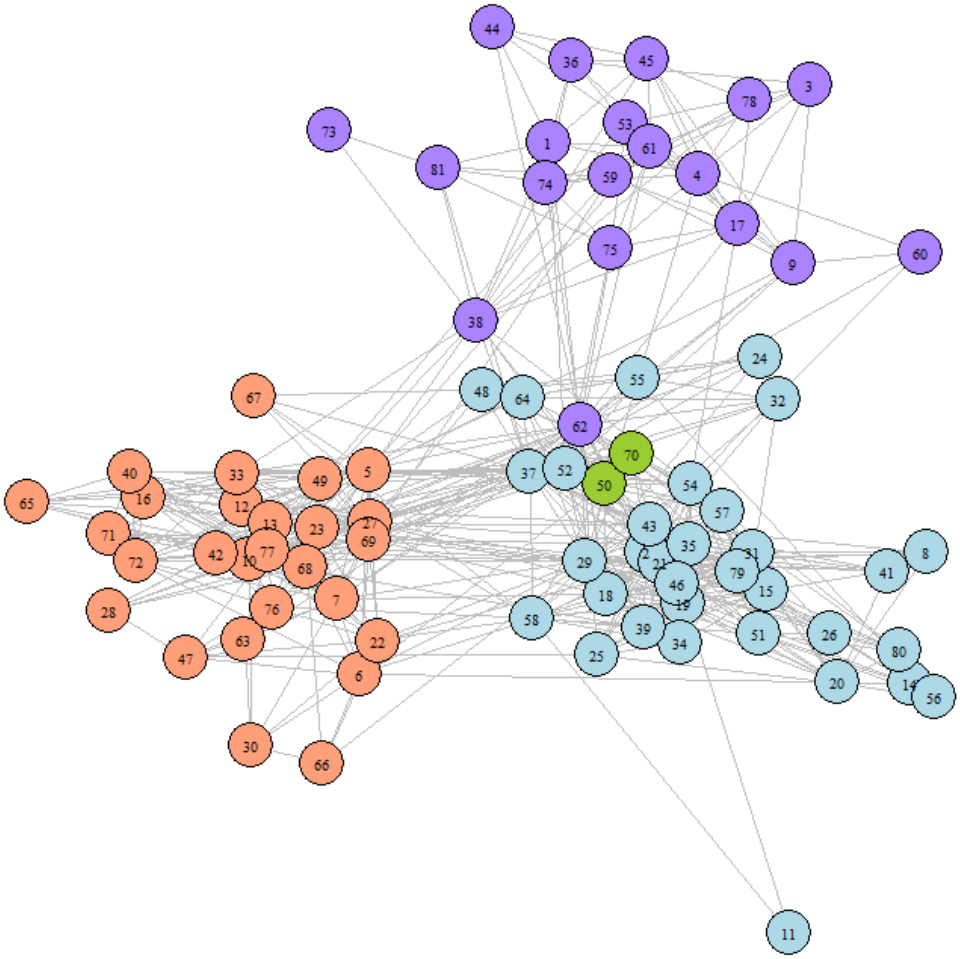



Figura 7 – Grafo com comunidades de docentes em uma universidade

### 2.4.4 Redes ponderadas

Uma rede com pesos é uma estrutura na qual as conexões entre os vértices não se limitam à presença ou ausência de conexão, como ocorre em uma rede binária, mas são associadas a valores numéricos (discretos ou contínuos) que representam a intensidade da conexão.

Para modelar pesos discretos, podemos assumir a distribuição de Poisson para os pesos entre os vértices, então $A_{ij}$ assumirá valores inteiros não negativos. Consideramos grafos aleatórios simples, não direcionados e sem laços, então $\mathbf{A}$ é uma matriz com diagonal de elementos zero e simétrica.

O Modelo Estocástico de Blocos para grafos ponderados estuda a existência de arestas com pesos entre os vértices das comunidades e utilizamos alguns resultados da Seção 2.4.1.

Seja $\theta$ uma matriz de pesos esperados nas arestas de vértices entre comunidades de dimensão $K \times K$, em que $\theta_{lm}$ representa o valor esperado de uma aresta com peso entre vértices da comunidade $l$ e $m$.

Desse modo, as atribuições dos vértices as comunidades será representada pelo vetor $\mathbf{Z}_i$.

A probabilidade de um certo vértice pertencer a comunidade $K$ será o vetor $\pi$. Segue que $\mathbf{Z}_i$ tem distribuição Multinomial com parâmetros 1 e $(\pi_1,...,\pi_k)$. Portanto, temos o modelo:

$$\mathbf{Z}_i \sim \mathscr{M}(1;(\pi_1,...,\pi_k)),$$

em que $\pi_1 + ... + \pi_K = 1$ e $(\mathbf{Z}_i)_{1\leq i\leq n}$ são independentes.

Condicionado a $\mathbf{Z}$, as arestas $(A_{ij})_{1\leq i\leq j\leq n}$ são variáveis aleatórias independentes, como $\mathbf{A}$ é um grafo ponderado, segue que:

$$A_{ij} \mid Z_{il}Z_{jm} = 1 \sim \text{Poisson}(\theta_{lm}).$$

### 2.4.5 Modelagem e verossimilhança

Como temos a seguinte que $A_{ij} \mid Z_{il}Z_{jm} = 1 \sim \text{Poisson}(\theta_{lm})$, a probabilidade é dada por:

$$\mathbb{P}(A_{ij} = a_{ij} \mid Z_{il}Z_{jm} = 1) = \frac{e^{-\theta_{lm}}\theta_{lm}^{a_{ij}}}{a_{ij}!}, \quad a_{ij} \in \{0,1,2,\ldots\}.$$

Deste modo, calculando a probabilidade condicional de se observar o grafo $\mathbf{a}$ dada as comunidades $\mathbf{z}$:

$$\begin{aligned}
\mathbb{P}(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = \mathbf{z}) &= \prod_{1\leq i<j\leq n} \mathbb{P}(A_{ij} = a_{ij} \mid \mathbf{Z} = \mathbf{z}) \\
&= \prod_{1\leq i<j\leq n} \prod_{l,m=1}^{K} \frac{e^{-\theta_{lm}\mathbb{1}\{z_{il}=1,z_{jm}=1\}}\theta_{lm}^{a_{ij}\mathbb{1}\{z_{il}=1,z_{jm}=1\}}}{a_{ij}!} \\
&= \frac{\prod_{l,m=1}^{K} e^{-\theta_{lm}\sum_{1\leq i<j\leq n}\mathbb{1}\{z_{il}=1,z_{jm}=1\}} \theta_{lm}^{\sum_{1\leq i<j\leq n} a_{ij}\mathbb{1}\{z_{il}=1,z_{jm}=1\}}}{\prod_{1\leq i<j\leq n}\prod_{l,m=1}^{K} a_{ij}!}.
\end{aligned}$$

Por (2.4) e (2.3), a distribuição do grafo $\mathbf{A}$ dada as comunidades $\mathbf{Z}$:

$$\mathbb{P}(\mathbf{A} = \mathbf{a} \mid \mathbf{Z} = \mathbf{z}) = \frac{\prod_{l,m=1}^{K} e^{-\theta_{lm}N_{lm}}\theta_{lm}^{E_{lm}}}{\prod_{1\leq i<j\leq n}\prod_{l,m=1}^{K} a_{ij}!}. \tag{2.7}$$

A probabilidade das comunidades $\mathbf{Z}_i$ é calculada da mesma forma que em redes binárias em (2.6).

CAPÍTULO

# 3

# TESTE DE HIPÓTESES PARA REDES BINÁRIAS

Neste capítulo, abordaremos em detalhes o teste de hipóteses proposto por Lei (2016), cujo objetivo é estimar o número de comunidades em uma rede binária por meio do Modelo Estocástico de Blocos. As seguintes seções apresentam sobre Estimador de Máxima Verossimilhança, resultados para matrizes aleatórias, o teste de hipóteses e sua versão ajustada utilizando a técnica de bootstrap.

## 3.1 Estimador de Máxima Verossimilhança

No Capítulo 2.4 apresentamos o Modelo Estocástico de Blocos dado as atribuições dos vértices as comunidades e o número de comunidades. No entanto, em aplicações para dados reais não são fornecidas algumas informações, como as associações dos grupos $\mathbf{Z}$ e a matriz de probabilidade entre os grupos $\mathbf{C}$.

Assim, para determinar o número de comunidades $K$, é possível através do teste de hipóteses de Lei (2016) estimar este valor. Se fosse dado o valor das associações dos grupos $\mathbf{Z}$ e a matriz de probabilidade $\mathbf{C}$, encontraríamos o valor de $K$ através do teste de hipóteses. Entretanto, não temos esses valores, então é necessário estimar os parâmetros, ou seja, $\hat{\mathbf{C}}$ e $\hat{\mathbf{Z}}$.

Para isto, será calculado o estimador de máxima verossimilhança através da distribuição de probabilidade da Seção 2.4.2.

A distribuição de probabilidade do grafo $\mathbf{A}$, condicionado as comunidades $\mathbf{Z}$, é dada através da equação (2.5):

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{C}) = \prod_{l,m=1}^{K} C_{lm}^{E_{lm}} (1 - C_{lm})^{N_{lm} - E_{lm}}.$$

Seja a função de verossimilhança:

$$\mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A}) = \prod_{l,m=1}^{K} C_{lm}^{E_{lm}}(1-C_{lm})^{N_{lm}-E_{lm}}. \tag{3.1}$$

Então, verificando se este é o estimador de máxima verossimilhança, temos:

i. Calculando o $\ln \mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A})$:

$$\begin{aligned}
\ln \mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A}) &= \ln\left(\prod_{l,m=1}^{K} C_{lm}^{E_{lm}}(1-C_{lm})^{N_{lm}-E_{lm}}\right) \\
&= \sum_{l,m=1}^{K} \ln C_{lm}^{E_{lm}} + \ln(1-C_{lm})^{N_{lm}-E_{lm}} \\
&= \sum_{l,m=1}^{K} E_{lm}\cdot \ln C_{lm} + (N_{lm}-E_{lm})\cdot \ln(1-C_{lm}).
\end{aligned}$$

ii. Derivando $\mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A})$ em relação a $C_{lm}$:

$$\begin{aligned}
\frac{\partial \mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A})}{\partial C_{lm}} &= \frac{\partial}{\partial C_{lm}}\left[\sum_{l,m=1}^{K} E_{lm}\cdot \ln C_{lm} + (N_{lm}-E_{lm})\cdot \ln(1-C_{lm})\right] \\
&= E_{lm}\cdot \frac{1}{C_{lm}} + (E_{lm}-N_{lm})\cdot \frac{1}{1-C_{lm}}.
\end{aligned}$$

iii. Resolvendo $\dfrac{\partial \mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A})}{\partial C_{lm}} = 0$:

$$\begin{gathered}
\frac{\partial \mathscr{L}(\mathbf{Z},\mathbf{C} \mid \mathbf{A})}{\partial C_{lm}} = 0 \\
E_{lm}\cdot \frac{1}{C_{lm}} + (E_{lm}-N_{lm})\cdot \frac{1}{1-C_{lm}} = 0 \\
\frac{E_{lm}}{C_{lm}} = \frac{(N_{lm}-E_{lm})}{1-C_{lm}} \\
E_{lm} - E_{lm}C_{lm} = N_{lm}C_{lm} - E_{lm}C_{lm} \\
E_{lm} = N_{lm}C_{lm} \\
\frac{E_{lm}}{N_{lm}} = C_{lm}
\end{gathered}$$

Logo, maximizando a log-verossimilhança com base em $\hat{z}$ que são as associações dos vértices às comunidades, podemos estimar as probabilidades $C_{lm}$:

$$\hat{C}_{lm} = \frac{\hat{E}_{lm}}{\hat{N}_{lm}}, \tag{3.2}$$

onde o número de conexões entre nós da comunidade $l$ e $m$, $\hat{E}_{lm}$, é obtido através da Equação (2.3) e é calculado usando as comunidades estimadas $\hat{z}$. Da mesma forma, $\hat{N}_{lm}$ é obtido usando as comunidades estimadas $\hat{\mathbf{Z}}$.

Como $\hat{\mathbf{Z}}$ são as comunidades estimadas, $\hat{N}_l$ é o número estimado de vértices na comunidade $l$, então o valor estimado para $\mathbf{C}$ será:

$$\hat{C}_{lm} = \begin{cases} \dfrac{\sum_{i<j}^{n} A_{ij} \mathbb{1}\{\hat{z}_{il} = 1, \hat{z}_{jm} = 1\}}{\hat{N}_l \hat{N}_m}, & l \neq m, \\ \dfrac{\sum_{i<j}^{n} A_{ij} \mathbb{1}\{\hat{z}_{il} = 1, \hat{z}_{jm} = 1\}}{\hat{N}_l (\hat{N}_l - 1)/2}, & l = m, \end{cases} \tag{3.3}$$

Para compreender o teste de hipóteses, tornou-se necessário estudar outros tópicos que o fundamentam, como os resultados assintóticos da distribuição do menor e maior autovalor que é obtida usando avanços recentes na teoria de matrizes aleatórias e apresentaremos nas próximas seções.

## 3.2 Resultados para matrizes aleatórias

### 3.2.1 Generalização da matriz Wigner

Seja uma matriz aleatória $\mathbf{A}$ de dimensão $n \times n$, cujas as entradas $A_{ij}$ com $i \leq j$ são variáveis aleatórias independentes. Assumindo a centralidade nas variáveis aleatórias, assume-se que essas variáveis tem média igual a zero, ou seja, $E(A_{ij}) = 0$.

Sabemos que a variância de uma variável aleatória pode ser escrita como:

$$\begin{aligned} Var(A_{ij}) &= E(A_{ij}^2) - E^2(A_{ij}) \\ &= E(A_{ij}^2). \end{aligned}$$

Para uma constante $H > 0$, suponha que a seguinte condição seja satisfeita:

$$H^{-1} \leq nVar(A_{ij}) \leq H, \qquad \sum_j Var(A_{ij}) = 1. \tag{3.4}$$

Uma matriz aleatória que satisfaz as condições do valor esperado das entradas serem iguais a zero e a soma das variâncias das entradas serem um, será chamada de matriz generalizada de Wigner.

### 3.2.2 Distribuição de Tracy-Widom

No estudo Tracy e Widom (2002) foi provado resultados teóricos de que os autovalores de matrizes aleatórias convergem para a Distribuição Tracy-Widom, no qual foram utilizadas implicitamente em Lei (2016).

