UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

MESTRADO ACADÊMICO EM FÍSICA

CARLOS GERMANO LIMA DE SOUSA

ENUMERAÇÃO EFETIVA DE CAMINHOS ENTRE PARES DE NÓS EM UMA REDE COMPLEXA

FORTALEZA

2023

CARLOS GERMANO LIMA DE SOUSA

ENUMERAÇÃO EFETIVA DE CAMINHOS ENTRE PARES DE NÓS EM UMA REDE COMPLEXA

Dissertação apresentada ao Curso de Mestrado Acadêmico em Física do Programa de Pós-Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Física. Área de Concentração: Física da matéria condensada.

Orientador: Prof. Dr. André Auto Moreira.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S696e Sousa, Carlos Germano Lima de.
Enumeração efetiva de caminhos entre pares de nós em uma rede complexa / Carlos Germano Lima de Sousa. – 2023.
62 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Física, Fortaleza, 2023.
Orientação: Prof. Dr. André Auto Moreira.

1. Matriz de adjacência. 2. Comunidades. 3. Caminhos. 4. Redes. I. Título.

CDD 530

CARLOS GERMANO LIMA DE SOUSA

ENUMERAÇÃO EFETIVA DE CAMINHOS ENTRE PARES DE NÓS EM UMA REDE COMPLEXA

Dissertação apresentada ao Curso de Mestrado Acadêmico em Física do Programa de Pós-Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Física. Área de Concentração: Física da matéria condensada.

Aprovada em: 23/02/2023.

BANCA EXAMINADORA

Prof. Dr. André Auto Moreira (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Saulo Davi Soares e Reis
Universidade Federal do Ceará (UFC)

Prof. Dr. Rilder de Sousa Pires
Universidade de Fortaleza (Unifor)

Aos meus pais, amigos e a todos que me ajudaram de alguma forma nesta jornada.

AGRADECIMENTOS

Gostaria de agradecer a todos aqueles que me ajudaram e me apoiaram nestes quase dois anos de mestrado. Primeiramente, ao meu pai Carlos, minha mãe Edna e meu padrasto Rodrigo por todo o apoio e torcida. Também aos meus irmãos Guilherme e Vitória pelo amor que é típico de irmãos. Além da minha namorada Raquel por todo o amor e carinho e por ter me dado apoio em todos os momentos, especialmente, nos mais difíceis.

Gostaria de agradecer ao meu orientador professor André Auto Moreira pela orientação, paciência e estímulo a este trabalho e ao professor Saulo Davi Soares e Reis por toda ajuda fornecida. Além de todos os professores que tive durante este percurso que deixaram marcas na minha vida. Além disso, gostaria de agradecer a todos os meus colegas e amigos do Laboratório de Sistemas Complexos por toda a convivência e companheirismo nesse período. Com certeza fizeram meus dias mais tranquilos e divertidos.

Gostaria de agradecer também a todos os meus amigos que me deram apoio e entenderam minha ausência em alguns momentos. Em especial, aos meus grandes amigos João Paulo, Lucas e Jaciel que durante todos estes anos tive o prazer de ter a amizade. E a todos os meus amigos que conheci ou me aproximei nesta jornada, em especial, Vasco, André e Lara.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Networks are present everywhere. All we need is an eye for them." (BARABASI, 2003, p. 7.)

RESUMO

Este trabalho tem como objetivo estudar os caminhos entre dois nós que estão cada um em uma comunidade e enumerar quantos caminhos efetivos existem entre eles. Para isso, inicialmente, introduzimos redes aleatórias abordando alguns conceitos, representação matemática, aplicações fundamentais e alguns tipos de comunidades. Em seguida, fizemos uma breve revisão sobre o método de detecção de comunidade baseado em inferência estatística e uma apresentação sobre o conceito de likelihood e como ela se aplica em tal método. Nós definimos um conceito novo que chamamos de caminhos efetivos para determinar a enumeração dos segmentos em paralelo entre dois nós utilizando o seguinte processo. Partindo de uma rede qualquer, supomos que todas as ligações são removidas. Depois disso reconstruímos a rede recolocando as ligações uma a uma aleatoriamente. Além disso, definimos um instante de aglutinação como sendo o número de ligações que são colocadas antes que os nós se aglutinem em um mesmo agregado. A enumeração dos caminhos efetivos determina o conjunto de segmentos em paralelo que tem maior probabilidade de produzir a mesma distribuição de instantes de aglutinação. Para obter essa enumeração, usamos dois métodos distintos. O primeiro método baseia-se na descoberta de uma equação que pode ser usada em um ajuste linear utilizando mínimos quadrados lineares gerais. E o segundo método utilizado foi a maximização da likelihood com o mesmo intuito de calcular a quantidade de segmentos. Finalmente, tínhamos almejado utilizar e aprimorar os resultados obtidos por nós a fim de detectar comunidades em redes. Para tanto, criamos redes artificiais com estruturas de comunidades e investigamos a enumeração no caso em que os pares de nós estão na mesma comunidade e no caso que estão em comunidades distintas. Porém, nossos resultados mostraram que esta abordagem é restrita demais para reproduzir perfeitamente a distribuição de instantes de aglutinação. Assim, sugerindo que modelos mais sofisticados precisam ser ainda elaborados.

Palavras-chave: matriz de adjacência; comunidades; caminhos; redes.

ABSTRACT

This work aims to study the paths between two nodes that are each in a community and count how many effective paths exist between them. For this, initially, we introduce random networks addressing some concepts, mathematical representation, fundamental applications and some kinds of communities. Then, we made a brief review of the method of community detection based on statistical inference and a presentation on the concept of likelihood and how it is applied in such a method. We define a new concept that we call effective paths to determine the enumeration of segments in parallel between two nodes using the following process. Starting from any network, we assume that all connections are removed. After that we replace them one by one at random. Furthermore, we define a merging instant to be the number of edges that are placed before the nodes merge into the same cluster. The enumeration of effective paths determines the set of parallel segments that are most likely to produce the same distribution of merging instants. To obtain this enumeration, we use two different methods. The first method relies on finding an equation that can be used in a linear fit using general linear least squares. And the second method used was likelihood maximization algorithm with the same purpose of calculating the number of segments. Finally, we aimed to use and improve the results obtained by us in order to detect communities in networks. For that, we created artificial networks with community structures and investigated the enumeration in the case where the pairs of nodes are in the same community and in the case that they are in different communities. However, our results showed that this approach is too restricted to perfectly reproduce the distribution of merging instants. Thus, suggesting that more sophisticated models still need to be elaborated.

Keywords: adjacency matrix; community; path; networks.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O estudo de redes em Física permite entender que modelos de redes podem representar de forma satisfatória redes reais ou no mínimo trazer sugestões de propriedades que se apresentam em redes reais. Exemplos de redes reais podem ser rede de amigos, redes de colaboração de pesquisadores, redes elétricas dentre outras.

Existem vários modelos matemáticos para redes a depender de sua estrutura. Neste trabalho, o único que precisamos nos atentar é o modelo de rede aleatória ou rede de Erdös-Rényi. Além disso, estendemos o conceito de redes aleatórias para dar conta de redes divididas em comunidades ou grupos. Tais redes, no nosso caso, são basicamente redes em que a densidade de ligações aleatórias intragrupos é bem maior que a densidade de ligações aleatórias intergrupos.

Assim, este trabalho está dividido da seguinte forma, o segundo capítulo é dedicado ao estudo de redes com ênfase em redes aleatórias, suas propriedades, abordagem matemática e a introdução do conceito de comunidades. O terceiro capítulo está relacionado ao estudo de detecção de comunidades especialmente o modelo de detecção de comunidades por inferência estatística em que se utiliza conhecimentos e aplicação de likelihood e de alguns conceitos aprendidos no capítulo anterior.

No último capítulo está presente a nossa contribuição direta neste trabalho. Tal pesquisa dedica-se ao estudo da enumeração efetiva dos caminhos que aglutinam pares de nós. Assim, pode-se pensar como uma extensão da matriz de adjacência. Tendo em vista que a matriz de adjacência apenas registra ligações diretas entre nós e neste trabalho se almeja encontrar caminhos entre nós que podem ter mais de uma ligação. Portanto, sendo mais geral que a matriz de adjacência convencional. Por fim, no último capítulo está nossas conclusões e perspectivas de trabalhos futuros seguindo este tema.

2 REDES ALEATÓRIAS

Um modelo simples de rede aleatória é aquele em que se fixa o número n de nós e m de ligações. Neste caso, cria-se n nós e se faz m ligações entre eles aleatoriamente. Tal modelo é, comumente, simbolizado por $G(n,m)$. Podemos pensar também de forma equivalente que há $\binom{n}{2}$ pares de nós que podem ser ligados de $\binom{\binom{n}{2}}{m}$ formas diferentes.

Estritamente falando, uma rede aleatória é definida em termos de ensemble de redes. Assim, o modelo é melhor definido pela distribuição de probabilidade $P(G)$ sobre todos as redes $G(n,m)$,

$$P(G) = \frac{1}{\binom{\binom{n}{2}}{m}}. \tag{2.1}$$

Portanto, ao se mencionar as propriedades da rede, refere-se, de fato, às propriedades de um ensemble de redes. Por exemplo, ao se referir a uma propriedade qualquer de uma rede aleatória, $O(G)$, usualmente, significa a média sobre o ensemble da forma:

$$\langle O \rangle = \sum_G P(G)O(G). \tag{2.2}$$

Há algumas vantagens em se expressar dessa forma, a primeira é que a propriedade é facilmente calculada principalmente para redes com grandes números de nós. A segunda diz respeito ao fato de mostrar exatamente as propriedades que procuramos, como o grau médio ou o diâmetro médio da rede. Afinal, se é um comportamento típico que se quer ao criar um modelo de rede, então a média do ensemble é um bom guia. Por fim, é demonstrável que para muitas medidas a distribuição de valores da rede tem pico acentuado quanto maior é a quantidade de nós. Assim, no limite de n grande, o valor é próximo da média.

Após se entender que a definição de uma rede aleatória se faz sobre um ensemble de redes com probabilidade sobre todas as redes do ensemble. Tem-se o ensemble $G(n,p)$ em que cada rede G apresenta probabilidade

$$P(G) = p^m(1-p)^{\binom{n}{2}-m}. \tag{2.3}$$

Este ensemble foi primeiramente estudado por (SOLOMONOFF; RAPOPORT, 1951). No entanto, é muito mais conhecido pelos trabalhos de Paul Erdős e Alfréd Rényi que publicaram uma série de artigos (ERDÖS; RÉNYI, 1959), (ERDOS; RENYI, 1960) e (ERDŐS; RÉNYI, 1961) sobre o modelo entre as décadas de 50 e 60 do século passado. A contribuição deste autores são tão importantes que usa-se a expressão Rede de Erdös-Renyi para se referir à rede

aleatória e este será o modelo usado ao longo do trabalho quando nos referimos a redes aleatórias. Vamos nas próximas seções estudar algumas propriedades e quantidades importantes a serem analisadas quando se estuda redes aleatórias.

2.1 número médio de ligações e grau médio

Um cálculo simples e muito útil ao se estudar redes aleatórias é o cálculo do número de ligações. Este número não é fixo, porém seu valor médio $\langle m \rangle$ (BARABáSI, 2016) pode ser calculado. Para calculá-lo, consideremos que o número médio de ligações entre um único par de sítios é p e sobre todos os $\binom{n}{2}$ pares é

$$\langle m \rangle = \binom{n}{2} p. \tag{2.4}$$

Além disso, pode-se calcular o grau médio da rede. Então, para uma rede de m ligações é $2m/n$ e o grau médio $\langle k \rangle$ de $G(n,p)$ é dado por

$$\langle k \rangle = (n-1)p. \tag{2.5}$$

Este resultado é muito útil, pois se obtém o grau médio apenas com dois parâmetros que são definidos antes de construir a rede. Ou melhor ainda, pode-se ajustar os parâmetros para se conseguir um rede com grau médio desejado. Esta é uma das formas mais comuns de se construir redes.

2.2 Distribuição de graus

Para se analisar a distribuição de graus de $G(n,p)$ (NEWMAN, 2018), tem-se que levar em consideração que cada nó tem probabilidade p de se ligar aos $n-1$ demais nós da rede. Daí, a probabilidade de se ligar a k sítios é dada por $p^k(1-p)^{n-1-k}$. Mas há $\binom{n-1}{k}$ formas de se escolher k sítios, consequentemente a probabilidade de um nó estar a outros k nós é dada por,

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \tag{2.6}$$

No limite em que n é grande, ou seja, $\langle k \rangle << n$, temos

$$p_k = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}, \tag{2.7}$$

Essa distribuição é conhecida como Distribuição de Poisson (KATTI; RAO, 1968). Ela é frequentemente chamada de distribuição de uma rede aleatória.

2.3 Coeficiente de agrupamento

O coeficiente de agrupamento captura o grau dos vizinhos de um nó i estarem conectados entre si (WATTS; STROGATZ, 1998). Para um dado nó i com grau k_i, o coeficiente de agrupamento local é dado por,

$$C_i = \frac{2L_i}{k_i(k_i-1)} \tag{2.8}$$

em que L_i representa o número de ligações entre os k_i vizinhos do nó i. Notemos que C_i representa a probabilidade de haver ligação entre dois nós vizinhos de i. Assim, se $C_i = 50\%$, então metade dos vizinhos do nó i estão conectados com ligações entre si.

O grau de agrupamento de uma rede é capturado pela média do coeficiente de agrupamento $\langle C \rangle$ que representa a média sobre todos C_i, assim

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i. \tag{2.9}$$

Em uma rede aleatória, esta probabilidade é sempre a mesma e já se sabe que é a probabilidade p de dois nós se ligarem, isto é

$$\langle C \rangle = \frac{\langle k \rangle}{n-1}. \tag{2.10}$$

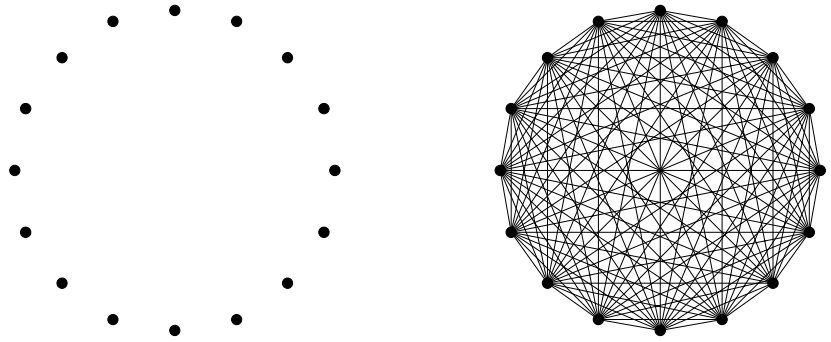
Ao se comparar este resultado com o esperado em redes reais, percebe-se uma grande diferença. Pois, em redes reais o coeficiente de agrupamento é, em geral, grande. Enquanto, se n na Equação (2.10) for grande, C será pequeno.

2.4 Componente gigante

Consideremos uma rede de Poisson $G(n,p)$ com $p = 0$ (Fig. 1A), neste caso cada nó está isolado, assim há n componentes de tamanho 1. Por outro lado, se $p = 1$(Fig. 1B), todos os nós estão ligados entre si e a rede tem apenas uma componente de tamanho n.

Poderia se pensar que a maior componente da rede cresceria de forma gradual $N_G = 1$ até atingir o tamanho da rede toda com $N_G = N$, se $\langle k \rangle$ cresce de 0 a $N-1$. No entanto, o que ocorre é que a fração N_G/N permanece nula para pequenos valores de $\langle k \rangle$. Então, quando $\langle k \rangle$ excede um valor crítico, a fração N_G/N cresce rapidamente sinalizando a emergência de um grande agregado denominado de componente gigante (ERDÖS; RÉNYI, 1959).

