UNIVERSIDADE DE SÃO PAULO

TRABALHO DE FORMATURA SUPERVISIONADO

Estudo de um método alternativo de limiarização para gerar grafos a partir de matrizes de dados

Autor:
Evandro Augusto Nunes
SANCHES

Supervisor:
Prof. Dr. Arnaldo
MANDEL

Trabalho de Conclusão de Curso submetido para atender aos requisitos para obtenção do grau de Bacharel em Ciência da Computação

Teoria dos Grafos
Departamento de Ciência da Computação

7 de Dezembro de 2016

UNIVERSIDADE DE SÃO PAULO

Resumo

Instituto de Matemática e Estatística
Departamento de Ciência da Computação

Bacharel em Ciência da Computação

Estudo de um método alternativo de limiarização para gerar grafos a partir de matrizes de dados

por Evandro Augusto Nunes SANCHES

Para modelar um problema no contexto da teoria dos grafos, mais precisamente na parte que trata de grafos sem pesos nas arestas, é comum gerar a matriz de adjacência do grafo a partir de uma matriz de dados experimentais, aplicando-se a ela um processo de limiarização. O presente texto é o detalhamento do estudo para elaborar um novo método capaz de gerar os valores dos limiares utilizados nesse processo. Esse método proposto foi desenvolvido de modo que o valor do limiar aplicado possa variar para cada elemento da matriz, i.e. aplicar $limiar_1$ para a_{ij} e $limiar_2$ para a_{ij+1}. Além disso o método visou garantir duas propriedades ao processo: primeiro, o método deve considerar características específicas do problema, recebendo, como um parâmetro, informações relevantes para gerar estes limiares (*propriedade 1*); e segundo, o conjunto de limiares utilizados devem ser, em média, iguais ao valor de limiar utilizado em uma limiarização com o método convencional (*propriedade 2*).

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

EEG **E**letro**e**ncefalo**g**rama

Lista de Símbolos

A	Matriz de entrada / matriz de dados
B	Matriz de saída / matriz de adjacência resultante do processo
E	Conjunto de arestas($edges$)
G	Abreviação para $G(V, E)$ quando V e E são subentendido
$G(V, E)$	Grafo formado pelo conjuntos V de vértices e E de arestas($edges$)
H	Matriz de interferência
J	Função de ajuste das proporções
k	Constante normalizadora
L	Matriz de limiares ajustados. i.e. $\mathit{limiar}_{ij} \in L$
l	Função de limiarização pelo método ajustado
l_c	Função de limiarização pelo método controle
limiar	Valor absoluto para decisão na limiarização. I.e. $\begin{pmatrix} \mathit{limiar} = p \\ \text{ou} \\ \mathit{limiar}_{ij} = \pi_{ij} \end{pmatrix}$
M	Matriz de interferências perturbadas
n	Número de vértices de um grafo
P	Matriz de proporções ajustadas
p	Proporção de corte
V	Conjunto de vértices
α	Potência da função de perturbação particular $f(d) = d^{\alpha}$
η_{ij}	Interferência em $\{ij\}$
μ_{ij}	Proporção perturbada de $\{ij\}$
π_{ij}	Proporção ajustada de $\{ij\}$

Capítulo 1

Motivação

1.1 Contextualização

A teoria dos grafos, desde sua origem na solução do problema das pontes de Königsberg, presta-se como ferramenta auxiliar para resolução dos mais variados problemas. Uma vez que um problema é visto sob a luz dessa teoria, pode-se usar resultados já provados para inferir certas propriedades. Grafos são usados para modelagem principalmente por formarem uma classe bem simples e intuitiva de modelos. Por essa razão coleciona casos de sucesso, em que as propriedades matemáticas dos grafos têm interpretação relevante dentro do objeto sendo modelado.(Garcia-Ramos et al., 2016)

Tentando apoiar-se nessa base teórica já sedimentada, muito se faz para procurar mecanismos que produzam grafos que modelem de forma útil essas diversas estruturas – os reais objetos de estudo. Para adequar um problema à parte da teoria que lida com grafos sem pesos nas arestas, podendo esse problema ser representado por uma matriz $A \in \mathbb{R}^{n \times n}$, é comum aplicar um limiar de corte às entradas de A.(Zhou, Thompson e Siegle, 2009)