Considere $\lambda_n$ o maior autovalor de uma matriz **A** de Wigner. Assim, a função densidade de $\lambda_n$ é representada pela Tracy-Widom (TW):

$$\lim_{n\to\infty} \mathbb{P}(n^{2/3}(\lambda_n - 2) \leq s) = F_\beta(s),$$

onde a função de distribuição acumulada de TW de $F_\beta$ pode ser descrita pela equação de Painlevé com $\beta = 1, 2$ e 4, correspondentes a ortogonal, unitário e conjunto de modelos.

A distribuição de Tracy-Widom, diferentemente de outras distribuições de probabilidade, não possui fórmula explícita já que sua distribuição depende da solução da equação de Painlevé (equações não lineares diferenciais de segunda ordem). Essa derivação que tem um carácter fortemente matemático foge do escopo deste trabalho, porém sua distribuição pode ser obtida no R através da função “dtw"no pacote “RMTstat".

O comportamento das distribuições de $TW_1$, $TW_2$ e $TW_4$ são influenciados através dos valores de $\beta = 1, 2$ e 4 respectivamente, como na Figura 8:

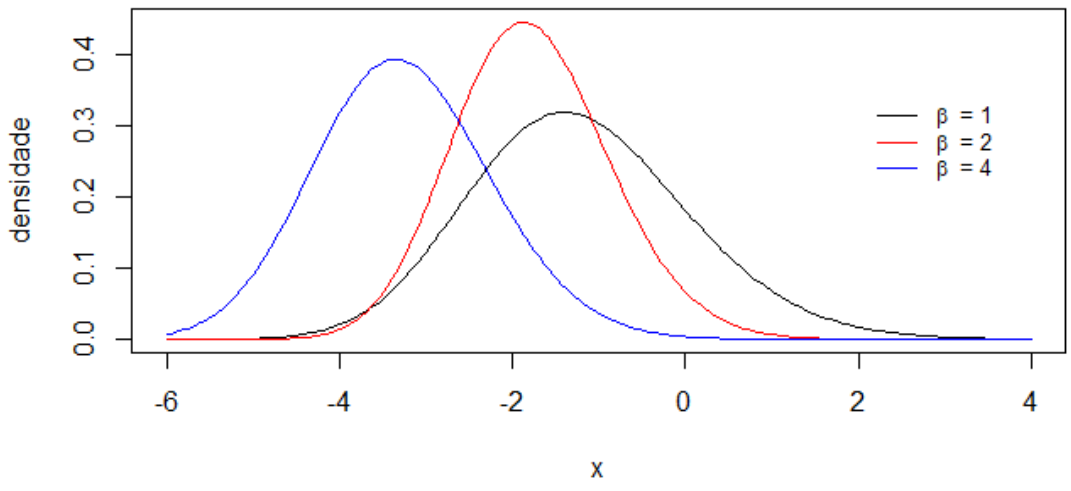



Figura 8 – Gráfico de densidade das distribuições Tracy-Widom

## 3.3 Teste de hipóteses

Nesta seção apresentaremos em detalhes um teste de hipóteses proposto por Lei (2016), no qual a proposta é estimar o número de comunidades da rede do Modelo estocástico de blocos,

para redes binárias.

Para estimar o número de comunidades, consideramos $K$ o número real de comunidades em um modelo estocástico de blocos e $K_0$ para denotar um número hipotético de comunidades, assim definimos o seguinte teste de hipóteses:

$$\begin{cases} H_{0,K_0}: & K = K_0 \\ H_{1,K_0}: & K > K_0, \end{cases} \tag{3.5}$$

sequencialmente para cada $K_0 \geq 1$ até que a hipótese nula não seja rejeitada. Quando a hipótese nula não é rejeitada, significa que o grafo possui o número de comunidades igual a $K_0$.

Considere uma matriz $\mathbf{P}$ de dimensão $n \times n$, no qual as entradas $P_{ij}$ representam a probabilidade de conexão entre os vértices $i$ e $j$ do grafo. Defina $g_i$ como a comunidade que o vértice $i$ pertence, ou seja, $g_i = a$ se, e somente se, $Z_{ia} = 1$, para $a = 1, \ldots, K$. Denotaremos como:

$$P_{ij} = C_{g_i g_j}, \tag{3.6}$$

para $i, j = 1, \ldots, n$, de modo que $E(\mathbf{A}) = \mathbf{P} - diag(\mathbf{P})$. Defina $\tilde{\mathbf{A}}^*$ como:

$$\tilde{A}^*_{ij} = \frac{A_{ij} - P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}},\ i \neq j \quad \text{e} \quad \tilde{A}^*_{ii} = 0,\ \forall i. \tag{3.7}$$

Sabemos que no modelo estocástico de blocos para redes binárias cada entrada da matriz de adjacência segue uma distribuição de Bernoulli, de modo que $E(A_{ij}) = P_{ij}$ e $Var(A_{ij}) = P_{ij}(1 - P_{ij})$.

Para utilizar resultados sobre o espectro de grafos, é necessário verificar se a matriz $\tilde{\mathbf{A}}^*$ é uma matriz generalizada de Wigner. Primeiramente verificaremos que o valor esperado das entradas da matriz $\tilde{\mathbf{A}}^*$ são iguais a zero:

$$
\begin{aligned}
E(\tilde{A}^*_{ij}) &= E\left(\frac{A_{ij}-P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= E\left(\frac{A_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}} - \frac{P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= E\left(\frac{A_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right) - \left(\frac{P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= \frac{1}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\cdot E(A_{ij}) - \left(\frac{P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= \frac{1}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\cdot (E(A_{ij})-P_{ij})\\
&= \frac{1}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\cdot (P_{ij}-P_{ij})\\
&= 0.
\end{aligned}
$$

Agora verificaremos que o somatório da variância das entradas da matriz $\tilde{\mathbf{A}}^*$ será igual a um. Calculando a $Var(\tilde{A}^*_{ij})$:

$$
\begin{aligned}
Var(\tilde{A}^*_{ij}) &= Var\left(\frac{A_{ij}-P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= Var\left(\frac{A_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}\right)\\
&= \frac{1}{(n-1)P_{ij}(1-P_{ij})}\cdot Var(A_{ij})\\
&= \frac{1}{(n-1)P_{ij}(1-P_{ij})}\cdot P_{ij}\cdot(1-P_{ij})\\
&= \frac{1}{n-1}.
\end{aligned}
$$

Assim, como $\tilde{\mathbf{A}}^*$ é uma matriz simétrica, temos que $Var(\tilde{A}^*_{ii}) = 0$. Logo:

$$
\sum_{j=1}^{n} Var(\tilde{A}^*_{ij}) = \sum_{j\neq i}^{n} Var(\tilde{A}^*_{ij}) + 0 = \sum_{j\neq i}^{n} \frac{1}{n-1} = \frac{n-1}{n-1} = 1.
$$

Note que $E(\tilde{A}^*_{ij}) = 0$ para todo $(i,j)$ e $\sum_{j\neq i} Var(\tilde{A}^*_{ij}) = 1$. Uma distribuição assintótica dos extremos autovalores de $\tilde{A}^*_{ij}$ tem sido estudado na teoria de matriz aleatórias, em particular combinando os resultados de Erdős, Yau e Yin (2012) e Lee e Yin (2014) temos:

$$
n^{2/3}[-\lambda_1(\tilde{A}^*_{ij}) - 2] \rightsquigarrow TW_1 \;\text{ e }\; n^{2/3}[\lambda_n(\tilde{A}^*_{ij}) - 2] \rightsquigarrow TW_1, \tag{3.8}
$$

onde $\lambda_1(\tilde{A}^*_{ij})$ é o menor autovalor de $\tilde{A}^*_{ij}$ e $\lambda_n(\tilde{A}^*_{ij})$ é o maior autovalor de $\tilde{A}^*_{ij}$. Denotamos $TW_1$ como a distribuição de Tracy-Widom com índice 1 e "$\rightsquigarrow$" para convergência da distribuição.

Note que $\mathbf{P}$ depende da matriz $\mathbf{C}$ que não é conhecida. Logo, ao invés de usarmos a matriz $\mathbf{C}$, usamos o estimador de $\mathbf{C}$, denotado por $\hat{C}$, dado na Equação 3.3. Como $\mathbf{C}$ depende de $g$ que não é conhecido, usamos a estimativa de $g$ para estimar $\mathbf{C}$, para assim, estimar $\mathbf{P}$.

Utilizando os estimadores, definimos a matriz $\tilde{\mathbf{A}}$:

$$\tilde{A}_{ij} = \begin{cases} \frac{A_{ij}-\hat{P}_{ij}}{\sqrt{(n-1)\hat{P}_{ij}(1-\hat{P}_{ij})}}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{3.9}$$

onde $\hat{P}_{ij} = \hat{C}_{\hat{g}_i\hat{g}_j}$.

Sob a hipótese nula e quando as estimativas $(\hat{g}, \hat{C})$ são precisas o suficiente, a convergência da relação (3.8) mostrará uma correspondência dos autovalores de $\tilde{\mathbf{A}}^*$, ou seja, o erro das estimativas será próximo de zero.

Para obter o maior intervalo na estatística do teste, Lei (2016) utiliza a correção de Bonferroni que é o máximo em módulo entre o menor e maior autovalor das estimativas na matriz $\tilde{\mathbf{A}}$:

$$\sigma_1(\tilde{A}) = \max(-\lambda_1(\tilde{\mathbf{A}}), \lambda_n(\tilde{\mathbf{A}})). \tag{3.10}$$

Utilizando os resultados dos maiores autovalores de matrizes aleatórias visto na Seção 3.2.2, temos que a estatística do teste é dada por:

$$T_{n,K_0} = n^{2/3}[\sigma_1(\tilde{\mathbf{A}}) - 2]. \tag{3.11}$$

Logo, quando combinamos as duas equações (3.11) e (3.10), resultam que a estatística do teste será:

$$T_{n,K_0} = \max[n^{2/3}(-\lambda_1(\tilde{\mathbf{A}}) - 2), n^{2/3}(\lambda_n(\tilde{\mathbf{A}}) - 2)]. \tag{3.12}$$

O nível de rejeição $\alpha$ para o problema do teste na equação (3.5) será:

$$\text{Rejeita-se } H_{0,K_0}, \text{ se } T_{n,K_0} \geq t_{\alpha/2}, \tag{3.13}$$

onde $t_{\alpha/2}$ é o quantil superior $\alpha/2$ da distribuição $TW_1$ para $\alpha \in (0,1)$.

Para o problema do teste, temos o teste sequencial para estimar o valor de $K$:

$$\hat{K} = \inf\{K_0 \geq 1 : T_{n,K_0} < t_{\alpha/2}\},$$

onde $T_{n,k_0} < t_{\alpha/2}$ seria a região crítica. É realizado o teste para $K_0 = 1, 2, ...$, até $H_0$ não seja rejeitada.

Para verificar a consistência do estimador $\hat{K}$, utilizamos dois resultados de Lei (2016) que consideram que se o estimador $\hat{\mathbf{Z}}$ das comunidades é consistente, ou seja, para uma rede $A$ com comunidades dada por $\mathbf{z}$ temos que $\mathbb{P}(\hat{\mathbf{Z}} = \mathbf{z})$ converge para 1 quando o número de vértices da rede cresce $(n \to \infty)$. Neste caso em que $\hat{Z}$ é consistente, Lei (2016) provou que a distribuição assintótica convergirá a uma Tracy-Widom:

$$n^{2/3}[-\lambda_1(\tilde{A}_{ij}) - 2] \rightsquigarrow TW_1 \text{ e } n^{2/3}[\lambda_n(\tilde{A}_{ij}) - 2] \rightsquigarrow TW_1. \tag{3.14}$$

Ainda, se $\hat{\mathbf{Z}}$ é consistente, então Lei (2016) prova em seu artigo o Teorema 3.4, que o estimador $\hat{K}$ é consistente, ou seja,

$$\mathbb{P}(\hat{K} = K) \xrightarrow{n\to\infty} 1.$$

Portanto, utilizar um estimador $\hat{\mathbf{Z}}$ consistente resulta que $\hat{K}$ se aproxime do $K$ verdadeiro, quando o número de vértices aumenta.

## 3.4 Teste de hipóteses com bootstrap

A construção do teste de hipóteses com bootstrap é feita seguindo os mesmos passos que foram descritos na Seção 3.3, entretanto a proposta de Lei (2016) é realizar um ajuste na estimativa do teste, utilizando a média e a variância dos extremos autovalores da matriz aleatória. Essa correção de Bootstrap é feita para garantir que, sob a hipótese nula, a distribuição da estatística do teste se aproxime da distribuição Tracy-Widom mesmo para redes com poucos vértices.

Dada a matriz de adjacência $\mathbf{A}$ com $n$ vértices, sob a hipótese nula de que $K = K_0$, a estatística teste com a correção do bootstrap é apresentado nas seguintes etapas:

1. Estimar as associações dos vértices com as comunidades estimadas, que corresponde a estimar $\hat{\mathbf{C}}$ e $\hat{\mathbf{P}}$ que corresponde as estimativas da equação (3.2).
2. Calcular $\tilde{\mathbf{A}}$ da equação (3.9) , como também o menor e maior autovalor de $\tilde{\mathbf{A}}$.
3. Seja $(\mu_1, \hat{s}_1^2)$ e $(\mu_n, \hat{s}_n^2)$ a média e a variância respectivamente de $\lambda_1(\tilde{\mathbf{A}})$ e $\lambda_n(\tilde{\mathbf{A}})$.
4. A correção através do bootstrap na estatística do teste será:

$$T_{n,K_0}^{(boot)} = \mu_{TW} + s_{TW} \max\left(-\frac{\lambda_1(\tilde{\mathbf{A}})}{\hat{s}_1}, \frac{\lambda_n(\tilde{\mathbf{A}})}{\hat{s}_n}\right), \tag{3.15}$$

onde $\mu_{TW}$ e $s_{TW}$ são a média e a desvio padrão da distribuição Tracy-Widom.

As simulações realizadas por Lei (2016), apresentam a densidade dos extremos autovalores de $\tilde{\mathbf{A}}$ calculado com base em 1000 realizações independentes utilizando a correção bootstrap. O modelo estocástico de blocos no qual queriam simular consistia em duas comunidades, na qual a probabilidade de conexão entre vértices da mesma comunidade é 0.7 ($C_{11} = C_{22} = 0.7$) e a probabilidade de conexão entre vértices de comunidades distintas é 0.3 ($C_{12} = C_{21} = 0.3$).