Analisando a situação acima descrita, temos que o tamanho S da maior componente passa por uma mudança repentina, chamada de transição de fase. Neste processo, a maior



(A) $p = 0$ (B) $p = 1$

Figura 1 – **Limites de redes aleatórias para** $p = 0$ **e** $p = 1$. **(A)** Quando $p = 0$, há somente ilhas. **(B)** Quando $p = 1$, todas as ligações possível estão presentes, todos os nós pertencem a somente uma componente. Adaptado de (NEWMAN, 2018).

componente passa por uma transição partindo de um tamanho nulo em comparação com a rede como um todo para um tamanho não nulo. A função que dá a fração S da componente gigante em relação à rede é dada por

$$S = 1 - e^{-\langle k \rangle S}. \tag{2.11}$$

O primeiro a estudar essa equação foi Erdös e Rényi em 1959 (ERDÖS; RÉNYI, 1959) e 1960 (ERDOS; RENYI, 1960), ela dá o tamanho da componente gigante em função do tamanho da rede no limite de n grande para qualquer valor de $\langle k \rangle$. Infelizmente, ela não tem solução fácil de forma fechada. No entanto, podemos ter alguma intuição se observarmos o seguinte gráfico da Fig. 2A, em que temos três curvas com distintos valores de $\langle k \rangle$, sendo 0, 1 e $1,5$ e há também a curva $y = S$ representada pela curva diagonal tracejada. Neste gráfico, onde a curva tracejada encontra a curva $y = S$, temos S como solução de (2.11).

Como visto na figura citada, dependendo do valor de $\langle k \rangle$, há uma solução ou duas. Para as duas curvas mais abaixo ($\langle k \rangle = 0$ e $\langle k \rangle = 1$), temos apenas solução em $S = 0$, ou seja, não há componente gigante na rede. Por outro lado, na curva mais acima ($\langle k \rangle) = 1,5$), existe solução para $S = 0$ e $S > 0$, implicando que neste caso há componente gigante. Além disso, a transição entre o regime que não há componente gigante para o regime em que há ocorre na curva de $\langle k \rangle = 1$.

A transição ocorre no momento em que o gradiente da linha tracejada coincide com o da curva em $S = 0$. Ou seja, quando

$$\frac{d}{dS}(1 - e^{-\langle k \rangle S}) = 1, \tag{2.12}$$

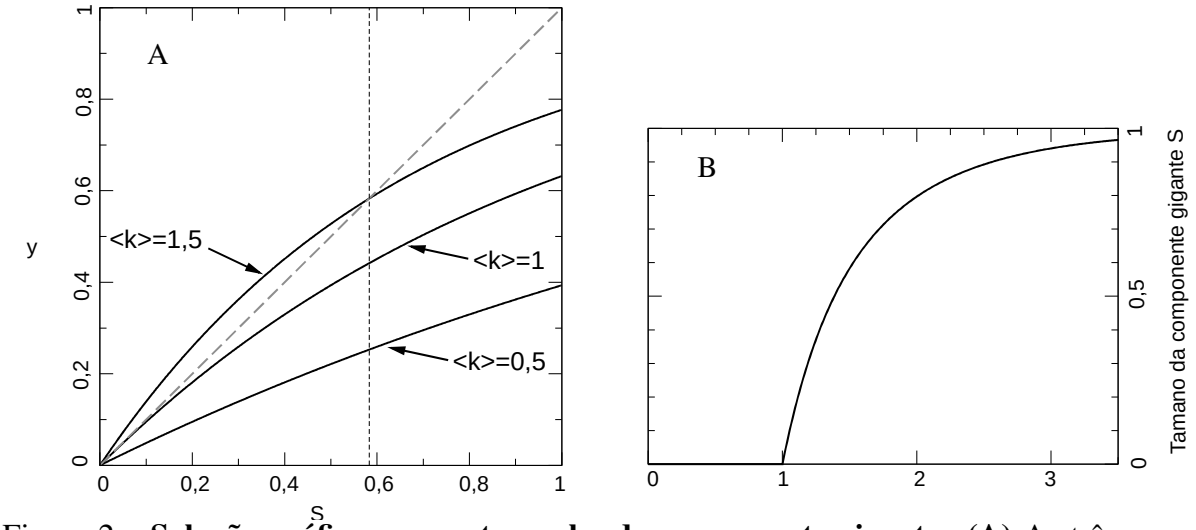


Figura 2 – **Solução gráfica para o tamanho da componente gigante. (A)** As três curvas $y = 1 - e^{\langle k \rangle S}$ para diferentes $\langle k \rangle$. A linha diagonal tracejada mostra $y = S$ e a interseção dá a solução para (2.11). Para a curva mais abaixo, somente há uma interseção em $S = 0$, então não há nenhuma componente gigante. Enquanto para a curva mais acima, há uma solução também marcada pela linha vertical tracejada. Por fim, a curva do meio representa o limiar entre os dois regimes. **(B)** O resultado para o tamanho da componente gigante S como função de $\langle k \rangle$. Figura adaptada de (NEWMAN, 2018).

ou

$$\langle k \rangle e^{-\langle k \rangle S} = 1. \tag{2.13}$$

Se $S = 0$, então a transição ocorre em $\langle k \rangle = 1$. Portanto, uma rede aleatória somente tem componente gigante, se $\langle k \rangle > 1$ e para valores abaixo de $\langle k \rangle \leq 1$, não há.

2.5 Comprimento do caminho

Para iniciar esta seção, é importante ter em mente algumas definições. A primeira é a definição de passeio (walk) em uma rede que é a sequência de ligações conectadas formando uma rota contínua na rede. Além disso, os passeios podem ser de três tipos, um trilha (trail) que é um passeio que não passa por nenhuma ligação mais que uma vez. Um caminho (path) que é um passeio que não passa por nenhum nó mais que uma vez. Por fim, um ciclo (cycle) que é um passeio que começa e termina no mesmo nó sem passar por nenhum nó mais que uma vez (CRAMER *et al.*, 2018).

Tendo essas definições, dizemos que o tamanho do caminho é a quantidade de ligações do caminho. Por sua vez, o menor comprimento de caminho entre dois nós é chamado de distância. Ademais, o diâmetro de uma rede é a maior distância entre dois nós de uma mesma componente. Estas últimas definições estão exemplificadas na Fig. 3. Em que, na Fig. 3A estão

representados em verde as ligações e em laranja um caminho que começa no nó 1 e termina no nó 6. Na Fig. 3B, temos em laranja e em marrom a distância entre os nós 1 e 7. Assim, percebemos que pode haver mais de um caminho que meça a distância entre dois nós. Além disso, o diâmetro d_{max} desta rede é 3.

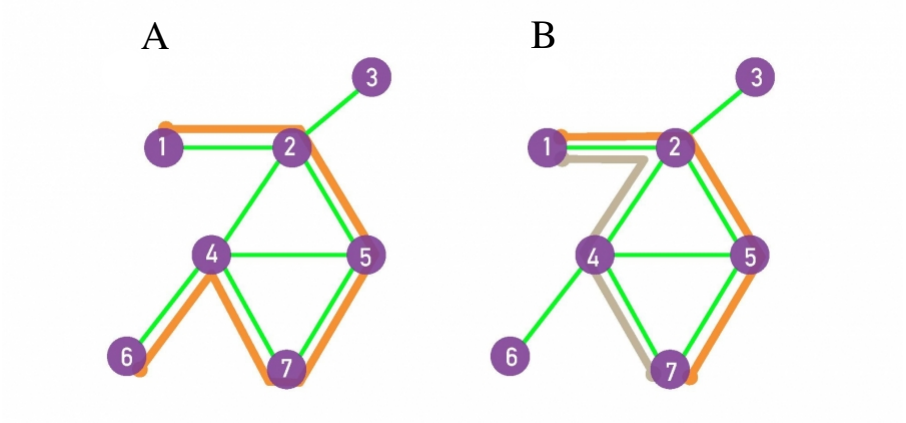


Figura 3 – **Caminhos. (A)** As ligações estão representadas em verde e em laranja está representado um caminho que sai do nó 1 e chega no nó 6. **(B)** Neste caso, temos em laranja e em marrom a distância entre os nós 1 e 7. Vemos que pode haver mais de um caminho que meça a distância entre dois nós. Por fim, o diâmetro d_{max} desta rede é 3, assim $d_{max} = 3$. Figura adaptada de (BARABáSI, 2016).

Pode-se estudar o diâmetro da seguinte forma. Dado um conjunto de nós de uma rede aleatória, imagine que será adicionado novos vizinhos a ele várias vezes. O número de vizinhos aumenta $\langle k \rangle$ vezes em média para cada passo. Daí iniciando esse processo com apenas um nó, com o tempo se completará toda rede. Assim, depois de s passos tem-se o tamanho da rede, aproximadamente, $\langle k \rangle^s \simeq n$ ou escrevendo de outra forma, $s = \ln(n)/\ln(\langle k \rangle)$. Grosso modo, como todos os nós estão dentro dos s pontos do ponto de partida dado, é implicado que o diâmetro da rede é aproximadamente $\ln(n)/\ln(\langle k \rangle)$.

O argumento seguido no parágrafo anterior foi somente uma aproximação. Tendo em vista que, pensando melhor não se está calculando um diâmetro, mas sim um “raio". Pois chega-se pelo processo descrito a cima à distância máxima de um de um nó que é ponto de partida a outro nó da rede e não à distância máxima da rede em si. Além disso, embora haja $\langle k \rangle^s$ nós decorridos s passos, este resultado não funciona à medida que $\langle k \rangle^s$ se torna comparável a n, pois significaria que o número de nós com distância s seria maior que o número de nós de toda a rede. O resultado correto desenvolvido, por exemplo, em (NEWMAN, 2018) para o diâmetro l

da rede é dado por

$$l = \frac{\ln(n)}{\ln(\langle k \rangle)}. \tag{2.14}$$

Ao observar a equação acima, percebe-se que o diâmetro da rede cresce somente com $\ln(n)$, ou seja, cresce bem mais devagar do que o tamanho da rede em si. Este fato, oferece uma explicação para o fenômeno do mundo pequeno. Dado que, por exemplo, se n for o número aproximado de pessoas no planeta Terra e que cada pessoa conhece 1.000 pessoas, tem-se que

$$l = \frac{\ln(n)}{\ln(\langle k \rangle)} = \frac{\ln(7{,}7x10^9)}{\ln(1.000)} = 3,29...,$$

este resultado é pequeno o suficiente para explicar os experimentos de mundo pequeno de (DODDS *et al.*, 2003), (MILGRAM, 1967) e (TRAVERS; MILGRAM, 1977). Na prática, a equação (2.14) é um bom guia para alguns comportamentos em redes reais. Por exemplo, na Fig. 4 tem-se as distâncias médias dos menores caminhos no Facebook entre estudantes de 100 universidades nos Estados Unidos (JACOBS *et al.*, 2015) ajustadas como função de $\ln(n)$. Pelos dados do gráfico, obtém-se, aproximadamente, uma linha reta o que segue o resultado da Eq. (2.14). Obviamente os resultados não são totalmente iguais, porém obtém-se algumas intuições ao se estudar redes aleatórias e compará-las a redes reais.

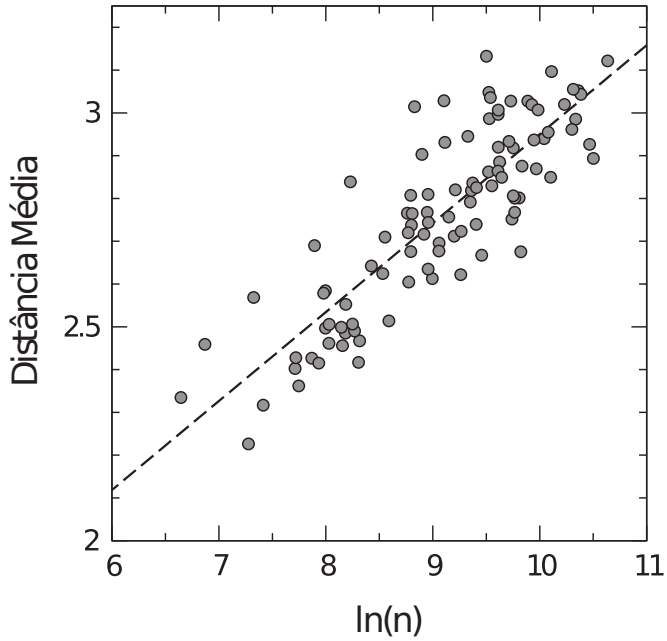


Figura 4 – **Distância média do caminho mais curto em uma rede de amizades no Facebook.** Os pontos representam as distâncias médias na rede de amizades do Facebook de estudantes de diferentes universidades do EUA como função de $\ln(n)$ e a linha tracejada é o melhor ajuste para esses pontos. Figura adaptada de (NEWMAN, 2018).

2.6 Problemas com redes aleatórias

A rede de Poisson é um dos modelos mais estudados e dele obtiveram-se várias noções de como se comportam as estruturas de outros modelos de redes tais como seus tamanhos e diâmetros particularmente. No entanto, redes aleatórias possuem vários problemas se o objetivo é compará-la com uma rede real. Pode-se citar, a título de exemplo, o agrupamento e a transitividade em redes aleatórias que é bastante diferente se comparados a redes reais. O coeficiente de agrupamento (2.10) tende a zero para n grande em redes aleatórias. Portanto, ao se pensar na população do planeta e supor que $\langle k \rangle = 1.000$, tem-se

$$C = \frac{\langle k \rangle}{n-1} \Rightarrow C = \frac{1000}{7 \cdot 10^9 - 1} = 1,29 \cdot 10^{-7}.$$

Este valor é bem distante do valor provável que está entre $0,1$ e $0,5$. Outra grande diferença está no fato de não haver correlação entre os graus de nós adjacentes, isso vem do fato das ligações serem aleatórias. No entanto, para redes reais há correlação. Além disso, em muitas redes há o fenômeno da criação de grupos ou comunidades que será estudado mais à frente.

Por fim, provavelmente a maior diferença está na distribuição de graus. Em redes reais há uma tendência para distribuição assimétrica à direita com muito nós tendo graus baixos e poucos "hubs" com alto grau, assim tendo graus altos na calda da distribuição. Ao contrário da distribuição de Poisson como pode ser visto na Fig. 5 para o exemplo de rede real a distribuição de internet no nível de sistemas autônomos.

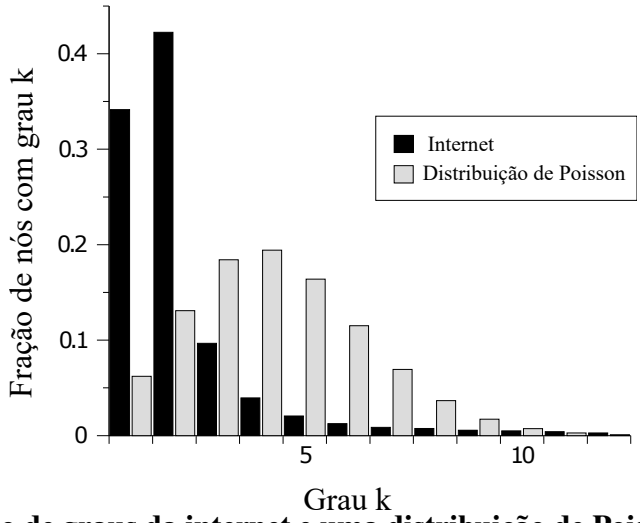


Figura 5 – **Distribuição de graus da internet e uma distribuição de Poisson.** As barras escuras representam as frações de nós com dado grau em uma representação da rede da internet no nível de sistemas autônomos. As barras claras representam a mesma medida para uma rede aleatória com mesmo grau médio da internet. Embora as distribuições tenham as mesmas médias, é claro que elas são inteiramente diferentes no formato. Figura adaptada de (NEWMAN, 2018).

2.7 Matriz de adjacência

A matriz de adjacência **A** de uma rede é a representação fundamental de uma rede e esta seção está dedicada ao seu estudo. Primeiramente, temos que esclarecer que pode haver mais de uma ligação entre nós. Assim, se há mais de uma ligação entre dos nós, o conjunto é chamado de ligação múltipla (multiple edge). Além disso, ligações que saem de um nó e chegam no mesmo nó são chamadas de autoligações (self-edge). Temos na Fig. 6A uma rede com apenas ligações simples, ou seja, sem ligações múltiplas ou autoligações. E na Fig. 6B está representada uma rede em que há não somente ligações simples, mas também ligações múltiplas e autoligações.