1.2 Limiarização

O processo de limiarização mencionado na seção anterior consiste em gerar uma nova matriz $B \in \{0,1\}^{n \times n}$ com cada entrada $b_{i,j}$ obedecendo:

$$b_{i,j} = \begin{cases} 1 & \text{, se } a_{i,j} \leq \textit{limiar} \\ 0 & \text{, c.c} \end{cases} \tag{1.1}$$

De fato, o processo de limiarização realiza a ponte entre o problema e a parte da teoria em questão. Entretanto a escolha do *limiar* adotado é uma parte sensível que pode modificar profundamente a representação do problema. Para melhor explicar este ponto, considere o seguinte caso:
Processo 1 ($\textit{limiar} = 0,6$):

$$A = \begin{bmatrix} 0,8 & 0,3 & 0,6 \\ 0,4 & 0,3 & 0,6 \\ 0,3 & 0,7 & 0,6 \end{bmatrix} \xRightarrow[\textit{limiarização}]{\textit{limiar}=0,6} B_1 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Processo 2 ($\textit{limiar} = 0,5$):

$$A = \begin{bmatrix} 0,8 & 0,3 & 0,6 \\ 0,4 & 0,3 & 0,6 \\ 0,3 & 0,7 & 0,6 \end{bmatrix} \xRightarrow[\textit{limiarização}]{\textit{limiar}=0,5} B_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Observe que a variação no limiar fez desaparecer 3 entradas iguais a 1 em B_2 que existiam em B_1. Pode ser que o problema apresente características bastante diferentes quando modelado por B_1 em comparação com as apresentadas quando modelado por B_2.

1.2.1 Métodos de escolha do limiar

Dentre os métodos existentes para escolha do valor do limiar, um deles é como segue: Considere que os dados de um determinado problema, quando limiarizados, produzam matrizes de adjacência de grafos com certa propriedade. Agora, procure o menor valor de limiar que garanta um grafo com esta propriedade.(Van Wijk, Stam e Daffertshofer, 2010)

Exemplo 1.1:

Suponha que já saibamos de antemão que o problema em questão, ao ser modelado, gere grafos conexos. Podemos determinar o valor para o limiar deste problema da seguinte forma: Escolha um valor de limiar pequeno, um valor que ao ser aplicado no processo de limiarização resulte em uma matriz de adjacência B de um grafo desconexo, agora incremente gradativamente o limiar até que B represente um grafo conexo.

Outro método para conseguir o limiar consiste em: defina um valor fixo baseando sua escolha em alguma propriedade do processo de criação da matriz A.

Exemplo 1.2:

Imagine que sejam coletados vetores de dados para diferentes partes do cérebro de um indivíduo, cada vetor contendo os níveis de oxigenação do sangue para uma das partes (uma medida indireta de atividade cerebral). Destas medições, calcula-se a correlação de Spearman sobre o conjunto de valores médios para cada par de vetores(regiões). Como os valores são calculados sobre médias do valor de cada entrada, pode-se utilizar o Teorema central do limite para assumir que cada correlação tem distribuição normal e executar um teste de hipótese calculando o *p-valor* de cada par. Ao final, como valor para o limiar, adota-se um ponto tal que o *p-valor* seja $0,05$. Em outras palavras, uma matriz de adjacência é construída considerando um *p-valor* $< 0,05$ como 1 e 0 caso contrário. (Takahashi et al., 2012)

Os métodos descritos acima compartilham um importante fato: um limiar adotado é único para todas as entradas da matriz. Ainda pode-se ressaltar que no primeiro caso, onde ocorre a escolha de menor limiar que gere grafos com uma dada propriedade, os limiares adotados não apresentam relação imediata entre cada escolha de limiar. Isto é, um limiar escolhido para a matriz A_1 não está diretamente vinculado a um limiar escolhido para uma outra matriz A_2.

Para muitas aplicações envolvendo o processo de limiarização, nem o fato de o limiar adotado ser único para todas as entradas da matriz A e nem o fato dele variar de amostra para amostra se define como um problema para o processo. Entretanto, podemos imaginar um cenário no qual deseja-se que cada entrada da matriz obtenha um limiar distinto. Nesse cenário, pode-se desejar ainda que os limiares sejam escolhidos de modo a agregar apenas informações do problema e do método de coleta ao processo (*propriedade 1*). Nesse caso os dois métodos mencionados anteriormente já não se aplicam.