Foram realizadas três simulações representadas na Figura 9, as duas primeiras simulações com a densidade dos extremos autovalores de $\tilde{\mathbf{A}}$ sem a correção bootstrap para $n = 200$ e 1600 e uma simulação com a densidade dos extremos autovalores de $\tilde{\mathbf{A}}$ usando a correção com bootstrap para $n = 200$:

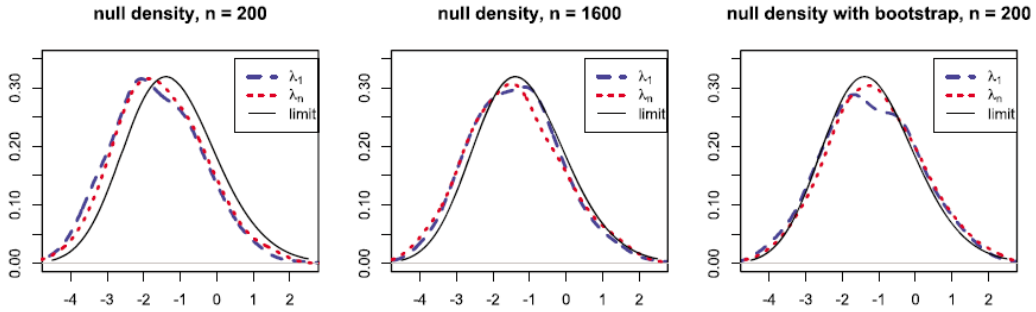



Figura 9 – As distribuições empíricas dos extremos autovalores de $\tilde{A}$ em 1000 repetições.
Fonte: LEI (2016, p. 411)

Note que comparando as duas simulações sem a correção bootstrap, conforme o número de vértices na rede aumenta, há uma aproximação na densidade dos extremos autovalores para distribuição Tracy-Widom. Entretanto, comparando a utilização da correção para $n = 200$, identificamos que este ajuste nos extremos autovalores das realizações da matriz aleatória aproximou-se da distribuição Tracy-Widom mesmo com um valor de $n$ relativamente pequeno quando comparado a $n = 1600$.

É interessante analisar que a simulação com a correção bootstrap para $n = 200$ obteve valores mais próximos da Tracy-Widom quando comparado sem a correção bootstrap para $n = 1600$, ou ainda, a distribuição empírica dos extremos autovalores convergem para $TW_1$. Consequentemente a correção com bootstrap é favorável para a questão computacional, utilizando-se uma menor quantidade de vértices para estimar os parâmetros e com um resultado mais adequado.

CAPÍTULO

# 4

# TESTE DE HIPÓTESES PARA REDES PONDERADAS

Nesta seção apresentaremos em detalhes uma adaptação do teste de hipóteses de Lei (2016), no qual a proposta será a estimar o número de comunidades da rede ponderada através do Modelo estocástico de blocos.

## 4.1 Estimador de Máxima Verossimilhança

Se fosse dado o valor das associações dos grupos $\mathbf{Z}$ e a matriz de probabilidade $\theta$, encontraríamos o valor de $K$ através do teste de hipóteses. Entretanto, não temos esses valores, então é necessário estimar os parâmetros, ou seja, $\hat{\theta}$ e $\hat{\mathbf{Z}}$.

Para isto, será calculado o estimador de máxima verossimilhança através da distribuição de probabilidade da Equação (2.4.4).

A distribuição de probabilidade do grafo $\mathbf{A}$, condicionada as comunidades $\mathbf{Z}$, é dada através da equação (4.1):

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \theta) = \frac{\prod_{l,m=1}^{K} e^{-\theta_{lm} N_{lm}} \theta_{lm}^{E_{lm}}}{\prod_{1 \leq i < j \leq n} \prod_{l,m=1}^{K} a_{ij}!}.$$

Podemos então escrever a função de verossimilhança da seguinte forma:

$$\mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A}) = \frac{\prod_{l,m=1}^{K} e^{-\theta_{lm} N_{lm}} \theta_{lm}^{E_{lm}}}{\prod_{1 \leq i < j \leq n} \prod_{l,m=1}^{K} a_{ij}!}. \tag{4.1}$$

Para obter o estimador de máxima verossimilhança fazemos:

i. Calculando o $\ln \mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A})$:

$$\begin{aligned}
\ln \mathscr{L}(Z, \theta \mid A) &= \ln \left( \frac{\prod_{l,m=1}^{K} e^{-\theta_{lm} N_{lm}} \theta_{lm}^{E_{lm}}}{\prod_{1 \le i < j \le n} \prod_{l,m=1}^{K} a_{ij}!} \right) \\
&= \sum_{l,m=1}^{K} \left( \ln e^{-\theta_{lm} N_{lm}} + \ln \theta_{lm}^{E_{lm}} - \sum_{1 \le i < j \le n} \ln a_{ij}! \right) \\
&= \sum_{l,m=1}^{K} \left( -\theta_{lm} N_{lm} + \ln \theta_{lm}^{E_{lm}} - \sum_{1 \le i < j \le n} \ln a_{ij}! \right) \\
&= \sum_{l,m=1}^{K} \left( -\theta_{lm} N_{lm} + E_{lm} \ln \theta_{lm} - \sum_{1 \le i < j \le n} \ln a_{ij}! \right).
\end{aligned}$$

ii. Derivando $\mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A})$ em relação a $\theta_{lm}$:

$$\begin{aligned}
\frac{\partial \mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A})}{\partial \theta_{lm}} &= \frac{\partial}{\partial \theta_{lm}} \left[ \sum_{l,m=1}^{K} \left( -\theta_{lm} N_{lm} + E_{lm} \ln \theta_{lm} - \sum_{1 \le i < j \le n} \ln a_{ij}! \right) \right] \\
&= -N_{lm} + \frac{E_{lm}}{\theta_{lm}}.
\end{aligned}$$

iii. Resolvendo $\dfrac{\partial \mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A})}{\partial \theta_{lm}} = 0$:

$$\begin{aligned}
\frac{\partial \mathscr{L}(\mathbf{Z}, \theta \mid \mathbf{A})}{\partial \theta_{lm}} &= 0 \\
-N_{lm} + \frac{E_{lm}}{\theta_{lm}} &= 0 \\
\frac{E_{lm}}{\theta_{lm}} &= N_{lm} \\
\frac{E_{lm}}{N_{lm}} &= \theta_{lm}.
\end{aligned}$$

Logo, maximizando a log-verossimilhança com base em $\hat{z}$, podemos estimar as probabilidades $\theta_{lm}$:

$$\hat{\theta}_{lm} = \frac{\hat{E}_{lm}}{\hat{N}_{lm}}, \tag{4.2}$$

onde a soma dos pesos entre nós da comunidade $l$ e $m$, $\hat{E}_{lm}$, é obtido através da Equação (2.3) e é calculado usando as comunidades estimadas $\hat{z}$. Da mesma forma, $\hat{N}_{lm}$ é obtido usando as comunidades estimadas $\hat{\mathbf{Z}}$.

Como o $\hat{\mathbf{Z}}$ as comunidades estimadas, $\hat{N}_l$ é o número de vértices estimado na comunidade $l$, então o valor estimado para $\theta_{lm}$ será:

$$\hat{\theta}_{lm} = \begin{cases} \dfrac{\sum_{i<j}^{n} A_{ij} \mathbb{1}\{\hat{z}_{il} = 1, \hat{z}_{jm} = 1\}}{\hat{N}_l \hat{N}_m}, & l \neq m, \\ \dfrac{\sum_{i<j}^{n} A_{ij} \mathbb{1}\{\hat{z}_{il} = 1, \hat{z}_{jm} = 1\}}{\hat{N}_l (\hat{N}_l - 1)/2}, & l = m, \end{cases} \tag{4.3}$$

## 4.2 Teste de hipóteses para redes ponderadas

Para estimar o número de comunidades através do teste de hipóteses para redes ponderadas, as hipóteses do teste são iguais ao teste de hipóteses para redes binárias como apresentada na Equação 3.5.

Considere uma matriz $\mathbf{Q}$ de dimensão $n \times n$, no qual as entradas $Q_{ij}$ representam o peso médio de conexão entre os vértices $i$ e $j$. Defina $g_i$ como a comunidade que o vértice $i$ pertence, ou seja, $g_i = a$ se, e somente se, $Z_{ia} = 1$, para $a = 1, \ldots, K$. Denotaremos como:

$$Q_{ij} = \theta_{g_i g_j}, \tag{4.4}$$

para $i, j = 1, \ldots, n$, de modo que $E(\mathbf{A}) = \mathbf{Q} - diag(\mathbf{Q})$. Defina $\tilde{\mathbf{A}}^*$ como:

$$\tilde{A}^*_{ij} = \frac{A_{ij} - Q_{ij}}{\sqrt{(n-1)Q_{ij}}}, \; i \neq j \quad \text{e} \quad \tilde{A}^*_{ii} = 0, \; \forall i. \tag{4.5}$$

Como cada variável aleatória da matriz de adjacência $A_{ij}$, segue uma distribuição de Poisson, então $E(A_{ij}) = Var(A_{ij}) = Q_{ij}$.

Para utilizar resultados sobre o espectro de grafos, é necessário verificar se a matriz $\tilde{\mathbf{A}}^*$ é uma matriz generalizada de Wigner. Primeiramente verificaremos que o valor esperado das entradas da matriz $\tilde{\mathbf{A}}^*$ são iguais a zero:

$$
\begin{aligned}
E(\tilde{A}^*_{ij}) &= E\left(\frac{A_{ij}-Q_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= E\left(\frac{A_{ij}}{\sqrt{(n-1)Q_{ij}}}-\frac{Q_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= E\left(\frac{A_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)-\left(\frac{Q_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= \frac{1}{\sqrt{(n-1)Q_{ij}}}\cdot E(A_{ij})-\left(\frac{Q_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= \frac{1}{\sqrt{(n-1)Q_{ij}}}\cdot (E(A_{ij})-Q_{ij})\\
&= \frac{1}{\sqrt{(n-1)Q_{ij}}}\cdot (Q_{ij}-Q_{ij})\\
&= 0.
\end{aligned}
$$

Agora verificaremos que o somatório da variância das entradas da matriz $\tilde{\mathbf{A}}^*$ será igual a um. Calculando a $Var(\tilde{A}^*_{ij})$:

$$
\begin{aligned}
Var(\tilde{A}^*_{ij}) &= Var\left(\frac{A_{ij}-Q_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= Var\left(\frac{A_{ij}}{\sqrt{(n-1)Q_{ij}}}\right)\\
&= \frac{1}{(n-1)Q_{ij}}\cdot Var(A_{ij})\\
&= \frac{1}{(n-1)Q_{ij}}\cdot Q_{ij}\\
&= \frac{1}{n-1}.
\end{aligned}
$$

Assim, como $\tilde{\mathbf{A}}^*$ é uma matriz simétrica, temos que $Var(\tilde{A}^*_{ii}) = 0$. Logo:

$$
\sum_{j=1}^{n} Var(\tilde{A}^*_{ij}) = \sum_{j\neq i}^{n} Var(\tilde{A}^*_{ij}) + 0 = \sum_{j\neq i}^{n} \frac{1}{n-1} = \frac{n-1}{n-1} = 1.
$$

Portanto, $E(\tilde{A}^*_{ij}) = 0$ para todo $(i,j)$ e $\sum_{j\neq i} Var(\tilde{A}^*_{ij}) = 1$. Utilizando novamente os resultados para matrizes aleatórias generalizada de Wigner, temos que:

$$
n^{2/3}[-\lambda_1(\tilde{A}^*_{ij}) - 2] \rightsquigarrow TW_1 \ \text{ e } \ n^{2/3}[\lambda_n(\tilde{A}^*_{ij}) - 2] \rightsquigarrow TW_1, \tag{4.6}
$$

onde $\lambda_1(\tilde{A}^*_{ij})$ é o menor autovalor de $\tilde{A}^*_{ij}$ e $\lambda_n(\tilde{A}^*_{ij})$ é o maior autovalor de $\tilde{A}^*_{ij}$. Denotamos $TW_1$ como a distribuição de Tracy-Widom com índice 1 e "$\rightsquigarrow$" para convergência da distribuição.

Note que $\mathbf{Q}$ depende da matriz $\theta$ que não é conhecida. Logo, ao invés de usarmos a matriz $\theta$, usamos o estimador de $\theta$ mostrada na Equação 4.3. Como $\theta$ depende de $g$ que não é conhecido, usamos o estimador de $g$, com base no estimador $\hat{Z}$, para então estimar $\mathbf{Q}$.

Definimos a matriz $\tilde{\mathbf{A}}$:

$$\tilde{A}_{ij} = \begin{cases} \frac{A_{ij}-\hat{Q}_{ij}}{\sqrt{(n-1)\hat{Q}_{ij}}}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{4.7}$$

onde $\hat{Q}_{ij} = \hat{\theta}_{\hat{g}_i \hat{g}_j}$.

Sob a hipótese nula e quando as estimativas $(\hat{g}, \hat{\theta})$ são precisas o suficiente, a convergência da relação (4.6) mostrará uma correspondência dos autovalores de $\tilde{\mathbf{A}}^*$, ou seja, o erro das estimativas será próximo de zero.

A estatística do teste é obtida da mesma maneira que descrito na Seção 3.3 na Equação (3.11).

Como no trabalho de Lei (2016) não foi abordado o teste de hipóteses para redes com pesos, não há resultados teóricos para o estimador $\hat{K}$. O foco desta dissertação é explorar via estudos de simulação a eficiência do estimador $\hat{K}$ para estimar o número de comunidades em redes com pesos discretos, descritos pela distribuição de Poisson.

CAPÍTULO

# 5

# SIMULAÇÕES

Para compreensão e validação dos testes de hipóteses realizamos estudos de simulação. Este capítulo está separado em três seções: as simulações das convergências dos extremos autovalores para a distribuição Tracy-Widom, as simulações para mensurar a eficiência na detecção do número de comunidades através dos testes de hipóteses para redes com pesos discretos e a comparação deste com outros métodos utilizados na estimação no número de comunidades.

## 5.1 Convergência para a Tracy-Widom

O teste de hipóteses para redes ponderadas abordada na Seção 4.2, utiliza resultados assintóticos da distribuição do menor e maior autovalor na teoria de matrizes aleatórias no qual convergem para a distribuição Tracy-Widom. Este resultado mostra que conforme o número de vértices no grafo aumenta, a frequência dos extremos autovalores se aproximam da distribuição $TW_1$ como na Equação (4.6).

Inicialmente construímos através do R a matriz de adjacência simétrica $\mathbf{A}$ com entradas aleatórias utilizando a distribuição Poisson ("rpois") com valor esperado 0.5 e $K = 1$. A ideia é simular 1000 vezes essa matriz, para analisar a frequência dos menores e maiores autovalores.