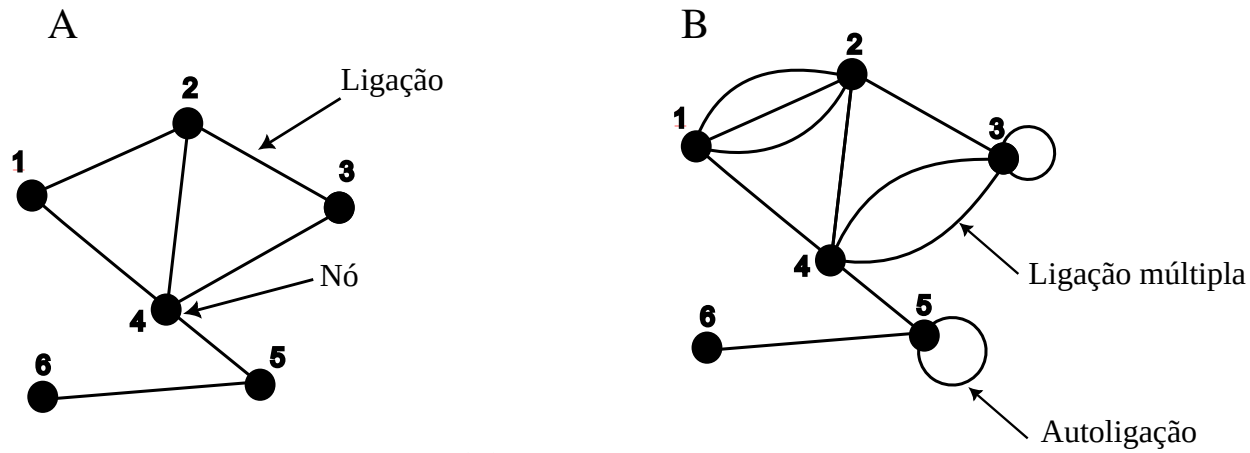


Figura 6 – **Duas redes pequenas. (A)** Uma rede somente com ligações simples, sem ligações múltiplas nem autoligações. **(B)** Uma rede com ligações múltiplas e autoligações. Figura de autoria própria.

Para a rede somente com ligações simples, vamos definir a matrize de adjacência da seguinte forma:

$$A_{ij} = \begin{cases} 1, & \textit{se há uma ligação entre os nós } i \textit{ e } j, \\ 0, & \textit{caso contrário}. \end{cases} \tag{2.15}$$

Por exemplo, para a Fig. 6A ficamos com a seguinte matriz representada na em (2.16)

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{2.16}$$

Vale a pena observar que para o caso de uma rede que só há ligações simples, a matriz de adjacência é simétrica e todos os termos da diagonal são nulos.

No entanto, também, conseguimos representar uma matriz com autoligações e ligações múltiplas com uma matriz de adjacência. Assim, ligações múltiplas são representadas pela multiplicidade de ligações. Por exemplo, se há 3 ligações entre os nós i e j, então $A_{ij} = 3$. Ademais, os termos A_{ii} representam as autoligações e cada autoligação é representada pelo número 2. Portanto, se há uma autoligação no nó i, temos o termo $A_{ii} = 2$, se há duas autoligações, $A_{ii} = 4$ e assim sucessivamente. Daí, a rede da Fig. 6B é representada por (2.17)

$$\begin{pmatrix} 0 & 3 & 0 & 1 & 0 & 0 \\ 3 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 2 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{2.17}$$

2.8 Grupos de nós

Muitas redes se dividem em grupos ou comunidades. Tem-se como exemplo, redes de pessoas que se dividem em grupos de amigos, colegas de trabalho, torcedores de uma mesmo time de futebol dentre outros, ou grupos funcionais de compostos orgânicos em Química Orgânica. No próximo capítulo será mais aprofundado este tema, mas antes disso, nesta seção, estudaremos alguns conceitos básicos sobre grupos em redes e veremos os conceitos de cliques, k-cores e k-cliques.

2.8.1 Cliques

Um clique é um conjunto de nós conectados entre si com ligações não direcionadas. Daí, se um grupo de 4 nós está todo ligado entre si, ou seja, cada nó tem 3 ligações com nós distintos, então o conjunto é considerado um clique. A ocorrência deste grupo em uma rede esparsa indica alta coesão do grupo. Ou seja, o grupo fortemente conectado. Podemos imaginar, por exemplo, uma família ou colegas de trabalho. Um exemplo pode ser visto na Fig.7.

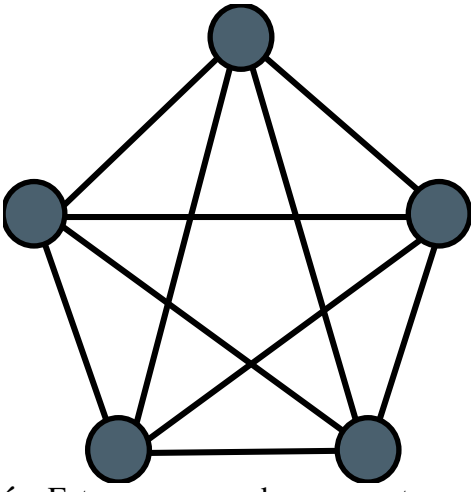


Figura 7 – **Um clique de 5 nós.** Esta pequena rede representa um clique com 5 nós. Figura de autoria própria.

2.8.2 Cores

Tendo em vista que o conceito de clique pode ser muito restritivo, surgiu o conceito mais flexível de k-core. Neste grupo, cada nó deve estar conectado a pelo menos k outros nós, como visto na Fig. 8 . Este conceito é bastante útil pela facilidade de ser detectado.

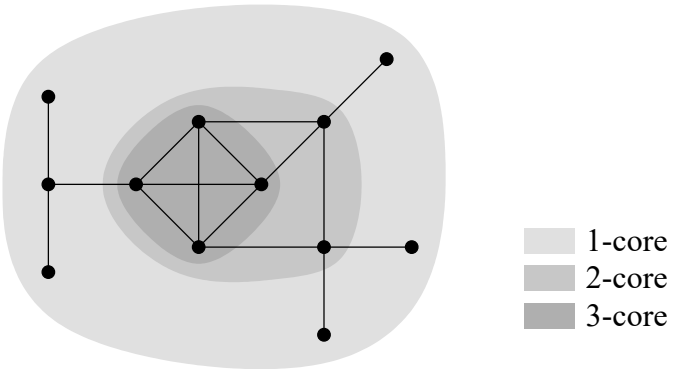


Figura 8 – **Os k-cores em uma rede pequena.** A regiões sombreadas denotam os k-cores para $k = 1, 2$ e 3 nesta rede. Não há k-cores para $k > 3$. Note como os k-cores estão um dentro do outro, o 3-core está dentro do 2-core que por sua vez está dentro do 1-core. Figura adaptada de (NEWMAN, 2018).

Uma forma simples de detectar um k-core é removendo todos os nós da rede que tenham grau menor que k. Assim, restará apenas cores com nós menores ou iguais a k. Obviamente, se a intenção é ter apenas os k-cores em que os nós tenham somente grau k, então pode-se remover também os demais nós com maior grau que k.

2.8.3 Componentes e k-componentes

A componente de uma rede não direcionada é definida como um conjunto de nós em que há caminhos entre todos os nós. Disto pode-se definir o conceito de k-componente que é um conjunto de nós aos quais há pelo menos k caminhos entre quaisquer pares de nós, como pode ser visto na Fig.9.

Sabendo deste novo conceito, pode-se agora analisar em uma rede seus k-cores e k-componentes como na Fig.10 em que há apenas um 2-core e duas 2-componentes. Portanto, cada nó é conectado a pelo menos 2 outros, mas há duas 2-componentes representadas pelas áreas sombreadas onde há pelo menos dois caminhos que ligam cada nó e essas 2-componentes se ligam por um caminho independente. A noção de k-componente é um jeito natural de analisar a robustez da rede. Por exemplo, se pensarmos na rede de internet, o número de caminhos entres dois nós pode representar duas rotas que dados podem percorrer.

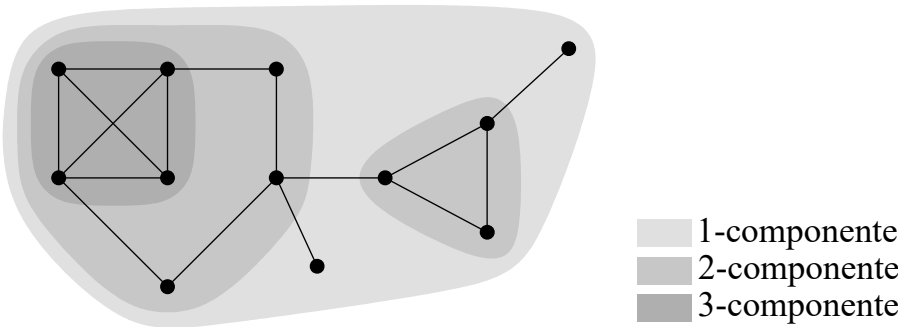


Figura 9 – **k-componentes em uma rede pequena.** As regiões sombreadas denotam as k-componentes na rede que possui uma 1-componente, duas 2-componentes e uma 3-componentes. Note que k-componentes estão umas dentro das outras. As 2-componentes estão dentro da 1-componente e a 3-componente estão dentro de umas das duas 2-componentes. Figura adaptada de (NEWMAN, 2018).

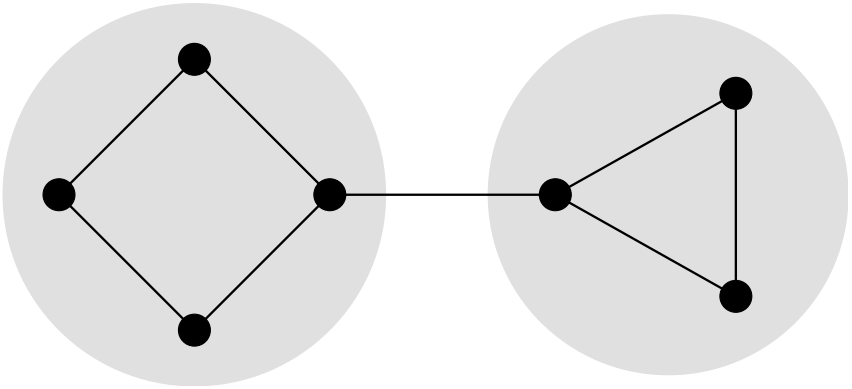


Figura 10 – **Uma rede pequena com um 2-cores e duas 2-componentes.** Há nesta rede um único 2-core, pois os todos os nós se ligam a pelo menos dois outros nós. Porém há duas 2-componentes separadas como indicada pelos círculos sombreados, provando que 2-cores e 2-componentes não são a mesma coisa. Figura adaptada de (NEWMAN, 2018).

3 ESTRUTURA DE COMUNIDADE

Redes representam sistemas reais que não são regulares. Estes sistemas são objetos em que há coexistência tanto de ordem quanto desordem. O paradigma de desordem é a rede aleatória em que a probabilidade de haver uma ligação entre qualquer par de nós é igual para todos os pares. Assim, em redes aleatórias a distribuição de ligações entre os nós é bastante homogênea. No entanto, redes reais não são redes aleatórias, pois redes reais apresentam grande heterogeneidade, o que revela um alto nível de ordem. Neste tipo de sistema, coexistem muitos nós com baixo grau e poucos nós com alto grau. Portanto, a distribuição de ligações não é somente heterogênea globalmente, mas também, localmente tendo grande concentração de ligações em grupos especiais e baixa concentração entre esses grupos. Esta característica de redes reais denominadas de estruturas de comunidades (GIRVAN; NEWMAN, 2002) ou agrupamento (clustering). Comunidades, também chamadas de grupos ou módulos, são grupos de nós que, provavelmente, compartilham propriedades em comum ou fazem papéis similares na rede (FORTUNATO, 2009).

Comunidades podem ter aplicações concretas. Por exemplo, ao se agrupar clientes da Web que têm interesses similares e estão geograficamente próximos, pode haver uma melhora na performance dos serviços fornecidos na World Wide Web (KRISHNAMURTHY; WANG, 2000). Outra aplicação é a identificação de grupos de usuários com interesses similares na rede das relações de clientes e produtos de uma empresa varejista como a Amazon para configurar um sistema que de forma eficiente recomende produtos (REDDY *et al.*, 2002). Podemos pensar de forma parecida para indicações de músicas ou podcasts no Spotify. Um último exemplo seria agrupar os nós para se permitir gerar tabelas de rotas compactas para escolher caminhos de comunicação eficientementes (STEENSTRUP, 2001).

Além das aplicações já citadas o problema denominado de detecção de comunidade (community detection) serve também para identificar grupos e seus limites a fim de classificá-los de acordo com a posição estrutural nos grupos. Assim, nós com uma posição central no grupo, por exemplo, compartilhando grande número de ligações com os outros nós do grupo podem ter papel importante de controle e estabilidade do grupo. Por outro lado, nós que estão nos limites entre grupos, fazem um importante papel de mediação e leva interações entre diferentes comunidades (CSERMELY, 2008).

Existem diversos métodos de detecção de comunidade e o objetivo deste capítulo é mostrar somente a abordagem de inferência estatística para detecção de comunidade. Além

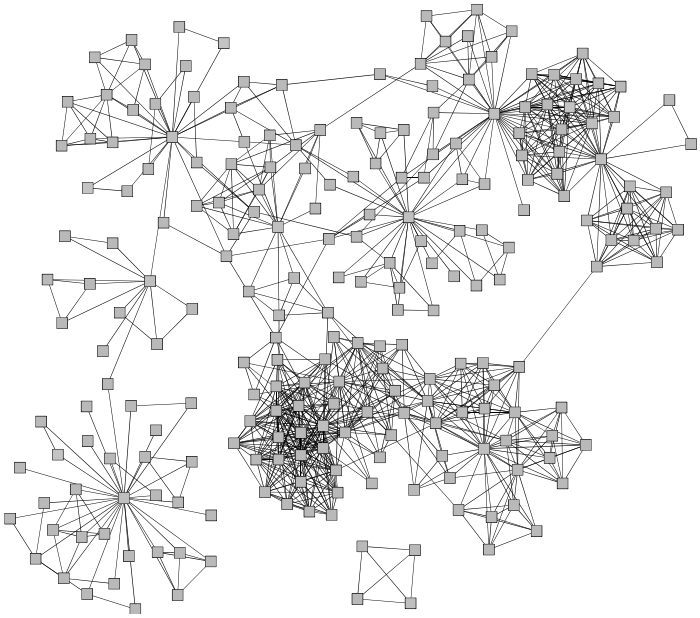


Figura 11 – **Rede de coautoria em um departamento de universidade.** Os nós, nesta rede, representam cientistas em um departamento de universidade e as ligações representam coautorias em artigos científicos (papers). Esta rede tem claramente estruturas de comunidade, presumivelmente refletindo a divisão de interesses e grupos de pesquisas. Figura adaptada de (NEWMAN, 2018).

disso, explicar sobre o método de maximização de likelihood para ajustes (fits).

3.1 Dividindo em grupos

Na Fig. 11 está representada a rede de colaboração de pesquisadores de um mesmo departamento de uma universidade. Na figura, cada nó representa um pesquisador e as ligações são formadas se dois pesquisadores são autores de um mesmo artigo científico (paper). Como somos conhecedores da organização de um departamento de universidade, percebemos rapidamente ser muito provável que os conjuntos de nós que possuem mais ligações entre si representem grupos de pesquisa ou laboratórios.

No entanto, agora podemos imaginar que não conhecemos essa estrutura de grupos de pesquisa ou que não sabemos sobre o que se trata a rede. Isto é, temos a rede em si em que sabemos os nós e as ligações com está representado na Fig. 11 e o objetivo é analisar a estrutura da rede e descobrir se ela se divide em grupos e quais grupos são esses. Ou seja, temos um problema de detecção de comunidades.