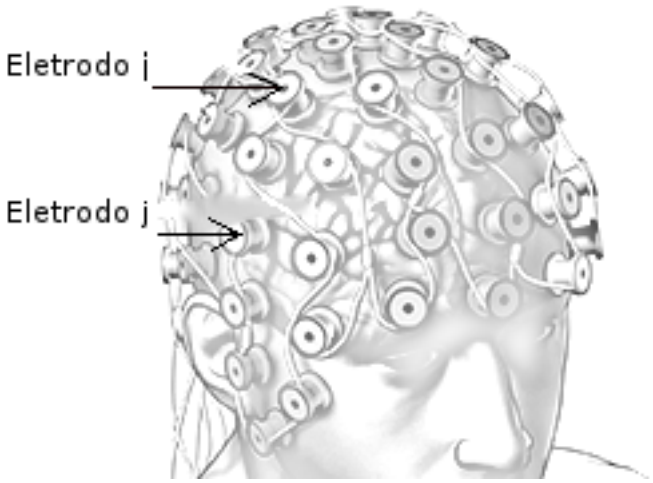


FIGURA 1.1: Destaque dos eletrodos número i e número j de uma coleta de EEG.
FONTE: modificado de
http://www.saintlukeshealthsystem.org/health-library/electroencephalogram-eeg

1.2.2 Método alternativo (ajustado)

Na missão de defender que existam casos como o que foi imaginado ao final da seção anterior, agora será ilustrado um caso no qual estas três propriedades são necessárias.

Exemplo 1.3:

Pensemos em uma coleta de eletroencefalograma(EEG) onde atribuímos números de 1 a n aos eletrodos (veja o esquema na figura 1.1). Suponha que matrizes $A \in \mathbb{R}^{n \times n}$ sejam criadas a partir dos dados experimentais, contendo na entrada $a_{i,j}$:= "*correlação de ativação entre os eletrodos i e j*". Suponha ainda que estas matrizes sejam adquiridas, todas, em um intervalo de tempo pequeno e com o mesmo conjunto de aparelho e acessórios. Nesse caso em particular, podemos imaginar que o intervalo reduzido e as condições semelhantes de coleta sejam suficientes para que adotemos os mesmos limiares para todas das matrizes A. Mais importante que isso, podemos imaginar que exista alguma interferência entre eletrodos posicionados em regiões muito próximas na cabeça do indivíduo. Sendo assim, o valor do limiar entre dois eletrodos muito próximos um do outro deveria considerar esta interferência e ser mais rigoroso. Por outro lado, para dois eletrodos muito distantes, o limiar deveria ser relaxado.

Tendo em vista essas características, um método adequado para a escolha dos valores limiares deve observar as seguintes propriedades:

Propriedade 1. *Cada limiar$_{i,j}$ utilizado no processo sugerido depende apenas do problema e do método de coleta de dados.*

Propriedade 2. *Para uma matriz A, com entradas $a_{ij} \sim U[0,1]$ i.i.d., o valor do grau médio do grafo limiarizado pelo processo sugerido deve coincidir com o grau médio de um grafo gerado na limiarização pelo processo controle.*

Capítulo 2

Objetivo

Como discutido no capítulo anterior, existem diversas formas de se escolher um valor para ser utilizado no processo de limiarização de uma matriz. Não obstante, um novo modelo de escolha pode ser útil se: permitir a adoção de limiares diferentes para cada entrada de uma mesma matriz a ser limiarizada; e considerar aspectos específicos de um dado problema em vez de valores obtidos em cada amostra coletada (*propriedade 1*).

Nesse sentido, o presente trabalho se dedicou à composição de um novo método que se ajustasse a essas exigências. Para alcançar esse feito, primeiramente foi preciso modificar a equação 1.1 da seguinte forma:

$$b_{ij} = \begin{cases} 1 & \text{, se } a_{ij} \leq \mathit{limiar}_{ij} \\ 0 & \text{, c.c} \end{cases} \tag{2.1}$$

Observe que a nova equação agora considera um limiar_{ij} específico para a entrada a_{ij} ao nosso método.