Assim, para analisar os extremos autovalores para os diferentes tamanhos de rede, foram simulados para grafos com vértices iguais a $n = 100, 300$ e 1000 como seguem nas Figuras 10 e 11.

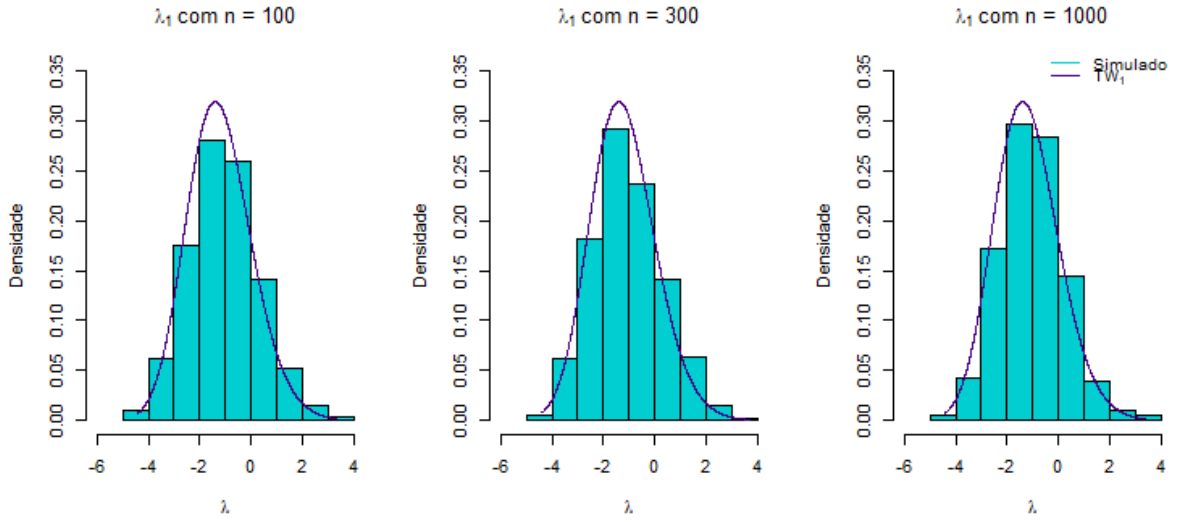



Figura 10 – Convergência dos menores autovalores

Na Figura 10 observamos que a frequência dos resultados dos menores autovalores das 1000 matrizes aleatórias geradas através da simulação, se aproximaram da distribuição $TW_1$, conforme aumentamos o número de vértices no grafo.

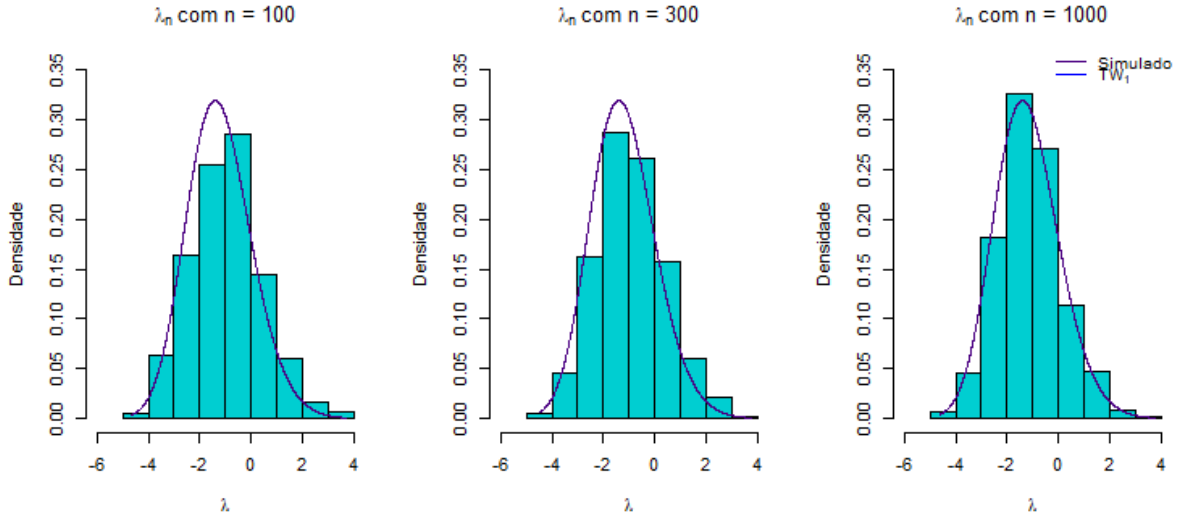



Figura 11 – Convergência dos maiores autovalores

Da mesma forma, na Figura 11 observamos que a frequência dos resultados dos maiores autovalores das 1000 matrizes aleatórias geradas através da simulação, se aproximaram da distribuição $TW_1$, conforme aumentamos o número de vértices no grafo.

Assim, através das simulações, há evidências que conforme o número de vértices em uma rede aumenta, a distribuição empírica do menor e maior autovalor de matrizes aleatórias convergem para a distribuição $TW_1$, como mostrava na Seção 4.2.

## 5.2 Estimação do número de comunidades via Teste de hipóteses

A ideia é estimar o número de comunidades utilizando os métodos do teste de hipótese de Lei (2016) através do Modelo estocástico de blocos. Para isso, adaptamos o código disponibilizado por Lei (2016), geramos uma matriz de adjacência simétrica com a diagonal de elementos 0 e com entradas seguindo a distribuição Poisson.

Realizamos a estimação das afiliações dos vértices às comunidades através do método "Spectral Clustering" proposto em Lei e Rinaldo (2015), no qual utiliza propriedades dos autovalores e autovetores da matriz de adjacência aplicando "*k-means*" para encontrar uma representação de dimensionalidade reduzida dos dados que facilita a separação das comunidades $\hat{z}_n$.

Para as simulações, é necessário escolher o número de comunidades que esperamos em uma rede e verificar a eficiência do método dos testes de hipóteses em relação à estimação das comunidades. Ainda, em redes aleatórias é comum estudar casos das atribuições dos vértices às comunidades seguindo uma distribuição de probabilidade $\pi = (\pi_1, \ldots, \pi_K)$, então quando a probabilidade de um vértice pertencer as comunidades da rede é $\frac{1}{K}$ temos o caso balanceado e caso contrário, temos o caso desbalanceado.

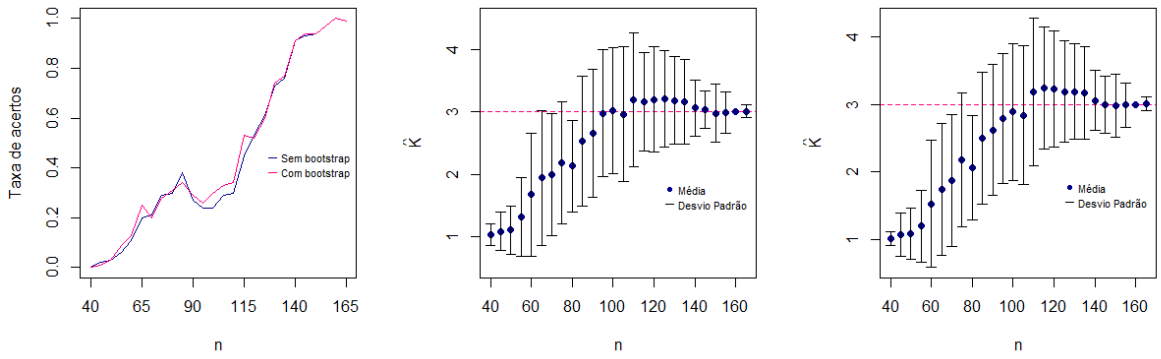
Neste capítulo, utilizamos redes simuladas com $K = 3$ e consideramos o caso balanceado com $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, o caso desbalanceado com $\pi = (0.4, 0.3, 0.3)$ e o caso com maior desbalanceamento com $\pi = (0.5, 0.3, 0.2)$.

A taxa de acertos do método baseado em teste de hipóteses para estimar o número de comunidades em relação a quantidade de vértices, foi calculado através da média de 100 simulações para cada $n$. Assim, para analisar este comportamento utilizamos os três casos de rede em relação ao balanceamento ou desbalanceamento como mostrado na Figura 12.

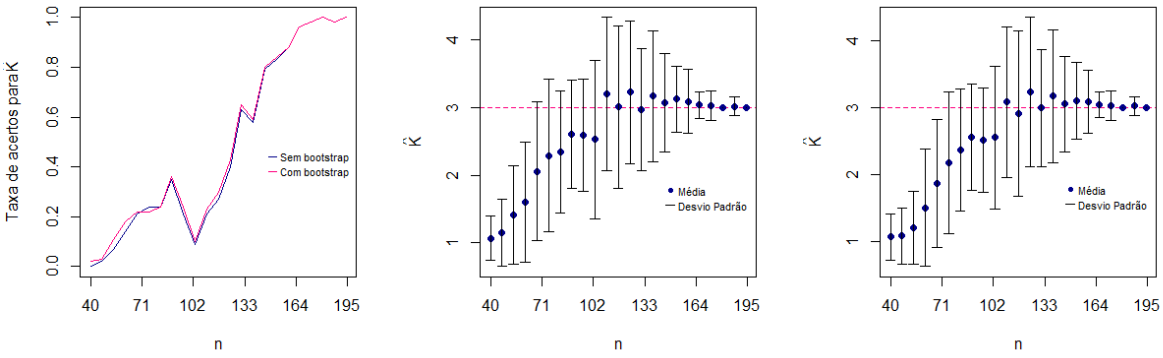
O número verdadeiro de comunidades simuladas foram $K = 3$ para as Figuras 12a, 12b e 12c, assim os gráficos da esquerda mostram a relação da taxa de acerto do número de comunidades dos testes de hipóteses conforme o aumento no número de vértices da rede, os gráficos no centro mostram a média e desvio padrão na estimação do número de comunidades no teste de hipótese sem bootstrap e os gráficos da direita mostram a média e desvio padrão na estimação do número de comunidades no teste de hipótese com bootstrap.

Em redes, a assortatividade é uma métrica utilizada para compreender a tendência de vértices se conectarem a outros semelhantes. Para o caso de redes ponderadas com comunidades, é esperado um peso maior nas arestas de vértices da mesma comunidade do que vértices entre comunidades, ou ainda, há mais assortatividade em vértices da mesma comunidade do que entre comunidades. Seguindo essa ideia, para as simulações na Figura 12 que exploram os pesos das arestas dos vértices em relação às comunidades, consideramos a distribuição Poisson com parâmetro $a = 4$ como o peso esperado das arestas de vértices da mesma comunidade e a

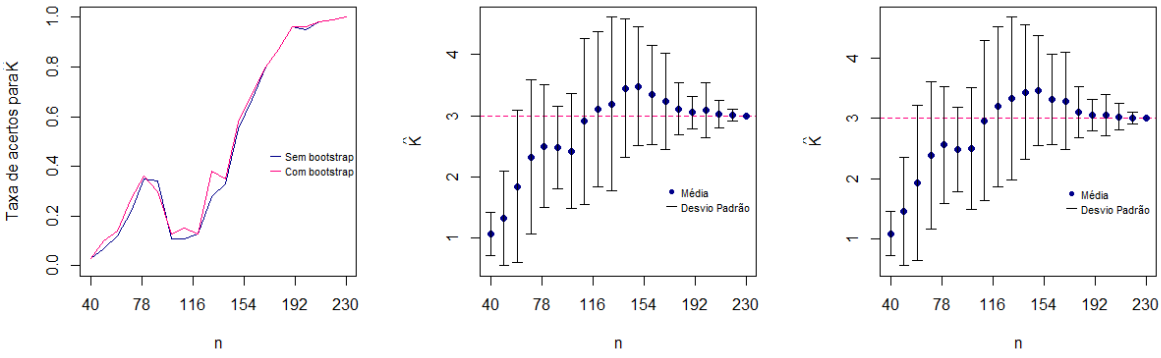
distribuição Poisson com parâmetro $b = 3$ como o peso esperado das arestas de vértices entre comunidades.




(a) Caso balanceado, onde $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$




(b) Caso desbalanceado, onde $\pi = (0.4, 0.3, 0.3)$




(c) Caso desbalanceado, onde $\pi = (0.5, 0.3, 0.2)$

Figura 12 – Simulações dos Testes de Hipóteses sem bootstrap e com bootstrap, variando o número $n$ de vértices do grafo, com o peso esperado das arestas dos vértices dentro e entre comunidades a=4 e b=3.

No caso 12a, para redes balanceadas o teste de hipóteses com bootstrap apresentou maior

taxa de acertos na estimação de comunidades em relação ao teste de hipóteses sem bootstrap. Comparando a média e o desvio padrão, o teste com bootstrap apresentou maior variabilidade na estimação de K, entretanto obteve a taxa de acertos maior.

No caso da Figura 12b, para redes desbalanceadas os dois testes de hipóteses apresentaram taxas de acertos semelhantes na estimação de comunidades. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

No caso da Figura 12c, para redes mais desbalanceadas o teste de hipóteses com bootstrap apresentou maior taxa de acertos na estimação de comunidades em relação ao teste de hipóteses sem bootstrap. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

Comparando os três casos das Figuras 12a, 12b e 12c, é possível notar que conforme a rede tem desbalanceamento, maior deve ser o intervalo de $n$ para que a taxa de acerto seja próxima de 1. Além disso, conforme o número de $n$ aumenta é esperado que a taxa de acerto aumente também, entretanto, aproximadamente na mediana dos intervalos de $n$, houve uma queda na taxa de acertos que é justificada através da média e variabilidade das simulações realizadas para cada $n$.

Vimos os três casos de redes balanceadas e desbalanceadas na Figura 12 em relação ao número de vértices na rede e a taxa de acerto na estimação no número de comunidades pelos testes de hipóteses, quando os pesos esperados entre os vértices eram fixos ($a = 4$ e $b = 3$). Entretanto, em redes reais o comportamento desses pesos médios não é regular, podendo existir uma diferença muito pequena entre $a$ e $b$ que influenciaria na estimação no número de comunidades, então investigamos qual deve ser a diferença mínima para um maior acerto através do teste.