A rede na Fig. 11 é muito pequena, porém redes podem ter milhões de nós o que torna o seu estudo bem complicado. Assim, uma possível solução para este problema é dividir a rede em grupos de forma a reduzi-la a um tamanho mais manejável. Como exemplo disso,

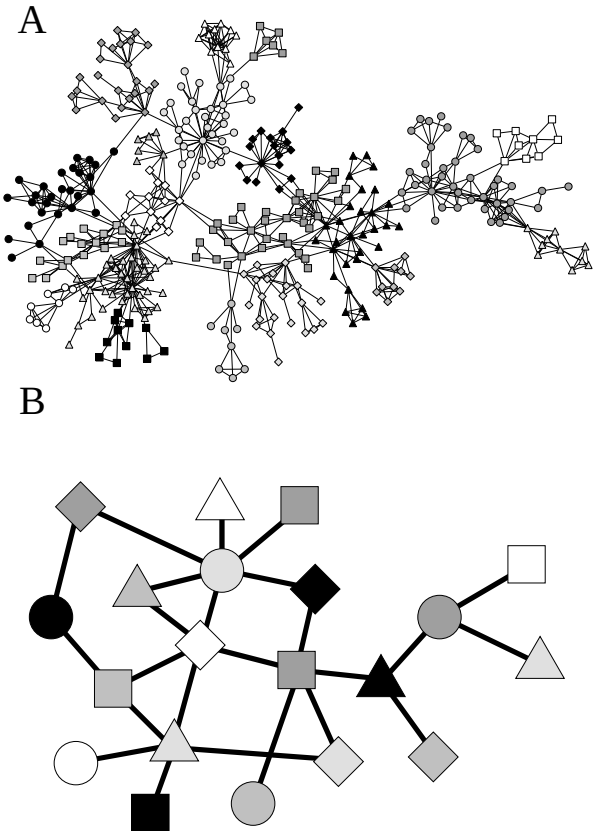


Figura 12 – **Visualização de estrutura de rede usando detecção de comunidade.** A rede em **(A)** está decomposta em suas comunidades constituintes. Em **(B)** cada comunidade na rede é representada por um único nó e as ligações indicam quais comunidades estão conectadas. Figura adaptada de (NEWMAN, 2018).

tem-se que a visualização pode ser possível para uma rede de milhões de nós se ela for dividida em grupos e cada grupo for representado por um nó apenas, bem como as conexões intergrupos como ligações, conforme é mostrado na Fig. 12 em que temos na Fig. 12A uma rede dividida em grupos representados por diferentes tonalidades e na Fig. 12B os mesmos grupos estão agora representados apenas por um nó e as ligações entre os grupos são representadas apenas por uma ligação.

3.2 Métodos baseados em inferência estatística

Alguns dos mais poderosos e flexíveis métodos para detecção de comunidades são baseados em inferência estatística que consiste em ajustar um modelo de rede para uma dada rede observada. Funciona da seguinte maneira, dado um modelo de rede, ou seja, qualquer processo que pode gerar uma rede, podemos ajustar aquele modelo para os dados de forma a encontrar os valores dos parâmetros que fornecem a likelihood máxima.

3.2.1 Apresentando o conceito de likelihood

Aqui vale uma pausa para comentar sobre o conceito de likelihood que pode ser encontrado por exemplo em (CASELLA; BERGER, 2021). Para isso, suponhamos que fizemos N medidas e denotamos os resultados por $x_1,..,x_N$. Além disso, consideremos que são medidas independentes e distribuídas aleatoriamente em torno do valor verdadeiro z seguindo uma distribuição normal com desvio padrão σ. Ademais, como as medidas são independentes, a probabilidade de todas coletivamente é dada por

$$P(x_1,...,x_N|z,\sigma) = \prod_{i=1}^{N} P(x_i|z,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-z)^2/2\sigma^2}. \tag{3.1}$$

Essa probabilidade é conhecida como likelihood de x_i dado z e σ. Dada esta expressão, podemos então nos perguntar quais os valores de z e σ são mais prováveis para os dados. A probabilidade de que valores específicos de z e σ produzir x_i pode ser calculada usando o teorema de Bayes também chamado de regra de Bayes (JOYCE, 2021):

$$P(z,\sigma|x_1,...,x_N) = P(x_1,...,x_N|z,\sigma)\frac{P(z)P(\sigma)}{P(x_1,...,x_N)}, \tag{3.2}$$

em que $P(z)$, $P(\sigma)$ e $P(x_1,...,x_N)$ são as chamadas probabilidade a priori de z, σ e dos dados respectivamente. Estas são as probabilidades de valores particulares de z, σ e $x_1,...,x_N$ se não se sabe o dado x_i.

Os mais prováveis valores de σ e z são, por definição, aqueles que $P(z,\sigma|x_1,...,x_N)$ é máxima, isto é, pode-se encontrar maximizando a Equação (3.2) com $x_1,...,x_N$ mantido fixos nos valores observados. Se os x_i's são fixos, $P(x_1,...,x_N)$ é constante e comumente podemos assumir que $P(z)$ e $P(\sigma)$ também são. Assim, ficando com

$$P(z,\sigma|x_1,...,x_N) \propto P(x_1,...,x_N|z,\sigma). \tag{3.3}$$

Portanto, para obtermos a probabilidade da esquerda da equação acima, basta maximizar a likelihood, pois o ponto de máximo das duas probabilidades é o mesmo. Isto pode ser feito simplesmente derivando em relação aos termos que se deseja encontrar e igualando a derivada a zero. É comumente procurado o máximo do logaritmo da likelihood que é chamado de log-likelihood, pois em geral é mais fácil de ser encontrado.

Vale a pena fazermos um adendo importante. Apesar da likelihood também ser uma probabilidade, denominamos ela por outro termo para diferenciar. Pois probabilidade refere-se a um chance particular de um evento ocorrer baseado nos parâmetros do modelo. Por outro lado,

a likelihood refere-se a quão bem uma amostra fornece suporte para valores específicos de um parâmetro em um modelo. Assim, quando calculamos a probabilidade de algum evento ocorrer, assumimos que os parâmetros de um modelo são confiáveis. No entanto, quando calculamos a likelihood, estamos tentando determinar se podemos confiar nos parâmetros de um modelo com base nos dados de amostra que observamos.

Como exemplo do uso da likelihood, podemos tomar como exemplo uma rede de Poisson de tamanho n e probabilidade de dois nós se ligarem p. A likelihood dessa rede definida pela matriz adjascente $\mathbf{A}$ é

$$P(\mathbf{A}|p) = p^m(1-p)^{\binom{n}{2}-m}, \tag{3.4}$$

em que m é o número de ligações. Porém, supondo que não se conhece p e tudo que se tem é apenas a rede em si. Podemos fazer uma estimativa de p empregando a regra de Bayes:

$$P(p|\mathbf{A}) = P(\mathbf{A}|p)\frac{P(p)}{P(\mathbf{A})}, \tag{3.5}$$

em que $P(p)$ e $P(\mathbf{A})$ são as probabilidades a priori de p e $\mathbf{A}$ respectivamente. O valor mais provável de p é agora, por definição, dado pela maximização dessa expressão com respeito a p enquanto $\mathbf{A}$ é o valor observado. Porém se $\mathbf{A}$ é constante, então o denominador não tem efeito e normalmente, se assume $P(p)$ também constante, isto é, todos os valores de p são igualmente prováveis. Assim, maximizar $P(p|\mathbf{A})$ é o mesmo que maximizar o likelihood $P(\mathbf{A}|p)$.

Daí, aplicando a derivada para maximizar (3.4), obtemos,

$$p = \frac{m}{\binom{n}{2}}. \tag{3.6}$$

Por fim, como é mais comum é calculado o logaritmo de (3.4). Assim, maximizando a seguinte equação:

$$\log(P(\mathbf{A}|p)) = m\log(p) + \left[\binom{n}{2} - m\right]\log(1-p) \tag{3.7}$$

3.2.2 Detecção de comunidade usando inferência estatística

Podemos usar o método de maximização da likelihood para detectar comunidades ajustando os dados da rede a um modelo que contém a estrutura da comunidade. O modelo a ser utilizado é o modelo de bloco estocástico com grau corrigido (degree-corrected stochastic block model) (YAN *et al.*, 2014) que é feito de uma forma um pouco diferente do modelo bloco estocástico ordinário (HOLLAND *et al.*, 1983) e (LEE; WILKINSON, 2019). No modelo,

que iremos trabalhar divide-se os nós em q grupos, em seguida, coloca-se as ligações com probabilidade $\omega_{g_i g_j} c_i c_j / 2m$, em que g_i e g_j são os grupo dos nós i e j. Bem como c_i é o grau desejado do nó i e a matriz q x q de parâmetros w_{rs} controla a estrutura da comunidade. Por exemplo, se w_{rr} é bem maior que os elementos fora da diagonal, então haverá muito mais ligações intragrupos que intergrupos. Isto é, já se supõe que a rede a ser gerada para o ajuste é dividida em comunidades.

Além disso, se c_i é igual graus dos nós em média, então

$$\sum_j \omega_{r g_j} c_j = 2m, \tag{3.8}$$

reescrevendo,

$$\sum_j \omega_{r g_j} c_j = \sum_s \omega_{rs} \kappa_s, \tag{3.9}$$

em que

$$\kappa_s = \sum_j \delta_{g_j s} c_j \tag{3.10}$$

que é a soma dos graus c_j de todos os nós do grupo s. Daí, ficando com

$$\sum_s \omega_{rs} \kappa_s = 2m. \tag{3.11}$$

Este resultado define completamente o modelo de bloco estocástico com grau corrigido. De fato, quando usado para detecção de comunidades o modelo é estudado adicionando ligações entre nós seguindo uma distribuição de Poisson com média $w_{g_i g_j} c_i c_j / 2m$ ou metade disso quando $i = j$.

Para este modelo, temos a likelihood de uma rede de matriz de adjacência $\mathbf{A}$ gerada por bloco estocástico de grau corrigido é dada por:

$$P(\mathbf{A}|\Omega, \mathbf{c}, \mathbf{g}) = \prod_{i<j} \frac{(\omega_{g_i g_j} c_i c_j / 2m)^{A_{ij}}}{A_{ij}!} e^{-\omega_{g_i g_j} c_i c_j / 2m} \times \prod_i \frac{(\omega_{g_i g_i} c_i^2 / 4m)^{A_{ii}/2}}{(\frac{1}{2} A_{ii})!} e^{-\omega_{g_i g_i} c_i^2 / 4m}, \tag{3.12}$$

em que a matriz Ω possuem os elementos de w_{rs}, $\mathbf{g}$ possui os elementos dos grupos g_i do qual os nós pertencem e $\mathbf{c}$ os graus do nós. O primeiro produtório é sobre os termos sem autoligações enquanto o outro produtório contém termos com autoligações. Como de costume vamos usar o logaritmo,

$$\begin{aligned} \log(P(\mathbf{A}|\Omega, \mathbf{c}, \mathbf{g})) = & \sum_{i<j} \left[A_{ij} \log\left(\frac{\omega_{g_i g_j} c_i c_j}{2m} \right) - \log\left(A_{ij}!\right) - \frac{\omega_{g_i g_j} c_i c_j}{2m} \right] \\ & + \sum_i \left[\tfrac{1}{2} A_{ii} \log\left(\frac{\omega_{g_i g_i} c_i^2}{4m} \right) - \log(\tfrac{1}{2} A_{ii}!) - \frac{\omega_{g_i g_i} c_i^2}{4m} \right]. \end{aligned} \tag{3.13}$$

Rearranjando,

$$\log(P(\mathbf{A}|\Omega,\mathbf{k},\mathbf{g})) = \frac{1}{2}\sum_{ij}\left[A_{ij}\log\left(\frac{\omega_{g_ig_j}c_ic_j}{2m}\right) - \frac{\omega_{g_ig_j}c_ic_j}{2m}\right] + constantes. \tag{3.14}$$

Além disso, chamando

$$m_{rs} = \sum_{ij}\delta_{g_ir}\delta_{g_js}A_{ij}, \tag{3.15}$$

obtém-se,

$$\log(P(\mathbf{A}|\Omega,\mathbf{k},\mathbf{g})) = \sum_i k_i\log(c_i) + \frac{1}{2}\sum_{rs}m_{rs}\log(\omega_{rs}) - \frac{1}{2}\sum_{ij}\frac{\omega_{g_ig_j}c_ic_j}{2m} + constantes. \tag{3.16}$$

Com esta expressão, podemos calcular os melhores parâmetros do ajuste ao maximizá-la.

Dando continuidade, derivando (3.16) com relação a c_i e igualando a zero,

$$\frac{\partial\log(P)}{\partial c_i} = \frac{k_i}{c_i} - \sum_j\frac{\omega_{g_ig_j}c_j}{2m} = \frac{k_i}{c_i} - 1, \tag{3.17}$$

chegamos em $c_i = k_i$, ou seja, o melhor valor esperado do conjunto dos graus c_i é igual ao graus k_i observados. Em seguida para derivar com relação à w_{rs}, inicialmente, reescreve-se

$$\sum_{ij}\frac{\omega_{g_ig_j}c_ic_j}{2m} = \sum_{ijrs}\delta_{g_ir}\delta_{g_js}\frac{\omega_{rs}c_ic_j}{2m} = \sum_{rs}\frac{\omega_{rs}\kappa_r\kappa_s}{2m} \tag{3.18}$$

em que κ_r e κ_s são os somatórios de c_i no grupos r e s, com este resultado e derivando com relação à ω_{rs} e igualando a zero. Tem-se,

$$\omega_{rs} = 2m\frac{m_{rs}}{\kappa_r\kappa_s}. \tag{3.19}$$

Assim, tendo obtido os melhores valores de c_i e ω_{rs} e substituindo em (3.16). Obtém-se a chamada profile likelihood:

$$\mathscr{L} = \frac{1}{2}\sum_{rs}m_{rs}\log\left(\frac{m_{rs}}{\kappa_r\kappa_s}\right) + constantes. \tag{3.20}$$

Essa é a expressão fundamental para detecção de comunidades usando maximização da likelihood. Como a equação já está maximizada com respeito aos parâmetros $\mathbf{c}$ e Ω, então resta apenas maximizar com relação à $\mathbf{g}$ onde entra as expressões m_{rs}, κ_r e κ_s que são substituídos em (3.20) para então ter $\mathscr{L}$ que maximiza todos os valores atribuídos.

Por fim, é importante salientar que apesar deste método ter como fraqueza a suposição de que a rede foi formada pelo modelo de bloco estocástico com grau corrigido e a desvantagem vinda do fato de que precisa-se especificar o número de grupos q. Para várias redes formadas

por outras abordagens, o método de maximização da likelihood dá resultados excelentes na prática. Assim apresentando resultados melhores que os dos métodos vistos até então. E embora tenha falhas para estruturas muito fracas (DECELLE *et al.*, 2011a), (DECELLE *et al.*, 2011b) foi provado por (BICKEL; CHEN, 2009) que chega a ser consistente assintoticamente. Isto é, consegue identificar comunidades em redes grandes já conhecidas.

4 ENUMERAÇÃO EFETIVA DE CAMINHOS

Neste capítulo encontra-se a nossa contribuição direta neste trabalho. Tal trabalho dedica-se ao estudo da enumeração efetiva dos caminhos que aglutinam pares de nós. Assim, pode-se pensar como uma extensão da matriz de adjacência inspirado por exemplo em (ANDRADE *et al.*, 2006) e (ANDRADE *et al.*, 2008). Tendo em vista que a matriz de adjacência apenas registra ligações diretas entre nós, ou seja, caminhos entre dois nós com uma ligação apenas. Por outro lado, neste trabalho se almeja encontrar caminhos entre pares de nós que, por sua vez, podem ter mais de uma ligação. Portanto, sendo mais geral que a matriz de adjacência convencional. Por fim, pretendíamos aperfeiçoar e adaptar os resultados desenvolvidos até aqui por nós a fim de futuramente usá-los como uma tentativa de detectar comunidades em redes. Contudo, ao tentar aplicar em redes divididas em comunidades, não obtivemos êxito.

4.1 Abordando o problema

Para dar início, definimos um conceito que chamamos de instante de aglutinação. Para determiná-lo, usamos o seguinte processo. Partindo de uma rede qualquer, supomos que todas as ligações são removidas e escolhemos um par de nós para analisar. Depois disso reconstruímos a rede recolocando as ligações uma a uma aleatoriamente. Note que, da forma como fazemos, a cada passo escolhemos uma das ligações da rede para ser incluída na reconstrução. Não há impedimento de que a mesma ligação seja escolhida mais de uma vez. Evidentemente, incluir uma ligação já presente não altera a reconstrução da rede. Definimos n como sendo o número de ligações incluídas na reconstrução e definimos o instante de aglutinação como sendo o número de ligações incluídas até que o par de nós analisados se aglutinem no mesmo agregado.