O próximo passo foi definir um processo sistemático de calcular os valores para limiar_{ij} de um modo que incorporasse as duas propriedades. Para isso, interpretou-se o processo de definir os limiares utilizados no processo como sendo a aplicação da seguinte função aos parâmetros do problema em questão:

$$\begin{aligned} &f : \mathbb{R} \text{ x } \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}^{n \times n} \\ &f(p, H) \longrightarrow P \end{aligned} \qquad \begin{aligned} &\text{, onde } p \text{ := valor base de corte,} \\ &H \text{ := matriz de interferência e} \\ &P \text{ := matriz de proporções ajustadas} \end{aligned} \tag{2.2}$$

Esta função calcula uma nova matriz $P \in \mathbb{R}^{n \times n}$ contendo um limiar_{ij} para cada entrada a_{ij} da matriz A a ser limiarizada. Para tanto, ela recebe um valor base de corte $p \in \mathbb{R}$ e uma matriz de interferência $H \in \mathbb{R}^{n \times n}$. A ideia é utilizar uma matriz de interferência H que de alguma forma esteja relacionada com a estrutura do problema. Assim, partindo de H podemos definir um limiar_{ij} para cada entrada $a_{ij} \in A$ como uma função dependente apenas da entrada $\eta_{ij} \in H$ e do valor base p. Como a matriz H detém informações do problema e a função f não depende dos valores da matriz A sendo limiarizada, os limiares gerados por f são tais que o método que a utilize possuirá a *propriedade 1*.

Para ilustrar como seria esse processo e deixar um pouco mais concreta a forma como H pode conter informações do problema, retomaremos o exemplo 1.3 para imaginar como seria construída a matriz H desse problema.

Exemplo 2.1:
Como vimos no exemplo 1.3, uma possível aplicação para o processo de

limiarização ajustada é no caso do processamento de dados de EEG. Nesse caso, uma matriz de correlação A é criada contendo na entrada a_{ij} a correlação entre o eletrodo i e o eletrodo j. Agora imagine que o posicionamento dos eletrodos sejam projetados em uma superfície que se assemelhe à cabeça do indivíduo. Uma forma de criar H é ter em cada entrada $\eta_{ij} \in H$ o valor da distância euclidiana entre o eletrodo i e o eletrodo j nessa superfície. Em um relaxamento simples desse processo podemos aproximar a cabeça do indivíduo por uma esfera. Veja um esquema na figura 2.1.

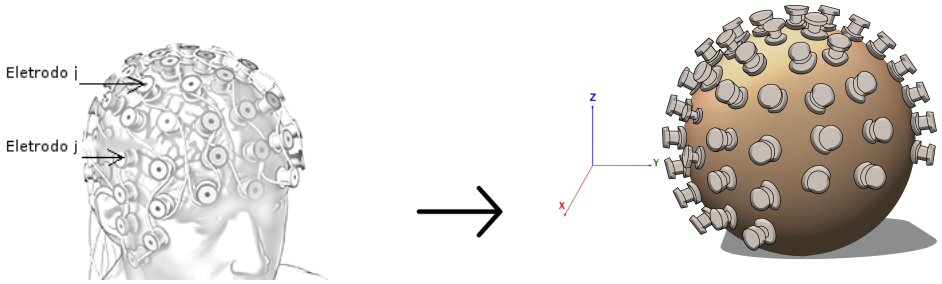


FIGURA 2.1: Posicionamento dos eletrodos em uma superfície esférica para o cálculo da distância euclidiana

O objetivo deste trabalho foi projetar um função f e estudar as possíveis propriedades alcançadas.

Capítulo 3

Materiais e Métodos

3.1 Visão geral

O presente estudo se caracteriza como uma pesquisa exploratória na definição de Antônio Carlos Gil, 1988 *apud* Bertucci, 2013. Sendo assim, este estudo tem por objetivo principal o aprofundamento das ideias sobre a limiarização. Para alcançar tal aprofundamento, este estudo se vale da proposição de um método alternativo aos encontrados hoje, o método ajustado.

O processo de criação do método ajustado foi dividido em duas etapas, uma etapa de formulação teórica e uma etapa de testes. A seguir uma breve descrição de cada uma delas:

3.1.1 Formulação teórica

Nesta etapa utilizou-se uma abordagem formal para descrever matematicamente o processo de limiarização pelo método ajustado. Desta forma, a definição 2.1 foi reformulada e a definição 2.2 foi decomposta para melhor investigar o processo.