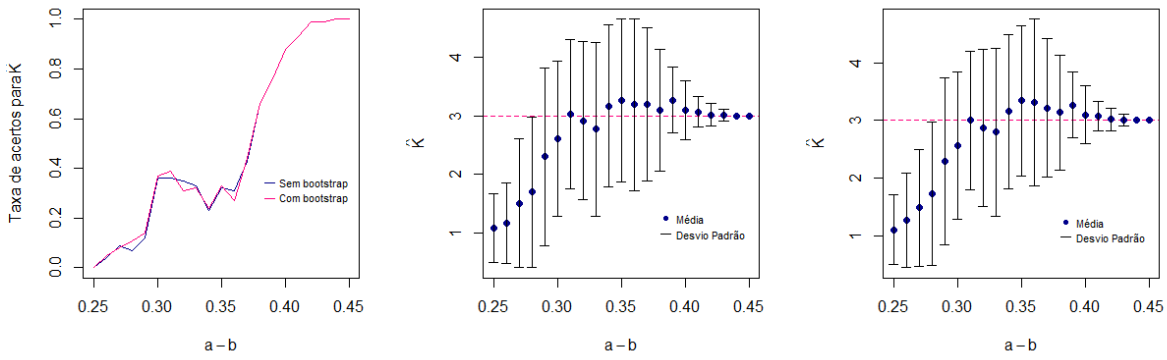
Para as simulações apresentadas nas Figuras 13, 14 e 15, fixamos os valores de $n = 500$, o valor real de comunidades para as simulações com $K = 3$ e variamos a diferença entre o peso esperado de vértices dentro e entre comunidades para verificar se os métodos detectariam corretamente o número de comunidades, simulando 100 vezes para cada diferença.

O parâmetro da distribuição Poisson, que representa o peso médio das arestas, poderia influenciar na detecção do número de comunidades, então para cada tipo de rede balanceada e desbalanceada mostrados na Figura 12, utilizamos dois casos: a distribuição Poisson com parâmetros $a \in (3,4)$ e $b = 3$ e a distribuição Poisson com parâmetros $a \in (5,6)$ e $b = 5$.

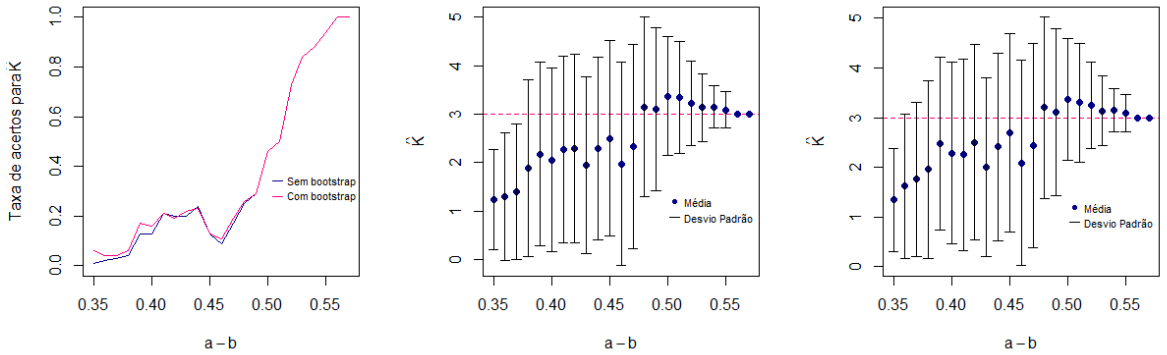
Para as três Figuras 13, 14 e 15, os gráficos da esquerda mostram a relação da taxa de acerto dos testes de hipóteses conforme o aumento da diferença entre o valor esperado nos pesos das arestas dos vértices, os dois gráficos restantes mostram a média e desvio padrão na estimação do número de comunidades conforme a diferença entre o valor esperado nos pesos das arestas dos vértices no teste de hipótese sem e com bootstrap.

Na Figura 13 temos uma rede balanceada, na Figura 13a usamos distribuição Poisson

com $a \in (3.25, 3.45)$ e $b = 3$. Na Figura 13b usamos distribuição Poisson com $a \in (5.35, 5.57)$ e $b = 5$.




(a) Caso balanceado, com Poisson variando entre 3.25 a 3.45, $|a-b| = (0.25, 0.45)$




(b) Caso balanceado, com Poisson variando entre 5.35 a 5.57, $|a-b| = (0.35, 0.57)$

Figura 13 – Simulações dos Testes de Hipóteses sem bootstrap e com bootstrap, variando a diferença entre o peso esperado das arestas dos vértices na rede balanceada e com o número de vértices fixo $n = 500$.

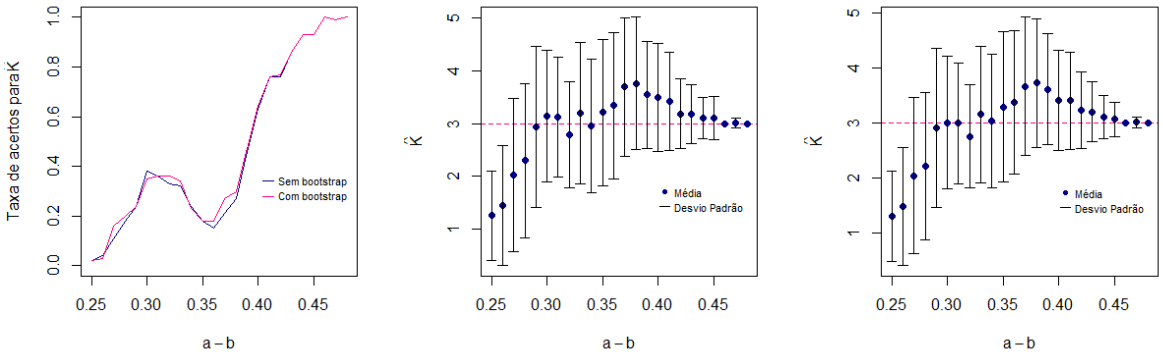
No caso da Figura 13a, para redes com menor peso esperado entre os vértices, os dois testes de hipóteses apresentaram taxas de acertos semelhantes na estimação de comunidades. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

No caso da Figura 13b, para redes com maior peso esperado entre os vértices, o teste de hipóteses com bootstrap apresentou maior taxa de acertos na estimação de comunidades em relação ao teste de hipóteses sem bootstrap, principalmente quando $|a-b| = (0.35, 0.45)$. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

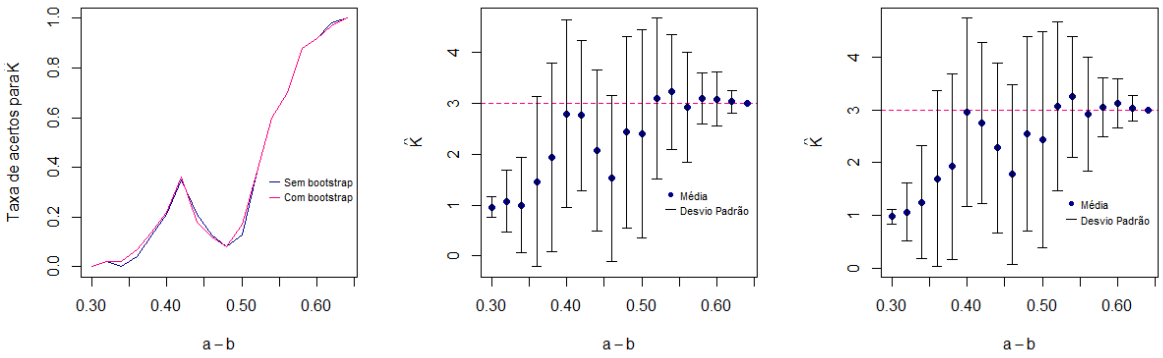
Comparando as duas Figuras 13a e 13b, é possível notar que conforme a rede tem um maior peso esperado das arestas entre os vértices a diferença entre $a$ e $b$ também deve ser maior, para que os testes de hipóteses estimem corretamente o número de comunidades. Além disso, conforme a diferença entre os pesos esperados aumenta é provável que a taxa de acerto aumente

também, entretanto, aproximadamente na mediana dos intervalos de $|a-b|$, houve uma pequena queda na taxa de acertos que é justificada através da média e variabilidade das simulações.

Na Figura 14 temos uma rede desbalanceada com $\pi = (0.4, 0.3, 0.3)$, na Figura 14a usamos distribuição Poisson com $a \in (3.25, 3.48)$ e $b = 3$. Na Figura 14b usamos distribuição Poisson com $a \in (5.3, 5.64)$ e $b = 5$.




(a) Caso desbalanceado, com $\pi = (0.4, 0.3, 0.3)$ e Poisson variando entre 3.25 a 3.48,$|a-b| = (0.25, 0.48)$




(b) Caso desbalanceado, com $\pi = (0.4, 0.3, 0.3)$ e Poisson variando entre 5.3 a 5.64, $|a-b| = (0.3, 0.64)$.

Figura 14 – Simulações dos Testes de Hipóteses sem bootstrap e com bootstrap, variando a diferença entre o peso esperado das arestas dos vértices na rede desbalanceada e com o número de vértices fixo $n = 500$.

No caso da Figura 14a, para redes com menor peso esperado entre os vértices, o teste de hipóteses com bootstrap apresentou maior taxa de acertos na estimação de comunidades em relação ao teste de hipóteses sem bootstrap, principalmente quando $|a-b| = (0.25, 0.37)$. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

No caso da Figura 14b, para redes com maior peso esperado entre os vértices, os dois testes de hipóteses apresentaram taxas de acertos semelhantes na estimação de comunidades. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

Comparando as duas Figuras 14a e 14b, é possível notar que conforme a rede tem um

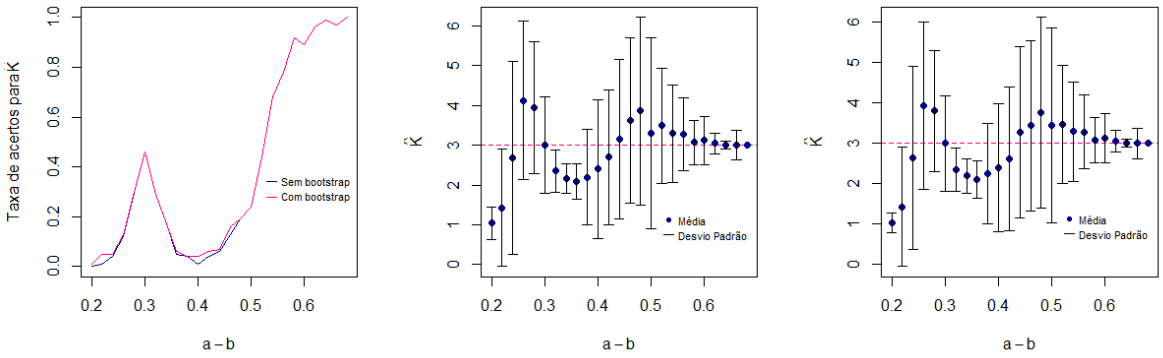
maior peso esperado das arestas entre os vértices a diferença entre $a$ e $b$ também deve ser maior, para que os testes de hipóteses estimem corretamente o número de comunidades. Além disso, conforme a diferença entre os pesos esperados aumenta é provável que a taxa de acerto aumente também, entretanto, aproximadamente na mediana dos intervalos de $|a-b|$, houve uma queda na taxa de acertos, sendo justificada através da média e variabilidade das simulações.

Na Figura 15 temos uma rede desbalanceada com $\pi = (0.5, 0.3, 0.2)$, na Figura 15a usamos distribuição Poisson com $a \in (3.25, 3.68)$ e $b = 3$. Na Figura 14b usamos distribuição Poisson com $a \in (5.32, 5.87)$ e $b = 5$.

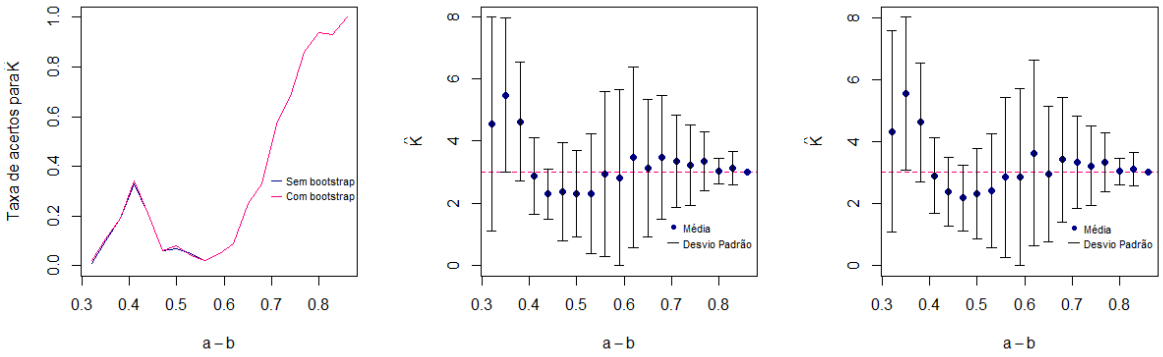
No caso da Figura 15a, para redes com menor peso esperado entre os vértices, os dois testes de hipóteses apresentaram taxas de acertos semelhantes na estimação de comunidades. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

No caso da Figura 15b, para redes com maior peso esperado entre os vértices, o teste de hipóteses com bootstrap apresentou maior taxa de acertos na estimação de comunidades em relação ao teste de hipóteses sem bootstrap, principalmente quando $|a-b| = (0.35, 0.45)$. Comparando a média e o desvio padrão, os dois testes apresentam valores semelhantes.

Comparando as duas Figuras 15a e 15b, é possível notar que conforme a rede tem um maior peso esperado das arestas entre os vértices a diferença entre $a$ e $b$ também deve ser maior, para que os testes de hipóteses estimem corretamente o número de comunidades. Além disso, conforme a diferença entre os pesos esperados aumenta é provável que a taxa de acerto aumente também, entretanto, aproximadamente na mediana dos intervalos de $|a-b|$, houve uma grande queda na taxa de acertos que é justificada através da média e variabilidade das simulações.




(a) Caso desbalanceado, com $\pi = (0.5, 0.3, 0.2)$ e Poisson variando entre 3.2 a 3.68,$|a-b| = (0.2, 0.68)$




(b) Caso desbalanceado, com $\pi = (0.5, 0.3, 0.2)$ e Poisson variando entre 5.32 a 5.87, $|a-b| = (0.32, 0.87)$.

Figura 15 – Simulações dos Testes de Hipóteses sem bootstrap e com bootstrap, variando a diferença entre o peso esperado das arestas dos vértices na rede desbalanceada e com o número de vértices fixo $n = 500$.