Assim, estudemos a seguinte situação esboçada na Fig. 13A em que temos uma rede aleatória e na Fig. 13C o gráfico do logaritmo da distribuição da probabilidade de dois nós escolhidos não se aglutinarem depois de incluídas n ligações. Na Fig. 13B, temos uma rede dividida em duas comunidades bem evidentes, ou seja, os grupos possuem nós densamente conectados entre si e há poucas ligações que partem de um nó de um grupo e chegam em um nó de outro grupo. Também fizemos o gráfico do logaritmo da probabilidade de não se aglutinarem depois de n ligações incluídas como pode ser observado na Fig. 13D. Assim, percebemos que nos dois casos as probabilidades decaem de forma aparentemente similar formando uma curva que temos o interesse de estudar.

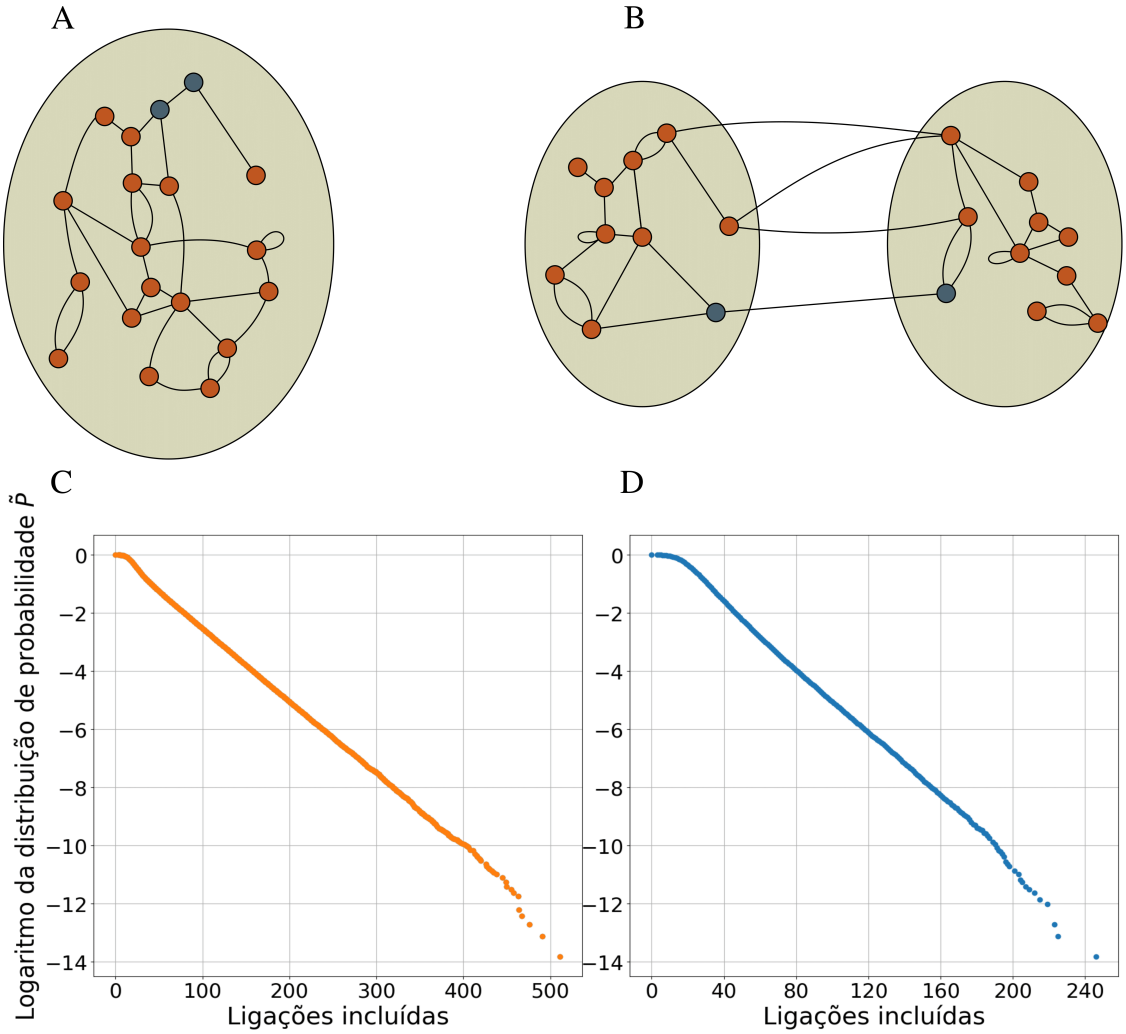


Figura 13 – **Rede aleatória e rede dividida em dois grupos.** Nesta figura, esboçamos os dois sistemas. **(A)** Uma rede aleatória convencional com a possibilidade de múltiplas ligações entre nós e autoligações. **(B)** Uma rede com ligações aleatórias divididas em dois grupos também com possibilidade de ligações múltiplas e autoligações. **(C)** Gráfico do logaritmo da distribuição de probabilidade $\tilde{P}$ de dois nós escolhidos não se aglutinarem depois de n ligações incluídas no caso da rede aleatória. **(D)** Gráfico do logaritmo da distribuição de probabilidade $\tilde{P}$ de dois nós escolhidos e que não pertencem ao mesmo grupo não se aglutinarem depois de n ligações incluídas. Figura de autoria própria.

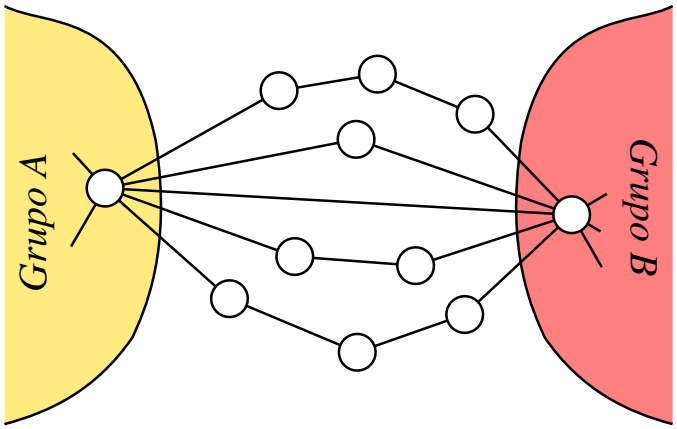


Figura 14 – **Dois grupos conectados apenas por dois nós.** Na imagem estão representados dois grupos que estão conectados apenas por dois nós. O conjunto de caminhos que conectam esses nós é chamado de bridge e os nós apenas se conectam pela bridge. Nesta bridge, estão representados cinco segmentos em paralelo, sendo um com uma ligação em série, um com duas ligações em série, um com três ligações em série e dois com quatro ligações em série. Figura de autoria própria.

Tendo como alvo entender o dado descrito no parágrafo anterior, vamos, inicialmente, abordar o seguinte problema. Imaginemos dois grupos que estão conectados através somente de dois nós que estão um em cada grupo. O conjunto de diferentes segmentos que aglutinam em um mesmo agregado esses dois nós será chamado de bridge. Na bridge, cada sequência de ligações serão denominadas ligações em série e os conjuntos de ligações em série serão chamados de segmentos em paralelo. Na Fig. 14, temos, assim, cinco segmentos em paralelo, sendo apenas um segmento para os casos de uma, duas e três ligações em série e dois segmentos com quatro ligações em série.

Iremos, primeiramente, trabalhar com o caso em que há somente um segmento formando a bridge como na Fig 15. Queremos aqui analisar a probabilidade $\tilde{P}_n{}^{(1)}$ dos dois nós que estão um em cada grupo não se aglutinarem depois de n ligações incluídas, considerando que eles só se aglutinam por único segmento, diremos que é o segmento 1. Por isso o número um entre parênteses na probabilidade $\tilde{P}_n^{(1)}$.

Assim, dado m o tamanho deste único segmento que na Fig. 15 é $m = 4$, k o número de ligações removidas l_i e a rede tendo N ligações totais. Além disso, trabalhando com redes sequenciais que se definem como redes nas quais as ligações são enumeradas em sequência como pode ser visto na Fig. 16 em que as duas redes têm os mesmos pares de nós conectados, porém em uma sequência diferente. Portanto, formando duas redes sequenciais distintas.

A probabilidade dos dois nós que estão nas extremidades da bridge não se aglutina-

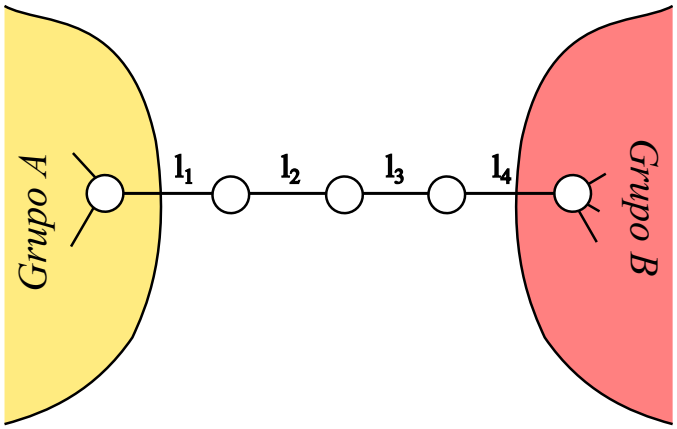


Figura 15 – **Bridge com apenas um segmento.** Seja l_i uma ligação em série do único seguimento paralelo da bridge. Para que não haja conexão entre os nós terminais da bridge, pelo menos uma das ligações l_i deve ser removida. Portanto, se definirmos A_C como o conjunto das redes em que não conectam os nós terminais. Além disso, sendo A_i o conjunto das redes nas quais a ligação l_i foi removida. Então, $N_C = |A_c| = |\cup_i A_i|$. Isso permite usar o princípio da inclusão-exclusão para encontrar N_C. Figura de autoria própria.

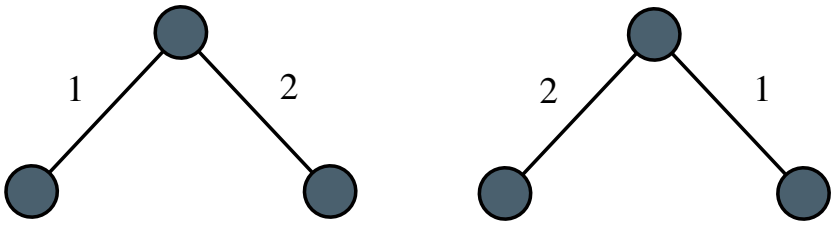


Figura 16 – **Duas redes sequenciais.** Na imagem, estão representadas duas redes sequenciais, em que os números ao lados das ligações indicam a sequência em que as ligações foram feitas. Assim, como são redes sequenciais, apesar dos mesmo pares de nós estarem conectados, como foram conectados em uma sequência diferentes, então são consideradas redes diferentes. Figura de autoria própria.

rem depois de n ligações incluídas é dada por

$$\tilde{P}_n^{(1)} = \frac{N_C}{N_T}, \tag{4.1}$$

em que N_C é o número de possíveis reconstruções da rede na qual os nós não se aglutinam depois de n ligações incluídas e $N_T = N^n$ é o número total de possíveis reconstruções de redes sequenciais possíveis. Como N_T já é conhecido, então basta calcular N_C.

Para encontrar N_C, devemos observar, primeiramente, que para os nós terminais não se aglutinarem pelo menos uma das l_i ligações não deve estar presente. Daí, definindo A_C como o conjunto de possíveis reconstruções da rede em que não há a aglutinação em um mesmo agregado dos dois nós terminais da bridge. Além disso, definindo A_i como o conjunto das reconstruções na qual a ligação l_i não foi recolocada. Como para não haver uma aglutinação através do segmento pelo menos uma das ligações deve não estar presente, temos, $N_C = |A_C| = |\cup_i A_i|$. Portanto, o conjunto total é a união dos conjuntos A_i, dessa forma permitindo utilizar o princípio da

inclusão-exclusão (ALLENBY; SLOMSON, 2010) para encontrar N_C.

O princípio da Inclusão-Exclusão está exemplificado na Fig. 17. A partir dele, podemos contar o número total dos elementos de conjuntos levando em consideração que, por exemplo, quando contamos os elementos do conjunto A_1 e A_2, também contamos elementos da interseção de A_1 e A_2. Portanto, sendo necessário excluir a interseção de A_1 e A_2 da contagem. No entanto, ao fazermos isso, estaremos retirando elementos da interseção de A_1, A_2 e A_3. Logo, precisando adicioná-los na contagem e assim por diante segue-se somando e subtraindo interseções até se contar todos os elementos dos conjuntos. Daí, ficamos com a seguinte expressão:

$$N_C = -\sum_{k=1}^{m} (-1)^k \binom{m}{k} (N-k)^n. \tag{4.2}$$

Note que $\binom{m}{k}$ é o numero de possíveis combinações de k grupos A_i de reconstruções em que a ligação l_i está ausente, $(N-k)^n$ é o tamanho da interseção entre esses grupos e os termos da soma alternam de sinal como esperado pelo princípio da inclusão-exclusão. Então, a probabilidade dos nós terminais não se aglutinarem para o caso em que há somente um segmento é dada por

$$\tilde{P}_n^{(1)} = -\sum_{k=1}^{m} (-1)^k \binom{m}{k} \left(1 - \frac{k}{N}\right)^n. \tag{4.3}$$

Analisado o caso em que há apenas um segmento paralelo, agora temos que generalizar para vários segmentos j. Assim, considerando que há N ligações totais, cada ligação tem uma probabilidade de ρ_j de ser colocado no segmento j. Quanto mais ligações são colocadas no segmento j, menos ligações poderão ser colocadas em outro segmento j'. Portanto, há correlação e as probabilidades ρ_j's não são independentes. No entanto, se o número total de ligações na bridge é muito menor que N, a correlação é desprezível. Além disso, quando existem p possíveis segmentos em paralelo, para que não exista a aglutinação através da bridge é preciso que não exista aglutinação através de nenhum dos segmentos em paralelo. Como desconsideramos correlações, podemos considerar que a probabilidade de não aglutinar através da bridge é o produto das probabilidades de não aglutinar através de cada segmento, ou seja,

$$\tilde{P}_n = \prod_{j=1}^{p} \tilde{P}_n^{(j)}. \tag{4.4}$$

Ademais, dado p o número total de segmentos em paralelo, m_j e k_j o tamanho do segmento e o número de ligações removidas do segmento j respectivamente. A probabilidade dos nós terminais não se conectarem é dada por

$$\tilde{P}_n = \prod_{j=1}^{p} \sum_{k_j=1}^{m_j} \left[-(-1)^{k_j} \binom{m_j}{k_j} \left(1 - \frac{k_j}{N}\right)^n \right]. \tag{4.5}$$

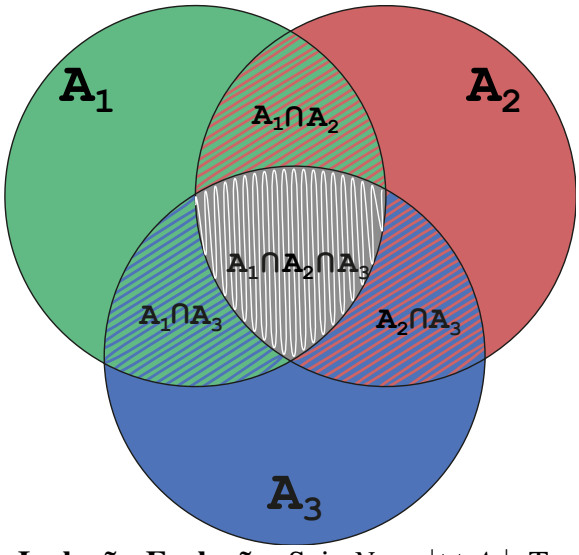


Figura 17 – **Princípio da Inclusão-Exclusão.** Seja $N_C = |\cup_i A_i|$. Tomando inicialmente $N_C = \sum_i |A_i|$ cometemos o erro de contar múltiplas vezes as interseções dos conjuntos A_i. Corrigindo isso na forma $N_C = \sum_i |A_i| - \sum_{i \neq j} |A_i \cap A_j|$, ainda erramos, porque nesse caso a interseção dos 3, $|A_1 \cap A_2 \cap A_3|$, não foi contada. O princípio da inclusão-exclusão diz que seguimos tomando em sequencias as interseções de todos as possíveis seleções de c conjuntos A_i, incluindo (tomando sinal positivo) as interseções de c ímpar e excluindo (tomando sinal negativo) as interseções de c par. Figura de autoria própria.