Após concluídas as novas definições, o modelo teórico alcançado foi utilizado para estimar o valor da distância de Hamming para algumas matrizes de interferência selecionadas. Estas estimativas serviram para inferir os efeitos do método ajustado.

3.1.2 Etapa de testes

Nesta etapa, o processo estudado na formulação teórica foi implementado em uma linguagem de programação para observar seus efeitos sobre um conjunto de dados em particular.

A linguagem selecionada foi R e utilizou-se o ambiente de desenvolvimento R versão 3.3.1-1~jess que é oferecida no repositório CRAN - *Comprehensive R Archive Network*. (R-project, 2016)

Além do pacote básico do ambiente de desenvolvimento R, foi instalado o módulo igraph versão 1.0.0 e suas dependências. (Igraph, 2015)

Os códigos escritos na linguagem R foram executados na máquina com hostname *brucutu* da rede de computadores do Instituto de Matemática e Estatística (*redeime*). Segue a configuração da máquina:

- Processadores: 2 x (Intel(R) Xeon(R) CPU E5645 @ 2.40GHz)
- Memória RAM: 14 x (8GB) - totalizando 112 GB
- Motherboard: Supermicro X8DTN

- Kernel Linux: 3.16.7-ckt25-2+deb8u3 (2016-07-02)
- Arquitetura: x86_64

Estes códigos foram executados sobre um conjunto de matrizes de dados $A \in \mathbb{R}^{n \times n}$ geradas pseudo aleatoriamente. Para gerar as matrizes A, utilizou-se a função runif() do pacote básico do R. Desta forma, cada entrada $a_{ij} \in A$ foi sorteada uniformemente no intervalo $]0, 1[$.

Para o número de vértices submetidos ao processo, adotou-se $n = n_1 \cdot n_2 = 15 \cdot 15 = 225$.

Foram selecionados cinco casos de estudo, cada um deles referente a um espaço métrico diferente. Os espaços métricos adotados foram gerados considerando o conjunto dos n vértices e como função distância uma função que dado o par (ij) retorna a distância entre o vértice i e o vértice j em um grafo conhecido. Como grafo conhecido, utilizou-se os seguintes grafos [1]:

- Caminho (*Path*)
- Circuito (*Cycle*)
- Estrela (*Star*)
- Tabuleiro (*Board*)
- Tabuleiro Toroidal (*Toroidal Board*)

Para cada caso, realizou-se o processo de limiarização ajustada, como definido no capítulo 4, com proporções de corte $p = 0,25$, $p = 0,5$ e $p = 0,75$. Também foi realizada uma limiarização pelo procedimento convencional, aqui batizada de limiarização controle, com as mesmas proporções de corte. A partir do grafo gerado em cada limiarização, calculou-se as seguintes estatísticas [2]:

- Grau médio (*Degree Average*)
- Índice de Hamming (*Hamming Index*)
- Fechamento (*Closeness*)
- Centralidade (*Coreness*)
- Transitividade (*Transitivity*)
- Agrupamento (*Clustering*)
- Modularidade (*Modularity*)
- Tamanho do cluster

Esse processo de criar a matriz A, definir os limiares ajustados e limiarizar a matriz A pelos métodos controle e ajustado, foi repetido 100 vezes. Os resultados de cada repetição foi armazenado e a média e o desvios padrão de cada estatística foram calculados. Estes resultados podem ser vistos nas tabelas 7.1, 7.2 e 7.3. As rotinas de produção e análise dos dados utilizadas nessa etapa serão explicadas em detalhes no capítulo 6.

[1] Uma descrição de cada grafo pode ser encontrada no apêndice A.

[2] Uma descrição de cada estatística pode ser encontrada no apêndice B.

Capítulo 4

Formulação teórica

4.1 Processo genérico e o *limiar*

O processo de limiarização se resume a, dada uma matriz de entrada A, decidir para cada elemento a_{ij} de A um valor $b_{ij} \in \{0, 1\}$ para uma matriz de saída B. Esse método de decisão consiste em avaliar se a_{ij} é menor, maior ou igual a um valor de corte. Tal valor de corte recebe o nome de limiar.