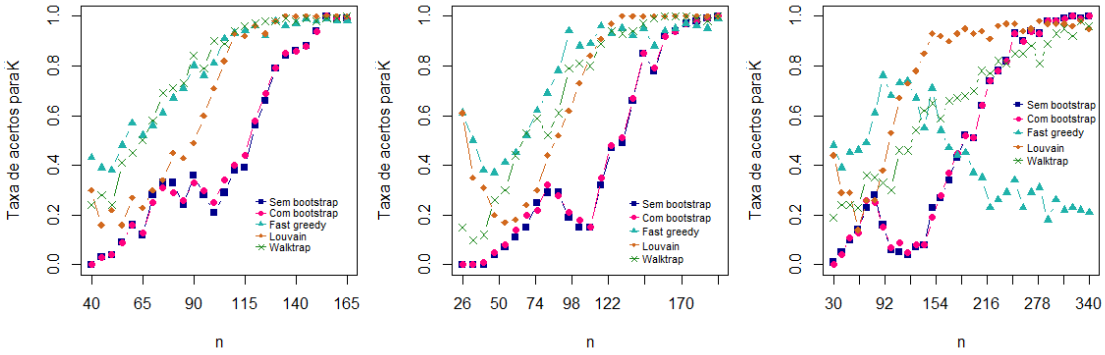
As Figuras 13, 14 e 15 evidenciam tanto pequenas quanto grandes variações na taxa de acerto do número estimado de comunidades em torno da mediana, sendo que diversos fatores podem influenciar esses resultados. Uma possível explicação para essas oscilações seria o fato de as estimativas das atribuições dos vértices às comunidades $\hat{\mathbf{Z}}$ não estarem suficientemente próximas das atribuições reais $\mathbf{Z}$. Dessa forma, para intervalos pequenos de $|a-b|$ as estimativas de K apresentam um comportamento aleatório, podendo ou não se aproximar do valor real que fundamenta a taxa de acerto no intervalo anterior à mediana.

No geral, o teste se mostrou eficiente na estimação no número de comunidades em redes com pesos discretos, quando temos o número de vértices sendo maior que pesos médios de arestas entre vértices de comunidades sendo $a = 4$ e $b = 3$.

## 5.3 Comparando métodos

Nesta seção, realizaremos a comparação de métodos de estimação no número de comunidades com os testes de hipóteses apresentados na Seção 4. Selecionamos os métodos mais utilizados na literatura de acordo com Hajibabaei, Seydi e Koochari (2023), He *et al.* (2021), Palowitch, Bhamidi e Nobel (2017), que são: *Fast greedy*, *Louvain* e *Walktrap*.

Comparamos os modelos através dos tipos de redes em relação ao seu balanceamento e desbalanceamento como mostrado na Figura 16. Para isto, fixamos $a = 4$ e $b = 3$, analisamos conforme aumentamos o número de vértices da rede como seria a taxa de acertos para a estimação do número de comunidades através dos métodos e simulamos 100 vezes para cada $n$.




(a) Caso balanceado, onde $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (b) Caso desbalanceado, onde $\pi = (0.4, 0.3, 0.3)$ (c) Caso desbalanceado, onde $\pi = (0.5, 0.3, 0.2)$

Figura 16 – Comparação dos métodos para estimar o número de comunidades, variando o número de vértices do grafo, com o peso esperado das arestas dos vértices dentro e entre comunidades a=4 e b=3.

Para as Figuras 16a e 16b, temos os casos balanceado e desbalanceado em que todos os métodos *Fast greedy*, *Louvain* e *Walktrap* estimaram melhor o número de comunidades ao longo do aumento no tamanho da rede quando comparamos aos métodos do teste de hipóteses.

Para a Figura 16c, caso mais desbalanceado, o método *Louvain* estimou melhor o número de comunidades ao longo do aumento no tamanho da rede quando comparamos aos métodos do teste de hipóteses. O método *Walktrap* estimou bem o número de comunidades quando $n \in (30, 180)$, mas após isso, teve uma oscilação na taxa de acertos enquanto os métodos *Louvain* e testes de hipóteses tinham uma taxa de acerto igual a 100%.

Ainda sobre a Figura 16c, inicialmente o *Fast greedy* mostrou a melhor estimação no número de comunidades em relação à todos os métodos quando temos uma rede pequena $n \in (30, 100)$ e após esse aumento no número de vértices, este método se torna o menos adequado para realizar a estimação.

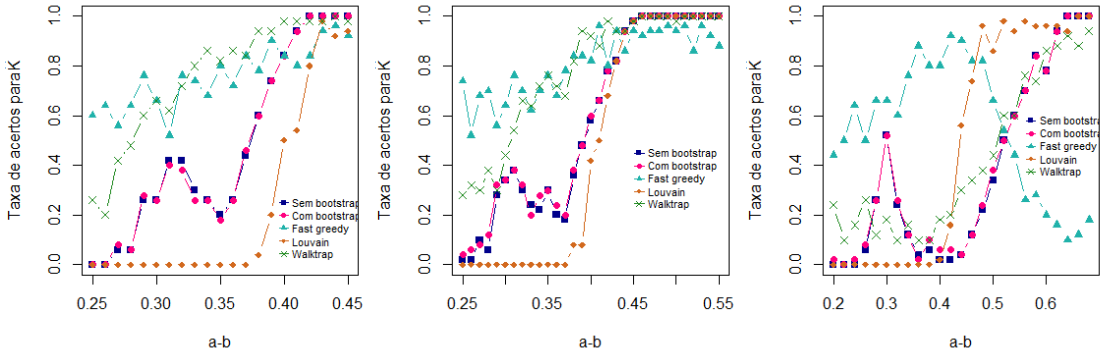
A Figura 16 mostra que para uma rede pequena em torno de $n = (20, 40)$, os métodos

possuem uma taxa de acerto significativamente alta de 20% a 50% quando comparamos aos testes de hipóteses com uma taxa de 0 a 10%. Então para uma rede deste tamanho, seria adequado os métodos *Fast greedy*, *Louvain* e *Walktrap*.

Quanto a Figura 16c, o *Louvain* se mostrou melhor em todo valor de $n$ quando comparamos aos testes de hipóteses, mas as estimativas do *Walktrap* e principalmente do *Fast greedy*, são métodos melhores em somente alguns intervalos de $n$.

Como visto na seção 5.2, o peso do parâmetro da distibuição Poisson poderia influenciar na detecção do número de comunidades, então escolhemos o caso da distribuição Poisson com parâmetros $a \in (3,4)$ e $b = 3$ representadas na Figura 17.

As simulações da Figura 17 foram para redes com $n = 500$, em que uma diferença muito pequena entre $a$ e $b$ pode influenciar na estimação no número de comunidades, então investigamos qual deve ser a diferença mínima para um maior acerto através do teste.




(a) Caso balanceado, onde $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (b) Caso desbalanceado, onde $\pi = (0.4, 0.3, 0.3)$ (c) Caso desbalanceado, onde $\pi = (0.5, 0.3, 0.2)$

Figura 17 – Comparação dos métodos para estimar o número de comunidades variando a diferença entre o peso esperado das arestas dos vértices, com o número de vértices fixo $n = 500$.

No caso balanceado na Figura 17a, os métodos *Fast greedy* e *Walktrap* estimaram melhor o número de comunidades ao longo da diferença dos pesos médios de arestas entre vértices quando comparamos aos métodos do teste de hipóteses e *Louvain*. Ainda, os testes de hipóteses estimaram melhor o número de comunidades que o método *Louvain* no intervalo de $|a-b| \in (0.25, 0.45)$.

No caso desbalanceado na Figura 17b, os métodos *Fast greedy* e *Walktrap* estimaram melhor o número de comunidades ao longo da diferença dos pesos médios de arestas entre vértices quando comparamos aos métodos do teste de hipóteses e *Louvain*, entretanto o *Fast greedy* teve uma oscilação na taxa de acertos após $|a-b| = 0.45$. Ainda, os testes de hipóteses estimaram melhor o número de comunidades que o método *Louvain* no intervalo de $|a-b| \in (0.25, 0.45)$.

No caso mais desbalanceado na Figura 17c tivemos melhores métodos locais dependendo

do intervalo $|a-b|$, assim não houve um melhor método na estimação do número de comunidades em todo o intervalo proposto. Podemos separar em intervalos e indicar qual seria o método mais adequado para cada caso:

- No intervalo $|a-b| \in (0.25, 0.45)$, o método mais adequado é o *Fast greedy*;
- No intervalo $|a-b| \in (0.45, 0.6)$, o método mais adequado é o *Louvain*;
- No intervalo $|a-b| \in (0.6, 0.7)$, os métodos que mostraram uma maior taxa de acerto são *Louvain* e os testes de hipóteses;

Na Figura 17c o método *Walktrap* teve um comportamento similar aos testes de hipóteses, entretanto quando aumentamos a diferença entre $a$ e $b$ ainda teve uma oscilação na taxa de acertos após $|a-b| = 0.6$, enquanto os testes de hipóteses e *Louvain* estimavam corretamente a partir deste intervalo.

A Figura 17 mostra que para uma diferença pequena dos pesos esperados em torno de $|a-b| \in (0.25, 0.3)$, os métodos possuem uma taxa de acerto significativamente alta de 20% a 80% quando comparamos aos testes de hipóteses com uma taxa de 0 a 40%. Então para redes com uma pequena diferença entre os pesos esperados, seria adequado os métodos *Fast greedy* e *Walktrap*. Entretanto, para uma diferença dos pesos esperados acima de 0.4, em geral, seria adequado os métodos *Louvain* e os testes de hipóteses.

Para as simulações representadas na Figura 17, os testes de hipóteses tem uma taxa de acertos muito maior ao longo do intervalo $|a-b|$ quando comparado ao método *Louvain*. Ambos realizam as estimativas corretamente após $|a-b| = 0.4$, entretanto antes disso, o *Louvain* tem uma taxa de acerto em torno de 0.

Principalmente para os casos mais desbalanceados como nas Figuras 16c e 17c, o método *Fast greedy* parece ser o menos adequado, devido a grande queda na taxa de acertos no número de comunidades conforme aumentamos o número de vértices ou a diferença entre os pesos médios.

Esta pesquisa mostrou os testes de hipóteses para redes com pesos discretos utilizando a distribuição Poisson, sobre os casos mais recomendados para utilização deste método na estimação do número de comunidades. Assim, é natural pensarmos aplicações em dados reais para compreender o desempenho do teste de hipóteses na estimação do número de comunidades.

CAPÍTULO

# 6

# APLICAÇÃO

Para aplicação em dados reais dos testes de hipóteses, utilizamos a área de transito rápida da baía (BART) que é um sistema público de transporte ferroviário que conectam diversas regiões da Baía de São Francisco, na Califórnia. Este sistema conecta a Península de São Francisco a cidades como Berkeley, Oakland, Fremont, Walnut Creek, Dublin/Pleasanton, entre outras cidades do litoral leste da baía.

A coleta dos dados foi impulsionada pela pandemia COVID-19, que resultou em uma flexibilização das restrições de quarentena/lockdown nas áreas litorâneas, apesar do aumento nos casos de COVID-19, gerou mudanças significativas nos padrões de mobilidade urbana. Essas alterações afetaram os hábitos de viagem das pessoas e o uso do transporte público nas diversas regiões da Baía de São Francisco.

Os dados foram fundamentais para entender as variações no comportamento de deslocamento e como diferentes áreas da baía responderam às mudanças nas restrições sanitárias. Ao identificar essas diferenças, é possível avaliar o impacto da pandemia no transporte público e fornecer informações valiosas para o planejamento de políticas públicas mais eficazes, que atendam às necessidades de mobilidade da população em cenários futuros de situações de emergência. Os dados são abertos e estão disponíveis BART - Bay Area Rapid Transit (2024).

Neste estudo, utilizamos 50 estações de trens que conectam diversas regiões da Baía de São Francisco, totalizando 2.450 trajetos entre elas. Para as conexões, atribuímos um peso de acordo com a quantidade de repetições do trajeto entre as estações, ou seja, a quantidade de trens que passaram entre as estações, com o objetivo de aplicar os testes de hipóteses apresentados na Seção 4.

Na Tabela 1 foi realizada uma análise dos anos de 2020 à 2023, com o objetivo de avaliar a distribuição das estações, considerando os agrupamentos existentes e as mudanças ocorridas ao longo desse período. Apesar de não ter ocorrido um aumento no número de estações nem no total de trajetos disponíveis ao longo dos anos, observou-se um crescimento na quantidade de

trajetos e no peso médio, à medida que a rotina foi gradualmente retornando à normalidade após a pandemia de COVID-19. Em 2023, os dados coletados correspondem ao primeiro trimestre do ano, o que resulta em uma quantidade reduzida de trajetos.

Tabela 1 – Análise de Estações, Trajetos e Pesos Médios (2020-2023).

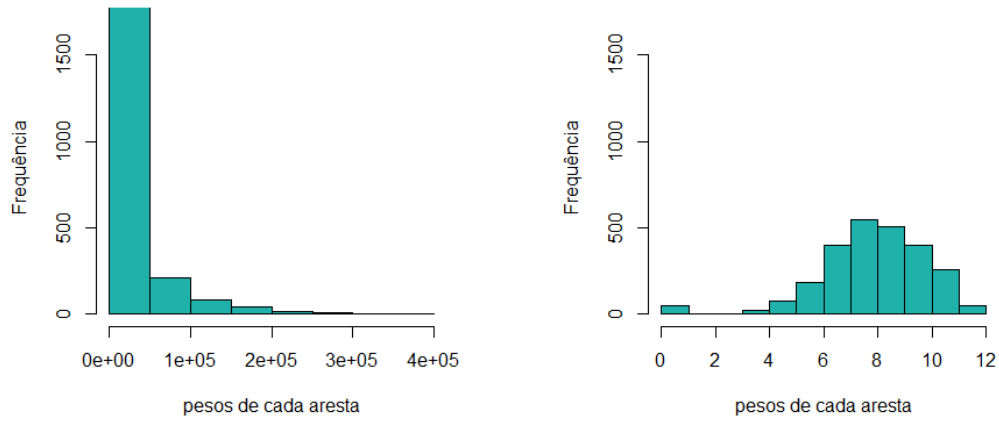
| | n° estações | n° trajetos | quantidade de trajetos | peso médio por trajeto |
|---|---|---|---|---|
| 2020 | 50 | 2450 | 5.811.863 | 2.372,19 |
| 2021 | 50 | 2450 | 6.598.064 | 2.693,09 |
| 2022 | 50 | 2450 | 8.245.370 | 3.365,46 |
| 2023 | 50 | 2450 | 2.815.982 | 1.149,38 |

Na Seção 4.2, temos a suposição de que o peso das arestas segue uma distribuição Poisson e apresentamos que a variância da matriz de adjacência desempenha um papel fundamental quando construímos a matriz generalizada, que é utilizada na estimação do número de comunidades por meio dos testes de hipóteses.