Da Equação (4.5), podemos agrupar os termos em que se tem s ligações em série. Assim, chamando M_s o número de segmentos em paralelo com s ligações em série e f o número de ligações que foram removidas, então ficamos com

$$\tilde{P}_n = \prod_{s=1} \left[\sum_{f=1}^{s} -(-1)^f \binom{s}{f} \left(1 - \frac{f}{N}\right)^n \right]^{M_s}. \tag{4.6}$$

Desenvolvendo o último termo entre parênteses, considerando $N \gg f$,

$$\left(1 - \frac{f}{N}\right)^n = \left(1 - \frac{f}{N}\right)^{\frac{n}{N}N} \simeq (e^{-n/N})^f, \tag{4.7}$$

em que e é o número de Euler. Chamando o somatório entre colchetes de C_{ns} e desenvolvendo

$$\begin{gathered} C_{ns} \simeq -\sum_{f=1}^{s} (-1)^f \binom{m}{f} (e^{-n/N})^f (1)^{s-f} - 1 + 1 \Rightarrow \\ \mathrm{C}_{ns} \simeq 1 - \sum_{f=0}^{s} (-1)^f \binom{m}{f} (e^{-n/N})^f (1)^{s-f} \Rightarrow \\ \mathrm{C}_{ns} \simeq 1 - (1 - e^{-n/N})^s. \end{gathered}$$

Conseguimos então,

$$C_{ns} \simeq 1 - (1 - e^{-n/N})^s. \tag{4.8}$$

Finalmente, substituindo este resultado na Equação (4.5), nossa equação final para a probabilidade de não haver caminhos que aglutinam os dois nós que estão nas extremidades da bridge é, então, dada por

$$\tilde{P}_n = \prod_{s=1} [C_{ns}]^{M_s}. \tag{4.9}$$

Usando logaritmo,

$$\ln \tilde{P}_n = \sum_s M_s \ln C_{ns}. \tag{4.10}$$

O que desejamos determinar no final são os valores de M_s que é exatamente o número de segmentos em paralelo de tamanho s. Assim, de fato, a enumeração efetiva dos caminhos é apenas uma forma de caracterizar as distribuições de instantes de aglutinação.

4.2 Obtenção da distribuição dos instantes de aglutinação

Nesta seção está explicado os métodos computacionais e os passos que tomamos até chegar à probabilidade $\tilde{P}_n$ empírica. Para isso, inicialmente, elaboramos um código em que todos os dados necessários eram armazenados em um grande arquivo que comportaria todas as realizações de reconstrução da rede e permitiria a obtenção rápida dos instantes de aglutinação de qualquer par de nós da rede.

Primeiramente, cria-se uma rede escolhendo-se a quantidade fixa de nós e de ligações. Logo após, escolhe-se aleatoriamente quais ligações serão feitas levando-se em conta que são permitidas autoligações e ligações múltiplas entre nós. Vale a pena salientar que, no primeiro momento, as ligações são somente listadas, pois elas só serão feitas realmente no processo que chamamos de reconstrução da rede. Além disso, tanto as quantidades de nós e ligações quanto as ligações listadas não se alteram durante todo o processo.

Tendo definido isso, as ligações começam a serem feitas aleatoriamente à medida que há duas contagens sendo realizadas. A primeira delas é a contagem de quantas ligações novas foram feitas seguindo a lista descrita no parágrafo anterior, ou seja, neste caso não se permite repetir nenhuma ligação da lista. Assim cada ligação é contada apenas uma vez. Por outro lado, na segunda contagem, leva-se em consideração que se pode repetir ligações da lista e é este valor o que estamos mais interessados. Portanto, ao passo que se contabiliza n_l ligações novas feitas, contabiliza-se também o número n de ligações feitas caso as ligações pudessem ser repetidas.

Além disso, quando uma ligação nova é feita e como resultado há uma aglutinação de nós em um mesmo agregado, registra-se a raiz do novo agregado. Para se registrar as raízes, utiliza-se o método union-find (NEWMAN; ZIFF, 2001) dando preferência à raiz nova para o novo agregado ser aquela com menor valor e não de maior agregado como é mais usual. Daí, o processo continua até que todos os nós estejam no mesmo agregado. Quando isso acontece, o processo para, desocupa-se todas as ligações e inicia-se a reconstrução novamente.

Portanto, ao final deste primeiro processo, tem-se os dados divididos em 4 colunas. Em duas delas estão registradas os valores de n_l e n à medida que cada ligação é feita e se tem uma aglutinação. As demais colunas registram as raízes que se aglutinaram.

Este programa e esta forma de salvar os dados mostra-se interessante, porque foram salvos os dados essenciais, que são as raízes e os valores n_l e n. Portanto, não sendo necessário salvar todas as ligações, por exemplo, que faria o arquivo tornar-se mais extenso. Além disso, basta agora criar outro programa que analisa os dados levando-se em consideração qual par de nós que se deseja estudar. Assim, não se faz necessário criar a rede de novo e fazer a simulação para cada par. É suficiente apenas analisar os dados que já estão à disposição.

Daí em diante, para analisar os dados, criou-se outro código no qual é definido previamente os nós os quais estamos interessados em estudar. O algoritmo abre o arquivo e busca nas colunas das raízes o momento em que elas se tornam iguais e a partir disso, encontra-se o número de ligações incluídas n na qual as raízes dos nós tornam-se as mesmas.

Tendo registrado todos os valores de n em que houve aglutinação dos nós para todas as amostras N_{sam}. Conta-se quantas vezes se repete cada valor de n e armazena-se na variável i a soma da quantidade de vezes que cada valor de n se repete com a quantidade de vezes que os outros valores menores que n também se repetem. Assim, por exemplo, para uma dado $n = 4$, a variável i registra a soma da quantidade de vezes que $n = 1,2,3,4$ apareceram. Tendo, portanto, que o valor de i sempre cresce até se atingir N_{sam}.

Daí, para se calcular $\tilde{P}_n$, calcula-se $N_{sam} - 1$ e divide-se pelo número total de amostras N_{sam},

$$\tilde{P}_n = \frac{N_{sam} - i}{N_{sam}}. \tag{4.11}$$

Assim, tendo os valores de $\tilde{P}_n$ que iremos utilizar para nosso objetivo também. Se quisermos calcular a probabilidade de dois nós se aglutinarem exatamente num valor de n escolhido, basta calcularmos o negativo da derivada.

4.3 Enumerando os segmentos em paralelo das bridges utilizando o método dos mínimos quadrados lineares gerais

Nesta seção, desenvolveremos os resultados empíricos considerando o problema da bridge abordado na seção 4.1. Fizemos isso, realizando simulações de redes da forma da Fig. 14 e obtemos as distribuições de probabilidades $\tilde{P}_n$ dos nós que estão nas extremidades da bridge não se aglutinarem após terem n ligações incluídas à rede. Para analisar estes resultados, usamos a Equação (4.10) para escrever

$$\begin{aligned} X_{is} &= \ln C_{n_i s} \\ Y_i &= \ln P_{n_i}. \end{aligned} \tag{4.12}$$

em que n_i e Y_i são os valores empíricos da quantidade de ligações incluídas e o logaritmo da probabilidade empírica respectivamente. Além, podemos escrever na forma vetorial,

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{M}. \tag{4.13}$$

Neste caso, vamos fornecer os valores dos tamanhos dos segmentos paralelos s existentes na rede e queremos descobrir os valores do vetor $\mathbf{M}$ dos valores de M_s que pode ser dado por

$$\mathbf{M} = \mathbf{X}^{-1}\mathbf{Y}, \tag{4.14}$$

no qual, fornecemos os elementos de $\mathbf{X}$ e $\mathbf{Y}$. Sabendo que o primeiro é dado por $C_{n_i s}$ e o segundo tiramos de dados empíricos. Neste caso, criamos uma rede que está dividida em dois grupos que possuem bridges com M_s segmentos em paralelo com s ligações em série e usamos o método de mínimos quadrados gerais que, pode ser encontrado em (TEUKOLSKY *et al.*, 1992) e descrito nos apêndices A e B, para encontrar os valores de $\mathbf{M}$. Assim, utilizamos a seguinte equação:

$$y_i(x) = \sum_{s=0}^{L-1} M_s X_s(n_i) \tag{4.15}$$

com o objetivo de minimizar a função de mérito χ na seguinte expressão,

$$\chi^2 = \sum_{i=0}^{N_a-1} \left[\frac{y_i - \sum_{k=0}^{L-1} M_s X_s(n_i)}{\varepsilon_i} \right]. \tag{4.16}$$

No entanto, desta equação não sabemos o valor do erro empírico ε da probabilidade empírica de não aglutinar os nós das extremidades da bridge e precisamos descobri-lo para realizar o ajuste. Para solucionar este problema, comecemos estudando a probabilidade empírica citada que é dada por

$$\tilde{P}_e = \frac{N_o}{N_s}, \tag{4.17}$$

em que N_o é o número de vezes em que os nós em questão não se aglutinaram em N_s amostras. Além disso, N_o é binomial, pois trata-se de um problema de duas opções que são: os nós em questão se aglutinam ou não se aglutinam. Portanto, devemos ter

$$\tilde{P}_e(N_o) = \binom{N_s}{N_o} \tilde{P}^{N_o}(1-\tilde{P})^{N_s-N_o} \tag{4.18}$$

em que $\tilde{P}$ é o valor verdadeiro da probabilidade. Sendo assim, tendo $\tilde{P}_e = \tilde{P} + \delta$, sendo δ o desvio da medida empírica $\delta = \sigma/\sqrt{N}$, no qual σ é o desvio padrão. Evidentemente, $\langle\delta\rangle = 0$, tendo em vista que $\langle\tilde{P}_e\rangle = \tilde{P}$. No entanto, o que desejamos é estimar o erro ε associado ao $\ln\tilde{P}_e$. Para isso, inicialmente, podemos escrever:

$$\ln\tilde{P}_e = \ln(\tilde{P}+\delta) = \ln\tilde{P} + \ln\left(1+\frac{\delta}{\tilde{P}}\right), \tag{4.19}$$

e a partir desta expressão, calculamos o erro de $\ln\tilde{P}_e$, dado por

$$\varepsilon = \langle(\ln\tilde{P}_e)^2\rangle - \langle\ln\tilde{P}_e\rangle^2. \tag{4.20}$$

Desenvolvendo os termos e usando $x = \frac{\delta}{\tilde{P}}$,

$$\langle(\ln\tilde{P}_e)^2\rangle - \langle\ln\tilde{P}_e\rangle^2 =$$
$$(\ln\tilde{P})^2 + 2\ln\tilde{P}\Big\langle\ln(1+x)\Big\rangle + \Big\langle(\ln(1+x))^2\Big\rangle - (\ln\tilde{P})^2 - 2\ln\tilde{P}\Big\langle\ln(1+x)\Big\rangle - \Big\langle(\ln(1+x))^2\Big\rangle \Rightarrow$$
$$\langle(\ln\tilde{P}_e)^2\rangle - \langle\ln\tilde{P}_e\rangle^2 = \langle(\ln(1+x))^2\rangle - \langle\ln(1+x)\rangle^2.$$

Dada a seguinte expansão em série de Taylor,

$$\ln(1+x) = \sum_{i=1}^{\infty}(-1)^{i+1}\frac{x^i}{i}, \tag{4.21}$$

temos,

$$\langle(\ln(1+x))^2\rangle - \langle\ln(1+x)\rangle^2 = \langle[x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3...]^2\rangle - [\langle x - \tfrac{1}{2} + \tfrac{1}{3}x^3..\rangle]^2 \Rightarrow$$
$$\varepsilon = \langle x^2 - x^3 + \tfrac{1}{4}x^4...\rangle - [\langle x\rangle - \tfrac{1}{2}\langle x^2\rangle + \tfrac{1}{3}\langle x^3\rangle...]^2 \Rightarrow$$

Como o termo $\langle x\rangle$ é nulo, pois $\langle\tilde{P}_e\rangle = \tilde{P}$ e desconsiderando os termos a partir da terceira ordem. Resta apenas

$$\varepsilon \simeq \left\langle\left(\frac{\delta}{\tilde{P}}\right)^2\right\rangle. \tag{4.22}$$

Finalmente, como estamos trabalhando com uma distribuição binomial, $\sigma^2 = N\tilde{P}(1-\tilde{P})$, temos

$$\varepsilon \simeq \frac{1-\tilde{P}}{\tilde{P}}. \tag{4.23}$$

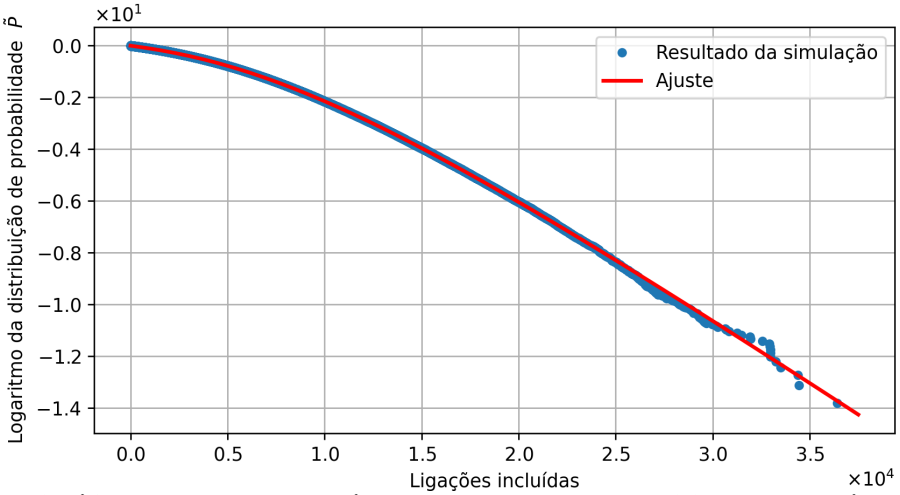


Figura 18 – **Gráfico dos dados empíricos e da simulação utilizando o método de mínimos quadrados gerais para o ajuste linear.** No caso abordado foi realizado uma simulação para uma rede com um segmento em paralelo para os casos de uma, duas e três ligações em série e dois segmentos em paralelo com quatro ligações em série como pode ser visto da Fig. 14. Além disso, usamos uma rede com 10.000 ligações e utilizamos 1.000.000 de amostras para construir a distribuição de probabilidade. Figura de autoria própria.

Ao realizarmos o ajuste linear para o mesmo caso da Fig. 14, colocando os valores dos tamanhos dos segmentos em paralelo que existem na rede, chegamos então em um resultado com erro menor que 10% da quantidade de segmentos em paralelo que há para cada tamanho de segmento. Sinalizando um bom ajuste como pode ser visto na Fig. 18.

4.4 Enumerando os segmentos em paralelo das bridges utilizando o método de maximização da likelihood

Nesta seção, desenvolveremos os resultados empíricos considerando o problema da bridge abordado na primeira seção deste capítulo. Porém, desta vez, utilizando o método de maximização da likelihood. Para isso, inicialmente, calculamos a probabilidade p_n dos nós se aglutinarem no instante de aglutinação n. Sabendo que $\tilde{P}_n$ é a probabilidade cumulativa dos nós não se aglutinarem após n ligações incluídas, basta calcular o negativo da derivada de $\tilde{P}_n$ em relação a n e encontra-se a probabilidade p_n dos nós se aglutinarem no instante de aglutinação n. Daí, derivando (4.9)

$$p_n = -\tilde{P}_n \sum_s \frac{\partial}{\partial n}[C_{ns}]^{M_s} \frac{1}{[C_{ns}]^{M_s}}, \tag{4.24}$$

desenvolvendo esta expressão, chegamos em

$$p_n = \tilde{P}_n \sum_s \frac{sM_s}{N} \frac{1}{[(1-e^{-n/N})^{-s}-1][e^{n/N}-1]}, \tag{4.25}$$

e usando a relação (4.8),

$$C_{ns} \simeq 1 - (1 - e^{-n/N})^s. \tag{4.26}$$

podemos simplificar e chegar na seguinte forma,

$$p_n = \tilde{P}_n \sum_s \frac{sM_s}{N} \frac{C_{ns} - 1}{C_{ns}(1 - e^{n/N})}. \tag{4.27}$$

Agora calculando a likelihood L,

$$\mathscr{L} = \prod_{i=1} \left\{ \prod_s [1 - (1 - e^{-n_i/N})^s]^{M_s} \sum_s \frac{sM_s}{N} \frac{1}{[(1 - e^{-n_i/N})^{-s} - 1][e^{n_i/N} - 1]} \right\}, \tag{4.28}$$

e como é de costume, usaremos o logaritmo de L. Assim,

$$\ln(\mathscr{L}) = \sum_{i=1} \left\{ \sum_s M_s \ln[1 - (1 - e^{-n_i/N})^s] + \ln \sum_s \frac{sM_s}{N} \frac{1}{[(1 - e^{-n_i/N})^{-s} - 1][e^{n_i/N} - 1]} \right\}. \tag{4.29}$$

Dada a likelihood, com o objetivo de maximizá-la em relação aos valores de M_s, vamos calcular sua derivada..

$$\frac{\partial \ln(\mathscr{L})}{\partial M_k} = \sum_{i=1} \left\{ \ln[1 - (1 - e^{-n_i/N})^k] + \frac{k}{[(1 - e^{-n_i/N})^{-s} - 1]} \frac{1}{\sum_s \frac{sM_s}{N} \frac{1}{[(1 - e^{-n_i/N})^{-s} - 1]}}, \right\} \tag{4.30}$$

podemos reescrever como,

$$\frac{\partial \ln(\mathscr{L})}{\partial M_k} = \sum_{i=1} \left\{ \ln[1 - (1 - e^{-n_i/N})^k] + k \frac{1 - C_{n_i k}}{C_{n_i k}} \frac{1}{\sum_k \frac{kM_k(1 - C_{n_i k})}{C_{n_i k}}} \right\}. \tag{4.31}$$

Utilizando esta equação e um algoritmo de gradiente ascendente semelhante ao método de gradiente descendente encontrado por exemplo em (POLYAK, 2020). Inicialmente, escolhemos valores arbitrários para os parâmetros que desejamos descobrir que no nosso caso são os M_k's, em seguida calculamos as derivadas para cada M_k aplicando os valores de n_i empírico na Equação (4.31) e atualizamos os valores de M_k empregando a seguinte relação,

$$M_k = M_k + \varepsilon \frac{\partial \ln(\mathscr{L})}{\partial M_k} \tag{4.32}$$

em que ε é a tamanho do passo escolhido pelo usuário. Este procedimento é feito até que as derivadas sejam nulas, ou seja, quando se atinge o valor do máximo da likelihood. Porém, no caso abordado neste trabalho, temos que colocar a restrição de evitar que o valor de algum M_k seja menor que zero, pois não faria sentido. Assim quando algum M_k é menor ou igual a zero, nós anulamos o valor dele e fazemos sua derivada também ser anulada.

Com este método conseguimos atingir os valores esperados de todos os M_k's com boa aproximação também. Como pode ser visto na Fig. 19.

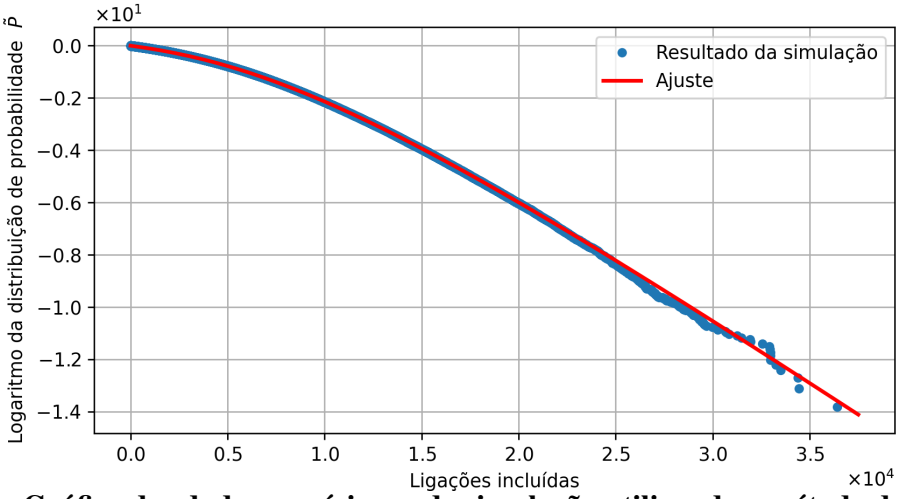


Figura 19 – **Gráfico dos dados empíricos e da simulação utilizando o método de maximização da likelihood.** No caso abordado foi realizado uma simulação para uma rede com um segmento paralelo para os casos de uma, duas e três ligações em série e dois caminhos paralelos com quatro ligações em série como na Fig. 14. Além disso, usamos uma rede com 10.000 ligações e 1.000.000 de amostras para construir a distribuição de probabilidade. Figura de autoria própria.

4.5 Limitações do modelo

Tínhamos almejado utilizar o método de enumeração dos caminhos efetivos a fim de detectar comunidades em rede. Para tanto, criamos redes artificiais com estruturas de comunidades e investigamos a enumeração no caso em que os nós estão na mesma comunidade e no caso em que não estão. Contudo, para os dois casos a abordagem não se mostrou eficaz. Assim, os algoritmos retornam enumerações efetivas não condizentes com as redes analisadas.

Como exemplo, temos representada na Fig. 20 uma rede pequena dividida em duas comunidades. Cada comunidade possui 10 nós e a rede tem 40 ligações totais, sendo 18 ligações intragrupos em cada grupo e 4 ligações intergrupos. Para esta rede foi aplicado os métodos de enumeração efetiva de caminhos tanto utilizando mínimos quadrados lineares gerais quanto maximização da likelihood, porém os resultados não foram satisfatórios.

Além disso, aplicamos também a abordagem em redes grandes com 10.000 ligações, porém mantendo-se as ligações e os nós da Fig. 20 e criando nós e ligações que não se conectam às comunidades que já estavam presentes. Chegamos também em resultados não adequados. Portanto, mostrando-se que esta abordagem é restrita demais para reproduzir perfeitamente a distribuição de instantes de aglutinação em redes divididas em comunidades sejam em redes pequenas com poucos nós e ligações, sejam em comunidades pequenas e isoladas dentro de grandes redes.

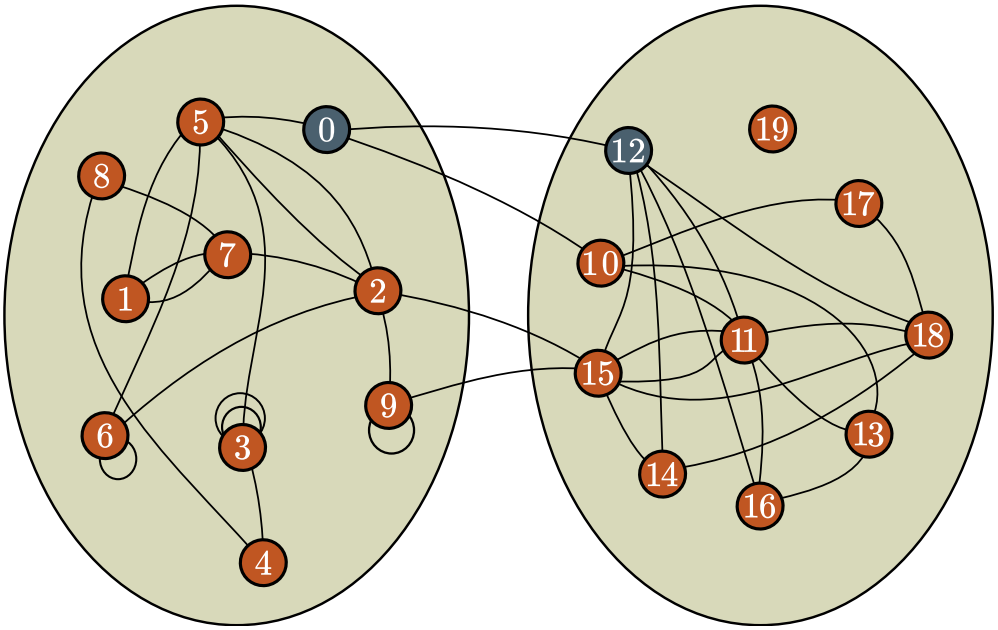


Figura 20 – **Rede divida em duas comunidades.** Uma rede com 20 nós e 40 ligações, podendo haver autoligações e ligações múltiplas. Há 18 ligações intragrupos em cada grupo e 4 ligações intergrupos.

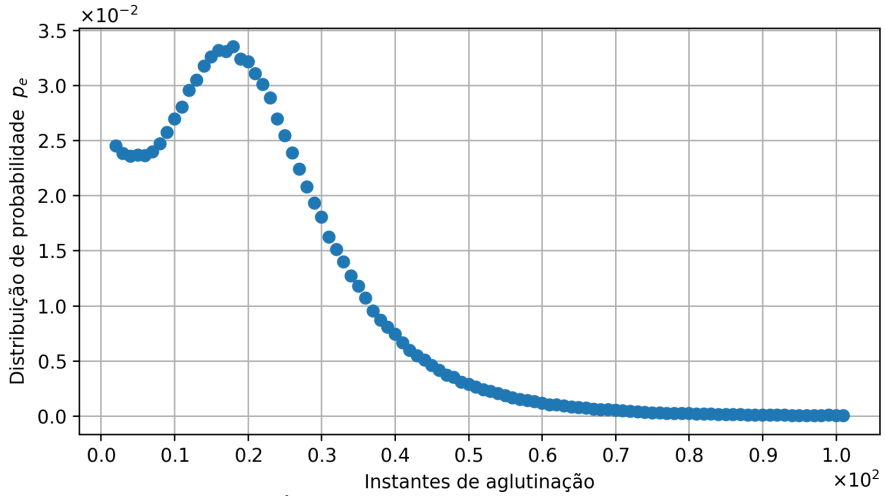


Figura 21 – **Probabilidade empírica p_e.** Gráfico da probabilidade empírica p_e dos nós 0 e 12 não se aglutinarem no instante de aglutinação n para uma rede pequena com dois grupos. A probabilidade foi calculada aos se derivar numericamente o negativo dos valores de $\tilde{P}$.

Uma das possíveis causas dos erros obtidos, pode se dever ao fato de que estamos trabalhando com distribuições de probabilidade que possuem derivada segunda negativa em todos os pontos, isto é, são funções côncavas como pode ser verificado pela Equação (4.27). Porém, os resultados empíricos das probabilidades revelam que as funções devem ter também derivada segunda positiva, ou seja, devem ser funções convexas em alguns pontos como pode ser vista na Fig. 21 na qual temos o gráfico da distribuição de probabilidade empírica p_e dos nós se aglutinarem no instante de aglutinação n para o par de nós destacados em verde na Fig. 20. Como o gráfico da derivada de p_e não é compatível com as equações de probabilidade que são obtidas, por sua vez, de $\tilde{P}_e$, então $\tilde{P}$ também se revela inconsistente.

4.6 Perspectivas

Nós pretendemos, contornar os problemas tidos até aqui ao se analisar redes divididas em comunidades mudando a abordagem a ser aplicada. O novo método terá como análise o fato de que no processo de reconstrução de uma rede pode haver ligações aglutinantes, isto é, ligações entre nós que ainda não pertencem ao mesmo agregado e ligações não aglutinantes que são ligações entre nós que já estão no mesmo agregado como pode ser visto na Fig. 22 . A lista de ligações feitas na reconstrução registra apenas ligações aglutinantes. Acreditamos que as ligações que aparecem com maior frequência na lista de aglutinantes são ligações que estão nas fronteiras das comunidades. Ou seja, ligações que saem de um grupo de chegam em outro. Essa abordagem tem relação com a definição de message passing (NEWMAN, 2023), que já é usada como uma forma de identificação de comunidades.

Assim, analisando as ligações aglutinantes que aparecem com maior frequência esperamos que elas sejam as ligações que conectam as comunidades. Tendo em vista que as ligações intragrupos, provavelmente, conectam com maior frequência apenas as ligações dos seus respectivos grupos e, em contrapartida, as ligações que conectam as comunidades devem conectar tanto os nós que estão em um mesmo grupo quanto em grupos distintos.

Portanto, contaremos a frequência com que cada ligação aglutinante conecta cada par de nós. E ao se somar todas as frequências, provavelmente, teremos sugestões de quais são as ligações intragrupos e quais são ligações intergrupos.

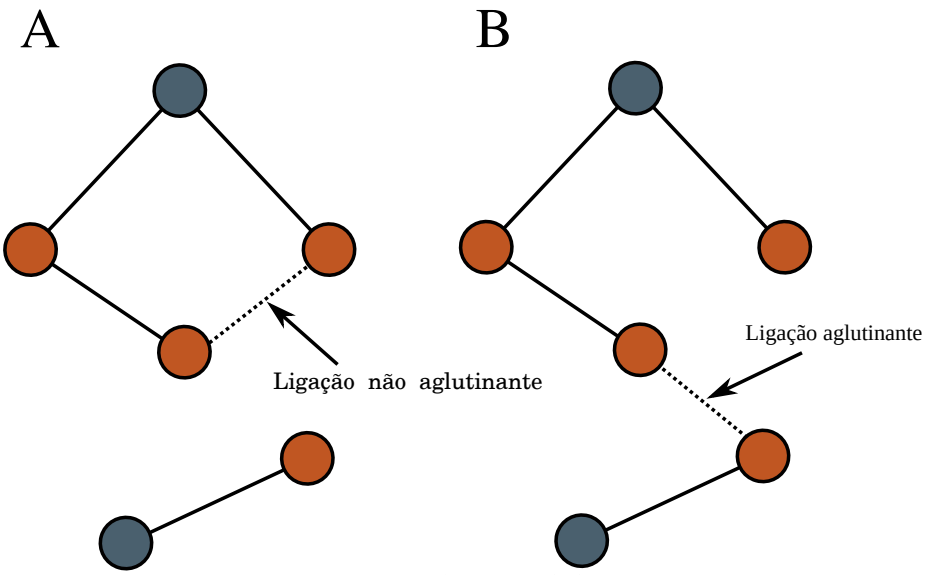


Figura 22 – **Ligação aglutinante e não aglutinante. (A)** A linha tracejada que simboliza uma ligação não aglutinante em relação aos dois nós destacados em azul. **(B)** A linha tracejada simboliza uma ligação aglutinante em relação aos nós destacados em azul.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, conseguimos construir duas abordagens satisfatórias na enumeração dos caminhos efetivos entre dois nós que estão ligados por caminhos de diferentes tamanhos em que cada nó está em um grupo diferente e que os grupos se conectam apenas por um único par de nós. Sendo que, no caso abordado, já foi dado os tamanhos dos segmentos em paralelo e descobriu-se o número de segmentos em paralelo para cada um dos tamanhos dos segmentos. Na primeira abordagem utilizamos o método de mínimos quadrados generalizados e na segunda abordagem maximização da likelihood. As duas funcionam bem para os casos estudados.

Tínhamos como perspectiva, utilizar e aperfeiçoar os resultados obtidos até aqui a fim de detectar comunidades em redes. No entanto, ao testarmos em redes divididas em grupos com mais que dois nós conectando as redes, os resultados não se mostraram adequados. Umas das possíveis causas de não termos conseguido, pode ser devido aos resultados empíricos da probabilidade em alguns momentos mostrarem derivada segunda positiva e nos nossos modelos as equações sempre possuíam derivada segunda negativa. Pretendemos contornar esses problemas, mudando a abordagem e dessa vez analisar as ligações que mais unem os nós, ou seja, ligações que quando feitas com maior frequência faz com que nós comecem a pertencer ao mesmo agregado.

REFERÊNCIAS

ALLENBY, R. B.; SLOMSON, A. **How to count**: an introduction to combinatorics. [*S. l.*]: Chapman and Hall/CRC, 2010.

ANDRADE, R. F.; MIRANDA, J. G.; PINHO, S. T.; LOBAO, T. P. Characterization of complex networks by higher order neighborhood properties. **The European Physical Journal B**, Springer, v. 61, p. 247–256, 2008.

ANDRADE, R. F. S.; MIRANDA, J. G. V.; AO, T. P. L. Neighborhood properties of complex networks. **Phys. Rev. E**, American Physical Society, v. 73, p. 046101, abr 2006.

ATKINSON, K. **An introduction to numerical analysis**. [*S. l.*]: John wiley & sons, 1991.

BARABáSI, A.-L. **Network Science**. [*S. l.*]: Cambridge University Press, 2016.

BICKEL, P. J.; CHEN, A. A nonparametric view of network models and newman–girvan and other modularities. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 106, n. 50, p. 21068–21073, 2009.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [*S. l.*]: Cengage Learning, 2021.