No entanto, nesta aplicação, o peso dos trajetos apresentaram uma variabilidade consideravelmente alta. Uma das opções para reduzir a dispersão devido à variação elevada é aplicar a transformação logarítmica. Para aplicá-la, utilizamos as entradas da matriz de adjacência com a seguinte regra:

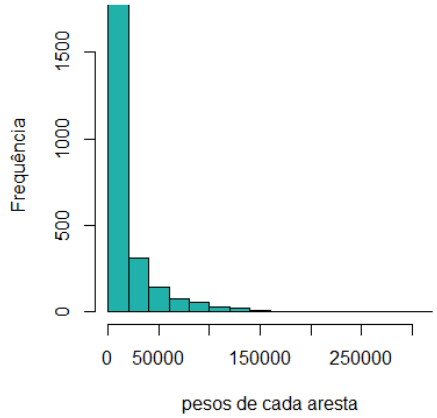
$$A_{ij}^{log} = \log(A_{ij} + 1).$$

Na Figura 18, para cada ano realizamos um comparativo dos histogramas de frequência em relação aos pesos das arestas, sem e com transformação, para mostrar a redução da variação.

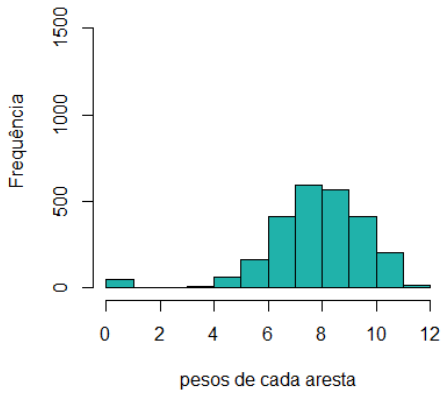



(a) 2020

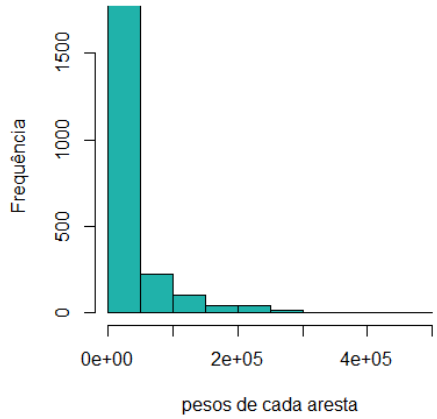
(b) 2020 - transformado



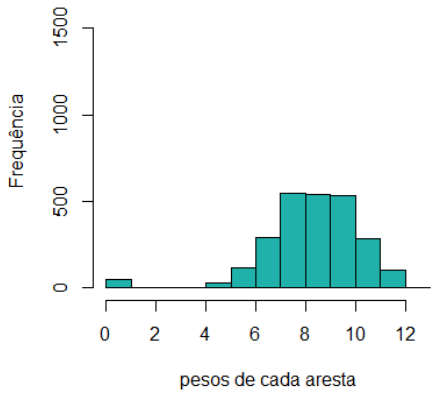

(c) 2021



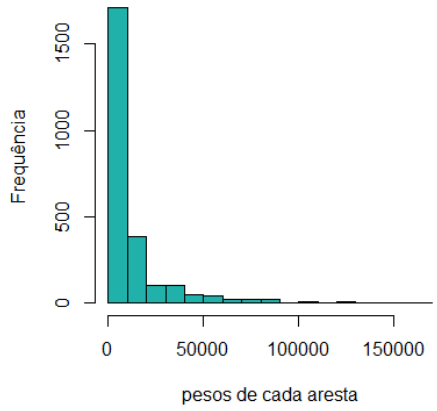

(d) 2021 - transformado



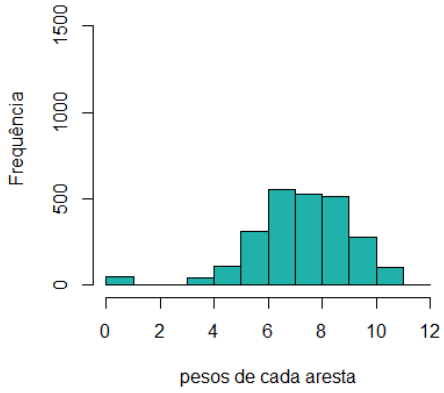

(e) 2022




(f) 2022 - transformado




(g) 2023




(h) 2023 - transformado

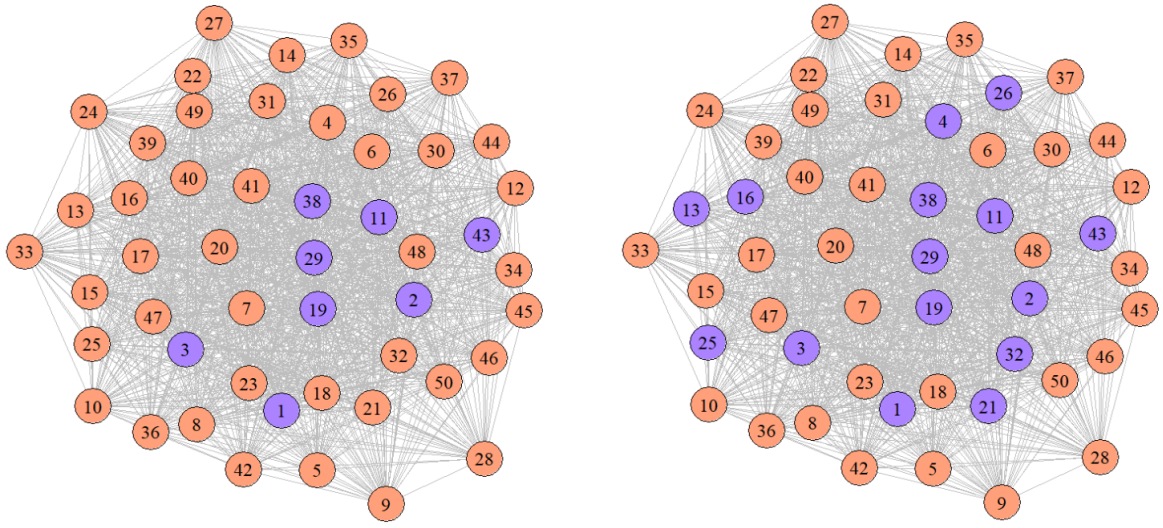
Figura 18 – Histogramas de frequência em relação aos pesos das arestas, comparativo da transformação por ano

Essa transformação foi crucial, pois reduz a variação dos dados e aproxima os valores da média e da variância, aproximando aos parâmetros de uma distribuição Poisson.

Com as premissas estabelecidas e a análise dos dados realizada, utilizamos o Teste de hipóteses para redes ponderadas apresentado na Seção 4, conforme o ano de referência. Ao aplicar esse método com nível de significância de 1%, foram estimadas duas comunidades. No entanto, as associações das estações às comunidades variam ao compararmos os métodos "Spectral Clustering" e "Spectral Clustering Spherical", como discutiremos a seguir ao analisarmos as atribuições às comunidades.

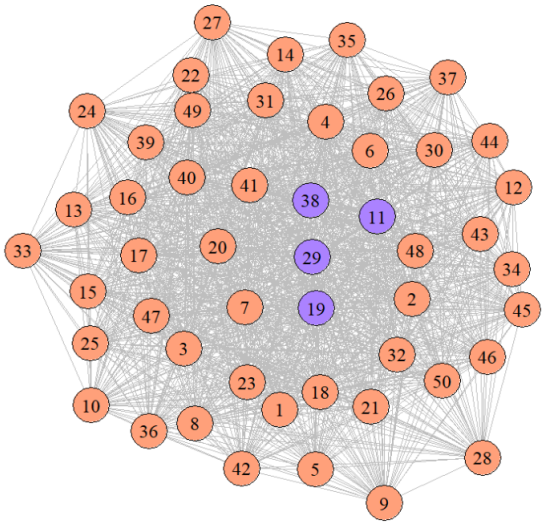
A estimação das afiliações dos vértices às comunidades é realizada por meio do método "Spectral Clustering", abordada na Seção 5.2. Uma modificação desse algoritmo proporciona uma melhoria significativa na precisão da estimação, especialmente quando utilizado em conjunto com o Modelo Estocástico de Blocos com grau corrigido. O "Spectral Clustering Spherical" é eficaz em redes com grau de arestas heterogêneas, o que torna essa abordagem mais aderente às características observadas em redes reais. Nesse método, os autovetores associados aos menores autovalores da matriz Laplaciana são explorados e normalizados, ou seja, cada autovetor resultante da decomposição espectral é projetado para ter norma unitária, o que facilita a redução da dimensionalidade dos dados. Após a normalização dos autovetores, o algoritmo "*k-means*" é aplicado como no método tradicional "Spectral Clustering", para realizar a divisão dos vértices às comunidades.

As redes dos dados BART estão representados na Figura 19, onde a cor roxa indica a comunidade *r*, que contém o menor número de estações, e a cor salmão representa a comunidade *s*, que agrupa as estações restantes. O método "Spectral Clustering" apresentado à esquerda nas Figuras 19a, 19c, 19e, 19g, enquanto o "Spectral Clustering Spherical" é representado à direita nas Figuras 19b, 19d, 19f, 19h.

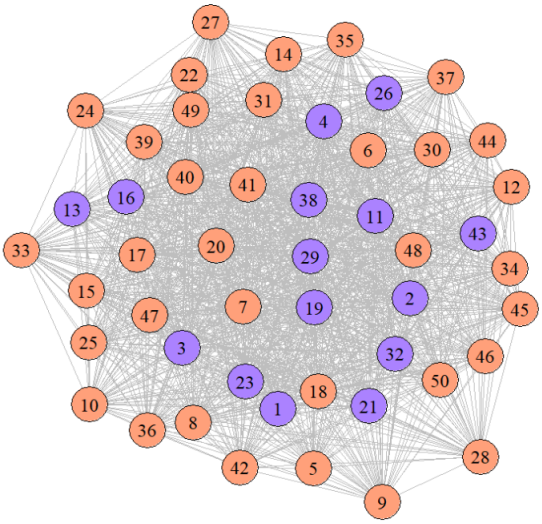


(a) Spectral Clustering (2020)

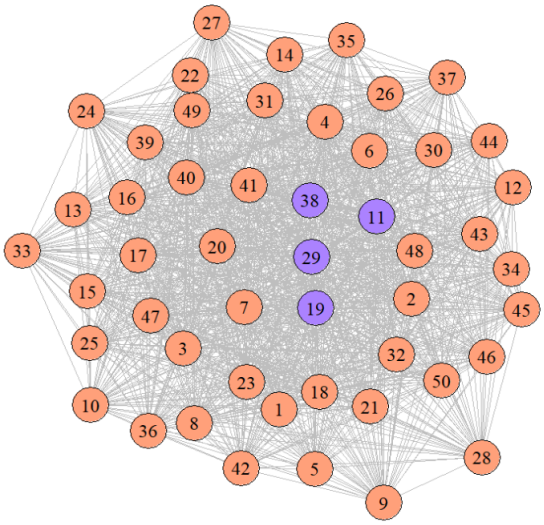
(b) Spectral Clustering Spherical (2020)



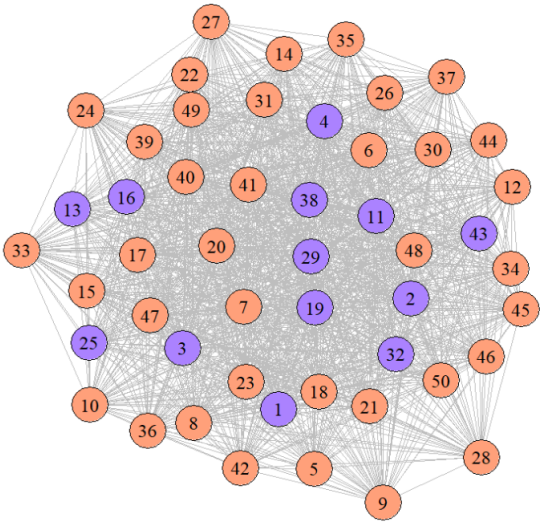
(c) Spectral Clustering (2021)



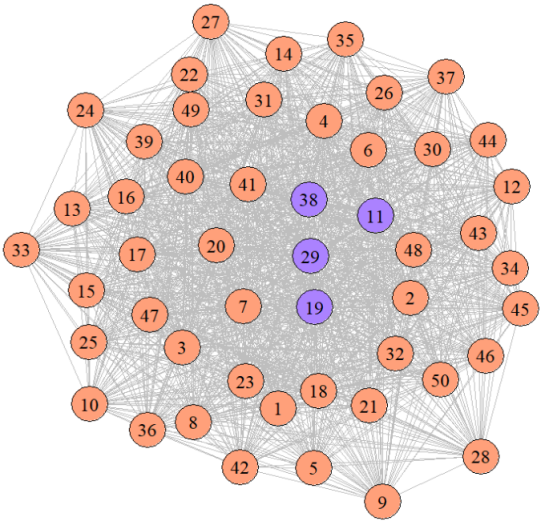
(d) Spectral Clustering Spherical (2021)



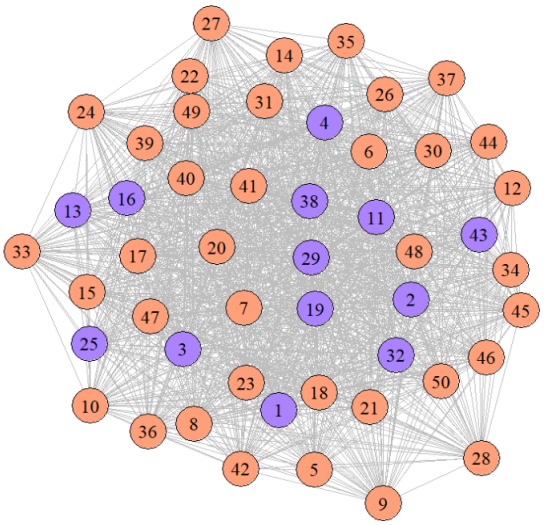
(e) Spectral Clustering (2022)



(f) Spectral Clustering Spherical (2022)



(g) Spectral Clustering (2023)



(h) Spectral Clustering Spherical (2023)

Figura 19 – Grafos com comunidades dos dados BART, mostrando as atribuições dos vértices às comunidades (normal e esférico) no método de teste de hipóteses

Construímos a Tabela 2 com a quantidade de estações nas comunidades $r$ e $s$ ao longo dos anos 2020 à 2023:

Tabela 2 – Quantidade de vértices em duas comunidades $r$ e $s$, utilizando os métodos Spectral Clustering e Spectral Clustering Spherical (2020-2023).

| | Spectral Clustering | | Spectral Clustering Sphere | |
|---|---|---|---|---|
| | r | s | r | s |
| 2020 | 8 | 42 | 15 | 35 |
| 2021 | 4 | 46 | 15 | 35 |
| 2022 | 4 | 46 | 13 | 37 |
| 2023 | 4 | 46 | 13 | 37 |

A Tabela 3 (Apêndice A) apresenta a relação entre os rótulos dos vértices, as estações do BART e as cidades em que essas estações estão localizadas. Enquanto a Tabela 4 (Apêndice A) detalha o tamanho das cidades em função de sua população. Ambas as tabelas servirão como base para as análises subsequentes.

Nas Figuras 19a, 19c, 19e, 19g foi aplicado o método "Spectral Clustering", levando na atribuição das estações 11, 19, 29 e 38 à mesma comunidade ao longo dos anos, localizadas na cidade de São Francisco. No entanto, observou-se que a comunidade $r$ apresentou um número reduzido de estações em comparação com o total de 50. Esse desequilíbrio pode indicar problemas na estimação das afiliações dos vértices, possivelmente relacionados à quantidade dos pesos das arestas, o que pode afetar a precisão da atribuição das comunidades.

Nas Figuras 19b, 19d, 19f, 19h foi utilizado o "Spectral Clustering Spherical" e as estações 1, 2, 3, 4, 11, 13, 16, 19, 29, 32, 38, 43 das cidades de São Francisco (7), Oakland (4) e Berkeley (1) foram atribuídas a mesma comunidade ao longo dos anos, enquanto as estações 21, 23, 25 e 26 das cidades de Oakland (3) e Hayward (1) alternaram nessa comunidade.

Ao comparar os dois algoritmos, observa-se que todas as estações geradas pela comunidade $r$ por meio do "Spectral Clustering" pertencem à mesma comunidade ao serem analisadas com o "Spectral Clustering Spherical". Esse resultado demonstra que o desempenho do Spectral Clustering Spherical determina com maior precisão e otimiza as associações dos vértices às comunidades.

A estimação de duas comunidades nos dados pode ser interpretada como as regiões $r$ (cidades maiores) e $s$ (cidades menores), que são em grande parte, resultado da alta densidade populacional e das características típicas das cidades metropolitanas, observado em áreas como São Francisco, Oakland, Hayward e Berkeley. Nessas áreas urbanas, as atividades econômicas geram uma demanda crescente por transportes públicos, além de intensificar a mobilidade de pessoas de cidades menores como San Bruno, Pittsburg, Orinda e South San Francisco às regiões metropolitanas, causando uma crescente concentração populacional.

CAPÍTULO

7

# CONCLUSÃO

Nesta dissertação foi feita uma revisão acerca dos conceitos à teoria de grafos e os principais modelos de grafos aleatórios Modelo de Erdös-Rényi e Modelo Estocástico de Blocos. Apresentamos os testes de hipóteses proposto por Lei (2016) para estimar o número de comunidades em redes binárias, no qual o objetivo principal desta pesquisa foi realizar uma adaptação desse método para estimar o número de comunidades em redes com pesos discretos via um estudo de simulação.

Selecionamos alguns tipos de redes nas simulações, no que diz respeito ao balanceado e desbalanceamento, ao número de vértices e em relação a diferença entre os pesos médios das arestas entre vértices das comunidades. Utilizamos o valor verdadeiro de $K = 3$ para todas as simulações.

Sobre os testes de hipóteses, quando observamos a taxa de acerto no número de comunidades em relação ao número de vértices da rede ponderada, é possível notar que conforme a rede tem desbalanceamento, maior deve ser o número de vértices $n$ para que a taxa de acerto seja próxima de 1. Além disso, conforme o número de $n$ aumenta aproximadamente na mediana dos intervalos de $n$, houve uma queda na taxa de acertos que é justificada através da média e variabilidade das simulações realizadas para cada $n$.

Vimos os três casos de redes balanceadas e desbalanceadas na Figura 12 em relação ao número de vértices na rede e a taxa de acerto na estimação no número de comunidades pelos testes de hipóteses, quando os pesos esperados entre os vértices eram fixos ($a = 4$ e $b = 3$). Entretanto, em redes reais o comportamento desses pesos médios não é regular, podendo existir uma diferença muito pequena entre $a$ e $b$ que influenciaria na estimação no número de comunidades, então investigamos sobre a diferença mínima para um maior acerto através do teste.

Quando observamos a taxa de acerto no número de comunidades em relação a diferença entre os pesos esperados $a$ e $b$, é possível notar que conforme a rede tem um maior peso esperado

das arestas entre os vértices, a diferença entre $a$ e $b$ também deve ser maior, para que os testes de hipóteses estimem corretamente o número de comunidades.

Nas Figuras 13, 14 e 15 em torno da mediana houveram pequenas e grandes quedas na taxa de acerto no número estimado de comunidades, existem muitos fatores que podem influenciar. Como as atribuições dos vértices as comunidades de maneira equívocada que pode desencadear uma conexão com menor ou maior peso de vértices que não são próximos da situação real.

No geral, o teste se mostrou eficiente na estimação no número de comunidades em redes com pesos discretos, quando temos o número de vértices sendo maior que os pesos médios de arestas entre vértices de comunidades sendo $a = 4$ e $b = 3$.

Realizamos a comparação dos métodos *Fast greedy*, *Louvain* e *Walktrap* em relação aos testes de hipóteses na estimação do número de comunidades.

Quando observamos a taxa de acerto no número de comunidades em relação ao número de vértices da rede ponderada, para uma rede pequena em torno de $n = (20, 40)$ seria adequado os métodos *Fast greedy*, *Louvain* e *Walktrap*. Para casos balanceado e desbalanceado, os métodos *Fast greedy*, *Louvain* e *Walktrap* são mais adequados. Entretanto, para o caso mais desbalanceado, o método *Louvain* estimou melhor o número de comunidades ao longo do aumento no tamanho da rede quando comparamos aos métodos do teste de hipóteses.

Quando observamos a taxa de acerto no número de comunidades em relação a diferença entre os pesos esperados, para redes com uma pequena diferença entre os pesos esperados, seria adequado os métodos *Fast greedy* e *Walktrap*. Entretanto, para uma diferença dos pesos esperados acima de 0.4, em geral, seria adequado os métodos *Louvain* e os testes de hipóteses. Os testes de hipóteses possuem uma taxa de acertos muito maior ao longo do intervalo $|a - b|$ quando comparado ao método *Louvain*. Ambos realizam as estimativas corretamente após $|a - b| = 0.4$, entretanto antes disso, o *Louvain* tem uma taxa de acerto em torno de 0.

Para a aplicação de testes de hipóteses em dados reais, utilizamos a área de transito rápida da baía (BART) referente ao período de 2020 a 2023, com o objetivo de estimar o número de comunidades e analisar as atribuições das estações a essas comunidades. Utilizamos o teste de hipóteses com os algoritmos “Spectral Clustering” e “Spectral Clustering Spherical”. Comparando os métodos, revelou que o “Spectral Clustering Spherical” oferece uma atribuição mais precisa das estações às comunidades, refletindo as características das regiões metropolitanas e a crescente demanda por transporte público nas áreas de maior densidade populacional.

Um trabalho futuro pode ser focado no estudo teórico da consistência da estimação do número de comunidades em redes ponderadas. Além disso, a pesquisa realizada nesta dissertação, pode ser expandida ao considerar diferentes distribuições de probabilidade para os pesos das arestas.

Os códigos utilizados no R nas simulações e nas aplicações em dados reais de redes

ponderadas foram adaptados do trabalho de Lei (2016), que disponibilizou os códigos utilizados em seu artigo para redes binárias por meio de seu website[1]. Os códigos produzidos nessa pesquisa, estão disponíveis no GitHub[2] da autora.

[1] Website - Jing Lei
[2] Github - Luana A. Tamura

# REFERÊNCIAS

BART - Bay Area Rapid Transit. **Base de dados BART Ridership Reports**. 2024. Acesso em: 03 jul. 2024. Disponível em: <https://www.bart.gov/about/reports/ridership>. Citado na página 59.

ERDŐS, L.; YAU, H.-T.; YIN, J. Rigidity of eigenvalues of generalized wigner matrices. **Advances in Mathematics**, Elsevier, v. 229, n. 3, p. 1435–1515, 2012. Citado na página 36.

HAJIBABAEI, H.; SEYDI, V.; KOOCHARI, A. Community detection in weighted networks using probabilistic generative model. **Journal of Intelligent Information Systems**, Springer, v. 60, n. 1, p. 119–136, 2023. Citado na página 56.

HE, Z.; CHEN, W.; WEI, X.; LIU, Y. On the statistical significance of communities from weighted graphs. **Scientific Reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 20304, 2021. Citado na página 56.

LEE, C.; WILKINSON, D. J. A review of stochastic block models and extensions for graph clustering. **Applied Network Science**, Springer, v. 4, n. 1, p. 1–50, 2019. Citado nas páginas 15 e 23.

LEE, J. O.; YIN, J. A necessary and sufficient condition for edge universality of wigner matrices. **Duke Mathematical Journal**, Duke University Press, v. 163, n. 1, p. 117–173, 2014. Citado na página 36.

LEI, J. A goodness-of-fit test for stochastic block models. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 44, n. 1, p. 401–424, 2016. Citado nas páginas 16, 27, 31, 34, 37, 38, 39, 41, 45, 49, 65 e 67.

LEI, J.; RINALDO, A. Consistency of spectral clustering in stochastic block models. 2015. Citado na página 49.

NEPUSZ, T.; PETRÓCZI, A.; NÉGYESSY, L.; BAZSÓ, F. Fuzzy communities and the concept of bridgeness in complex networks. **Physical Review E**, APS, v. 77, n. 1, p. 016107, 2008. Citado na página 28.

NEWMAN, M. E. J. **Networks: An Introduction**. [S.l.]: New York: Oxford University Press, 2010. Citado na página 15.

PALOWITCH, J.; BHAMIDI, S.; NOBEL, A. B. Significance-based community detection in weighted networks. **J. Mach. Learn. Res.**, v. 18, p. 188–1, 2017. Citado na página 56.

TRACY, C. A.; WIDOM, H. Distribution functions for largest eigenvalues and their applications. **arXiv preprint math-ph/0210034**, 2002. Citado na página 34.

World Population Review. **US Cities by Population 2024**. 2024. Acesso em: 10 dez. 2024. Disponível em: <https://worldpopulationreview.com/us-cities>. Citado na página 73.

ZACHARY, W. W. An information flow model for conflict and fission in small groups. **Journal of anthropological research**, University of New Mexico, v. 33, n. 4, p. 452–473, 1977. Citado na página 28.

APÊNDICE

# A

# TABELAS ESTAÇÕES BART

A Tabela 3 apresenta a numeração dos vértices utilizados na Figura 19, relacionando com o nome das estações do BART e as respectivas cidades em que estão localizadas.

Tabela 3 – Relação entre as estações BART, as cidades e os rótulos dos vértices.

| rótulo do vértice | nome da estação | cidade |
|---|---|---|
| 1 | 12th Street / Oakland City Center | Oakland |
| 2 | 16th Street Mission | São Francisco |
| 3 | 19th Street Oakland | Oakland |
| 4 | 24th Street Mission | São Francisco |
| 5 | Antioch | Antioch |
| 6 | Ashby | Berkeley |
| 7 | Balboa Park | São Francisco |
| 8 | Bayfair | San Leandro |
| 9 | Berryessa / North San José | San José |
| 10 | Castro Valley | Castro Valley |
| 11 | Civic Center | São Francisco |
| 12 | Colma | Colma |
| 13 | Coliseum | Oakland |
| 14 | Concord | Concord |
| 15 | Daly City | Daly City |

| rótulo do vértice | nome da estação | cidade |
|---|---|---|
| 16 | Berkeley | Berkeley |
| 17 | El Cerrito Del Norte | El Cerrito Del Norte |
| 18 | Dublin/Pleasanton | Dublin/Pleasanton |
| 19 | Embarcadero | São Francisco |
| 20 | Fremont | Fremont |
| 21 | Fruitvale | Oakland |
| 22 | Glen Park | São Francisco |
| 23 | Hayward | Hayward |
| 24 | Lafayette | Lafayette |
| 25 | Lake Merritt | Oakland |
| 26 | MacArthur | Oakland |
| 27 | Millbrae | Millbrae |
| 28 | Milpitas | Milpitas |
| 29 | Montgomery Street | São Francisco |
| 30 | North Berkeley | Berkeley |
| 31 | North Concord | Concord |
| 32 | Oakland International Airport | Oakland |
| 33 | Orinda | Orinda |
| 34 | Pittsburg/Bay Point | Pittsburg |
| 35 | Pleasant Hill | Pleasant Hill |
| 36 | Pittsburg Center | Pittsburg |
| 37 | El Cerrito Plaza | El Cerrito |
| 38 | Powell Street | São Francisco |
| 39 | Richmond | Richmond |
| 40 | Rockridge | Oakland |
| 41 | San Leandro | San Leandro |
| 42 | San Bruno | San Bruno |
| 43 | San Francisco International Airport | São Francisco |
| 44 | South Hayward | Hayward |
| 45 | South San Francisco | South San Francisco |
| 46 | Union City | Union City |
| 47 | Warm Springs | Fremont |
| 48 | Walnut Creek | Walnut Creek |
| 49 | West Dublin/Pleasanton | Dublin/Pleasanton |
| 50 | West Oakland | Oakland |

A Tabela 4 apresenta uma estimativa da população de cada cidade dos Estados Unidos que possui uma estação do BART.

Tabela 4 – População aproximada das cidades dos Estados Unidos.

| cidade | população |
|---|---|
| San José | 957.000 |
| São Francisco | 789.000 |
| San Leandro | 500.000 |
| Oakland | 435.000 |
| Fremont | 225.000 |
| Hayward | 155.000 |
| Concord | 122.000 |
| Berkeley | 118.000 |
| Antioch | 117.000 |
| Richmond | 113.000 |
| Daly City | 105.000 |
| Dublin/Pleasanton | 90.000 |
| San Leandro | 85.000 |
| Milpitas | 77.000 |
| Pittsburg | 76.000 |
| Walnut Creek | 69.000 |
| Castro Valley | 66.000 |
| Union City | 64.000 |
| South San Francisco | 62.000 |
| San Bruno | 40.000 |
| Pleasant Hill | 34.000 |
| Lafayette | 25.000 |
| El Cerrito | 25.000 |
| Millbrae | 22.000 |
| El Cerrito Del Norte | 20.000 |
| Orinda | 20.000 |
| Colma | 1.500 |

Fonte: World Population Review.


ICMC
USP
SÃO CARLOS