CRAMER, C. B.; PORTER, M. A.; SAYAMA, H.; SHEETZ, L.; UZZO, S. M. **Network science in education: Transformational approaches in teaching and learning**. [*S. l.*]: Springer, 2018.

CSERMELY, P. Creative elements: network-based predictions of active centres in proteins and cellular and social networks. **Trends in biochemical sciences**, Elsevier, v. 33, n. 12, p. 569–576, 2008.

DECELLE, A.; KRZAKALA, F.; MOORE, C.; ZDEBOROVÁ, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. **Physical Review E**, APS, v. 84, n. 6, p. 066106, 2011.

DECELLE, A.; KRZAKALA, F.; MOORE, C.; ZDEBOROVÁ, L. Inference and phase transitions in the detection of modules in sparse networks. **Physical Review Letters**, APS, v. 107, n. 6, p. 065701, 2011.

DODDS, P. S.; MUHAMAD, R.; WATTS, D. J. An experimental study of search in global social networks. **science**, American Association for the Advancement of Science, v. 301, n. 5634, p. 827–829, 2003.

ERDÖS, P.; RÉNYI, A. On random graphs i. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

ERDOS, P.; RENYI, A. On the evolution of random graphs. pubblication of the mathematical institute of the hungarian academy of sciences. 1960.

ERDŐS, P.; RÉNYI, A. On the strength of connectedness of a random graph. **Acta Mathematica Hungarica**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV . . . , v. 12, n. 1, p. 261–267, 1961.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, 2009.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002.

GOLUB, G. H.; LOAN, C. F. V. **Matrix computations**. [*S. l.*]: JHU press, 2013.

HOLLAND, P. W.; LASKEY, K. B.; LEINHARDT, S. Stochastic blockmodels: first steps. **Social networks**, Elsevier, v. 5, n. 2, p. 109–137, 1983.

JACOBS, A. Z.; WAY, S. F.; UGANDER, J.; CLAUSET, A. Assembling the facebook: sing heterogeneity to understand online social network assembly. *In*: **Proceedings of the ACM Web Science Conference**. [*S. l.*: *s. n.*], 2015. p. 1–10.

JOYCE, J. Bayes' Theorem. *In*: ZALTA, E. N. (ed.). **The Stanford Encyclopedia of Philosophy**. Fall 2021. [*S. l.*]: Metaphysics Research Lab, Stanford University, 2021.

KATTI, S.; RAO, A. V. **Handbook of the poisson distribution**. [*S. l.*]: Taylor & Francis, 1968.

KRISHNAMURTHY, B.; WANG, J. On network-aware clustering of web clients. *In*: **Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication**. [*S. l.*: *s. n.*], 2000. p. 97–110.

LEE, C.; WILKINSON, D. J. A review of stochastic block models and extensions for graph clustering. **Applied Network Science**, SpringerOpen, v. 4, n. 1, p. 1–50, 2019.

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967.

NEWMAN, M. **Networks**. 2nd. ed. [*S. l.*]: Oxford University Press, 2018.

NEWMAN, M. Message passing methods on complex networks. **Proceedings of the Royal Society A**, The Royal Society, v. 479, n. 2270, p. 20220774, 2023.

NEWMAN, M. E. J.; ZIFF, R. M. Fast monte carlo algorithm for site or bond percolation. **Phys. Rev. E**, American Physical Society, v. 64, p. 016706, jun 2001.

POLYAK, B. **Introduction to Optimization**. [*S. l.*: *s. n.*], 2020.

REDDY, P. K.; KITSUREGAWA, M.; SREEKANTH, P.; RAO, S. S. A graph based approach to extract a neighborhood customer community for collaborative filtering. *In*: **DATABASES IN NETWORKED INFORMATIO SYSTEMS: Second International Workshop, DNIS 2002 Aizu, Japan, dezembro 16–18, 2002 Proceedings 2**. [*S. l.*: *s. n.*], 2002. p. 188–200.

SOLOMONOFF, R.; RAPOPORT, A. Connectivity of random nets. **The bulletin of mathematical biophysics**, Springer, v. 13, n. 2, p. 107–117, 1951.

STEENSTRUP, M. Cluster-based networks. *In*: **Ad hoc networking**. [*S. l.*: *s. n.*], 2001. p. 75–138.

TAYLOR, J. **Introduction to error analysis, the study of uncertainties in physical measurements**. [*S. l.*: *s. n.*], 1997.

TEUKOLSKY, S. A.; FLANNERY, B. P.; PRESS, W.; VETTERLING, W. Numerical recipes in c. **SMR**, v. 693, n. 1, p. 59–70, 1992.

TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. *In*: **Social networks**. [*S. l.*]: Elsevier, 1977. p. 179–197.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **Nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

YAN, X.; SHALIZI, C.; JENSEN, J. E.; KRZAKALA, F.; MOORE, C.; ZDEBOROVÁ, L.; ZHANG, P.; ZHU, Y. Model selection for degree-corrected block models. **Journal of Statistical Mechanics:** theory and Experiment, IOP Publishing, v. 2014, n. 5, p. P05007, 2014.

APÊNDICE A – MÍNIMOS QUADRADOS GERAIS LINEARES

O método dos mínimos quadrados é bastante conhecido e utilizado na Física para se descobrir os parâmetros a e b de uma função que em geral é do tipo $a+bx$ dado somente o conjunto de dados (x_i, y_i) (TAYLOR, 1997). Assim, o objetivo ao utilizá-lo é tendo N dados (x_i, y_i) em que $i = 1, \ldots, N$, podemos modelar M parâmetros a_j, em que $j = 1, \ldots, M$. Para isso, supomos que função agora pode ser uma combinação linear de M funções de x. Por exemplo, no polinômio a seguir que tem grau $M-1$

$$y(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_M x^{M-1}. \tag{A.1}$$

No entanto, a função pode ser também uma combinação de senos e cossenos, por exemplo, e não somente um polinômio. Temos então que a forma geral deste modelo é

$$y(x) = \sum_{k=1}^{M} a_k X_k(x), \tag{A.2}$$

em que $X_1(x), \ldots, X_M(x)$ são funções fixas de x, chamadas de funções de base. Notemos algo importante, as funções $X_k(x)$ podem ser amplamente não lineares. Assim, neste método estamos estudando o termo "linear" refere-se apenas a dependência do modelo nos parâmetros a_k.

Dando sequência, para este modelo linear, usamos a denominada função mérito. Tal função serve para medir a concordância entre os dados e o modelo para uma certa escolha de parâmetros. Ela é definida de forma que ao minimizá-la, devemos ter os melhores valores para o ajuste. Ela é dada por

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2, \tag{A.3}$$

no qual o termo σ_i é uma medida de erro que em geral pode ser o desvio padrão que é presumível que se sabe. Porém, se as medidas de erros são desconhecidas, pode se usar $\sigma = 1$.

Há diferentes técnicas que minimizam χ^2, a técnica que nós iremos discutir é a solução por equações normais. Para este método, iremos usar a seguinte notação. Dada $\mathbf{A}$ uma matriz $N \times M$ componentes, em que temos M funções de base e N abscissas x_i e N medidas de erros σ_i, definimos

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i}, \tag{A.4}$$

em que **A** é denominada de matriz de design (design matrix) do problema de ajuste. Um modelo desta matriz está representado a seguir.

$$\begin{pmatrix} \frac{X_1(x_1)}{\sigma_1} & \frac{X_2(x_1)}{\sigma_1} & \cdots & \frac{X_M(x_1)}{\sigma_1} \\ \frac{X_1(x_2)}{\sigma_2} & \frac{X_2(x_2)}{\sigma_2} & \cdots & \frac{X_M(x_2)}{\sigma_2} \\ \vdots & & & \vdots \\ \frac{X_1(x_N)}{\sigma_N} & \frac{X_2(x_N)}{\sigma_1} & \cdots & \frac{X_M(x_N)}{\sigma_1} \end{pmatrix} \tag{A.5}$$

Também se define o vetor **b** de tamanho N por

$$b_i = \frac{y_i}{\sigma_i}. \tag{A.6}$$

Por fim, se denota um vetor **a** de tamanho M cuja as componentes são os parâmetros a serem ajustados.

Tendo tudo definido, vamos estudar como se chega no resulta almejado utilizando equações normais. Inicialmente, derivamos a equação (A.3) com respeito aos parâmetros a_k e igualamos a zero para se obter os valores de mínimos de χ^2, ficando com

$$\sum_{i=1}^{N} \frac{1}{\sigma^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(x_i) = 0, \tag{A.7}$$

no qual $k = 1, \ldots, M$. Podemos reescrever da seguinte forma

$$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k, \tag{A.8}$$

sendo

$$\alpha_{kj} = \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \tag{A.9a}$$

de forma equivalente,

$$[\alpha] = \mathbf{A}^T \cdot \mathbf{A} \tag{A.9b}$$

uma matriz $M \times M$. Além disso, temos

$$\beta_k = \sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2}, \tag{A.10a}$$

de forma equivalente,

$$[\beta] = \mathbf{A}^T \cdot \mathbf{b} \tag{A.10b}$$

um vetor de tamanho M.

As equações e (A.9) e (A.10) são chamadas de equações normais do problema de mínimos quadrados. Elas podem ser resolvidas para se chegar no parâmetro **a** por alguns métodos, dentre eles LU decomposition e backsubstitution (TEUKOLSKY *et al.*, 1992), Cholesky decomposition (GOLUB; LOAN, 2013) e Gauss-Jordan elimination (ATKINSON, 1991). No próximo apêndice está descrito o método Gauss-Jordan elemination que foi o utilizado neste trabalho. Dando continuidade, na forma matricial, as equações normais podem ser escritas como

$$[\alpha] \cdot \mathbf{a} = [\beta] \tag{A.11a}$$

ou

$$(\mathbf{A}^T \cdot \mathbf{A}) \cdot \mathbf{a} = \mathbf{A}^T \cdot \mathbf{b} \tag{A.11b}$$

A matriz inversa $C_{jk} \equiv [\alpha]^{-1}_{jk}$ está muito próxima da incerteza estimada do parâmetro **a**. Para se estimar essas incertezas, consideramos que

$$a_j = \sum_{k=1}^{M} [\alpha]^{-1}_{jk} \beta_k = \sum_{k=1}^{M} C_{jk} \left[\sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \right] \tag{A.12}$$

e a variância associada com a_j da seguinte forma

$$\sigma^2(a_j) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2. \tag{A.13}$$

Notemos que α_{jk} é independente de y_i, tal que

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=1}^{M} \frac{C_{jk} X_k(x_i)}{\sigma_i^2}. \tag{A.14}$$

Portanto,

$$\sigma^2(a_j) = \sum_{k=1}^{M} \sum_{l=1}^{M} C_{jk} C_{jl} \left[\sum_{i=1}^{N} \frac{X_k(x_i) X_l(x_i)}{\sigma_i^2} \right]. \tag{A.15}$$

O termo entre colchetes é a matriz $[\alpha]$ que é inversa de $[C]$, daí ficamos com a expressão reduzida a

$$\sigma^2(a_j) = C_{jj}. \tag{A.16}$$

Concluísse, portanto, que os elementos da diagonal de $[C]$ são as variâncias de parâmetro **a** ajustado.

APÊNDICE B – GAUSS-JORDAN ELIMINATION

Um conjunto de equações lineares algébricas é dada por:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \ldots & + & a_{1N}x_N = b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \ldots & + & a_{2N}x_N = b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & + & \ldots & + & a_{3N}x_N = b_3 \\ & & & & & & \ldots & & \ldots \\ a_{M1}x_1 & + & a_{M2}x_2 & + & a_{M3}x_3 & + & \ldots & + & a_{MN}x_N = b_M. \end{array} \tag{B.1}$$

Em que há N desconhecidos x_j com j variando de 1 a N que são relacionados com M equações. Além disso, os coeficientes, a_{ij} são valores conhecidos com i variando de 1 a M.

Para conhecedores um pouco de álgebra linear, sabemos que podemos reescrever a equação acima como,

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}, \tag{B.2}$$

sendo,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ a_{21} & a_{22} & \ldots & a_{2N} \\ & \ldots & & \\ a_{M1} & a_{M2} & \ldots & a_{MN} \end{pmatrix} \tag{B.3}$$

e

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \ldots \\ b_M \end{pmatrix} \tag{B.4}$$

Tendo esta notação, iniciemos o estudo sobre Gauss-Jordan elimination. Este método pode ser usado para conseguir a matriz inversa $\mathbf{A}^{-1}$. Ele é tão eficiente quanto a maioria dos outro métodos. Porém, ele tem como desvantagem o fato de todos os elementos de **b** terem que ser manipulados ao mesmo tempo.

Para evitar a utilização de muitas reticências $(\ldots)$ na notação, escreveremos apenas quatro equações e quatro incógnitas e com três vetores do lado direito já conhecidos. Assim, pode-se estender depois para matrizes maiores com tamanhos $N \times N$ e M conjuntos de vetores do lado direito.

Consideramos a seguinte notação:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \cdot \left[\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{41} \end{pmatrix} \bigsqcup \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \end{pmatrix} \bigsqcup \begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \\ x_{43} \end{pmatrix} \bigsqcup \begin{pmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \end{pmatrix} \right] = \left[\begin{pmatrix} b_{11} \\ b_{21} \\ b_{31} \\ b_{41} \end{pmatrix} \bigsqcup \begin{pmatrix} b_{12} \\ b_{22} \\ b_{32} \\ b_{42} \end{pmatrix} \bigsqcup \begin{pmatrix} b_{13} \\ b_{23} \\ b_{33} \\ b_{43} \end{pmatrix} \bigsqcup \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right] \tag{B.5}$$

em que o operador $\bigsqcup$ é um operador que representa uma matriz aumentada. Se ele é retirado a matriz faz-se uma matriz com os operandos do operador. Portanto, seria uma matriz 7×4 do lado direito, por exemplo. Além disso, podemos como sempre simplificar a equação acima, ficando com

$$[\mathbf{A}] \cdot [\mathbf{x}_1 \sqcup \mathbf{x}_2 \sqcup \mathbf{x}_3 \sqcup \mathbf{Y}] = [\mathbf{b}_1 \sqcup \mathbf{b}_2 \sqcup \mathbf{b}_3 \sqcup \mathbf{1}] \tag{B.6}$$

e solucionar o conjunto de equações a seguir.

$$\begin{gathered} \mathbf{A} \cdot \mathbf{x}_1 = \mathbf{b}_1 \quad \mathbf{A} \cdot \mathbf{x}_2 = \mathbf{b}_2 \quad \mathbf{A} \cdot \mathbf{x}_3 = \mathbf{b}_3 \\ \mathbf{A} \cdot \mathbf{Y} = \mathbf{1} \end{gathered} \tag{B.7}$$

Em que temos, $\mathbf{Y}$ como a inversa $\mathbf{A}^{-1}$ da matriz $\mathbf{A}$ por definição.

Usaremos, Gauss-Jordan elimination para calcular a matriz inversa $\mathbf{A}^{-1}$ da matriz $\mathbf{A}$. Este método consiste nos passos que serão explicados a seguir. Inicialmente, a primeira linha é dividida pelo elemento a_{11}, então subtrai-se múltiplos da primeira linha, em sua nova forma, das outras linhas a fim de tornar zero todos os elementos da primeira coluna com exceção do elemento a_{11}. Damos sequência e dividimos a segunda linha por a_{22} e realizamos o mesmo processo descrito para anular todos os termos da segunda coluna com exceção do termo a_{22}. Segue-se neste processo até que a matriz $\mathbf{A}$ fique na forma de uma matriz identidade. Além disso, à medida que as operações são feitas em $\mathbf{A}$, elas também devem ser feitas nos $\mathbf{b}$'s e $\mathbf{1}$ correspondente. Ao passo que o processo altera $\mathbf{A}$ até que a matriz se torne $\mathbf{1}$, a matriz identidade do lado direito se torna a solução que desejamos, isto é, a matriz inversa $\mathbf{A}^{-1}$.

O elemento que dividimos é chamado de pivô. Ao usarmos o método Gauss-Jordan elimination sem o pivô, ele fica numericamente instável na presença de qualquer erro de arredondamento. Além disso, outra observação é de como não desejamos mexer nos elementos da

matriz identidade que construímos, podemos escolher elementos que estão nas linhas abaixo daquelas linhas que estamos prestes a normalizar e também de colunas à direita daqui está sendo eliminada.