Retomando a equação 1.1, devemos lembrar que o método de decisão pode ser descrito genericamente como sendo o modo de determinar B por:

$$b_{ij} = \begin{cases} 1 & \text{, se } a_{ij} \leq \textit{limiar} \\ 0 & \text{, c.c.} \end{cases}$$

Definição 4.1. Uma *proporção de corte* é um real $p \in [0, 1]$.

O *limiar* utilizado na equação acima também pode ser entendido como o valor que represente uma porcentagem p do maior valor possível para as entradas a_{ij} de A. Sendo assim temos a seguinte definição:

Definição 4.2. Seja $A \in \mathbb{R}^{n \times n}$ t.q. $a_{ij} \in [0, 1]$. E seja $p \in [0, 1]$ uma proporção de corte. Então definimos *limiar* $:= p$.

Para esse trabalho, os termos proporção de corte e limiar são equivalentes.

Com o *limiar* definido desta forma, podemos definir o processo de limiarização pelo método controle e com algumas modificações, chegar na definição do processo pelo método ajustado.

4.2 Método controle

Partindo da definição 4.2, podemos enxergar o processo de limiarização, utilizando o método controle, como a aplicação de uma *função de limiarização* a uma matriz A. Chamamos essa função de l_c e a definimos como segue:

Definição 4.3. Seja $A \in \mathbb{R}^{n \times n}$ t.q. $a_{ij} \in [0, 1]$ e A represente uma matriz de dados coletados de um certo problema. Seja também $p \in [0, 1]$ uma proporção de corte. Nestas condições, a *função de limiarização pelo método controle* é definida como:

$$l_c : \mathbb{R}^{n \times n} \text{ x } \mathbb{R} \longrightarrow \mathbb{R}^{n \times n}$$
$$l_c(A, p) = (b_{ij})_{n \times n}, \text{ onde } b_{ij} = \begin{cases} 1 & \text{, se } a_{ij} \leq p \\ 0 & \text{, c.c.} \end{cases}$$

4.3 Método ajustado

Queremos agora sugerir um método que por meio de modificações na definição 4.3 consiga incorporar algumas propriedades ao processo descrito anteriormente.

O primeiro objetivo é fazer com que o processo considere limiares distintos para cada entrada a_{ij} da matriz. Para atingir este feito, vamos modificar o valor de p para cada entrada da matriz A formando assim uma *Matriz de proporções ajustadas*.

Definição 4.4. Uma *Matriz de proporções ajustadas* P é o resultado da multiplicação dos valores de interferência perturbada pelo limiar base ($\pi_{ij} = \mu_{ij} \cdot p$). Assim π_{ij} está em $]0, 1[$ e contém o valor da proporção de corte para a entrada a_{ij}.

Enquanto na definição 4.3 a função l_c recebe apenas um valor p de proporção de corte, nossa primeira definição da *função de limiarização pelo método ajustado* recebe uma matriz de proporções de corte P.

Definição 4.5. Seja $A \in \mathbb{R}^{n\times n}$ t.q. $a_{ij} \in [0, 1]$ e A represente uma matriz de dados coletados de um certo problema. Seja também $P \in \mathbb{R}^{n\times n}$ t.q. $\pi_{ij} \in]0, 1[$ e P seja uma *Matriz de proporções ajustadas* para o problema em questão. Nestas condições, a *função de limiarização pelo método ajustado* é definida como:

$$l : \mathbb{R}^{n\times n} \text{ x } \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$$
$$l(A, P) = (b_{ij})_{n\times n}, \text{ onde } b_{ij} = \begin{cases} 1 & \text{, se } a_{ij} \leq \pi_{ij} \\ 0 & \text{, c.c.} \end{cases}$$

Para conseguir a *propriedade 1* é preciso que a construção de P não dependa dos valores de A e também que os valores π expressem alguma característica já conhecida do tipo de matriz A em processamento. Então, a construção de P deve depender de algum parâmetro relacionado ao problema. Chamamos esse parâmetro de interferência e o denotamos por η. Como queremos manter as propriedades anteriores, cada entrada π_{ij}precisa de um parâmetro η_{ij} o que nos leva a formar uma *Matriz de interferência*.

Definição 4.6. Uma *Matriz de interferência* H é tal que na entrada $\eta_{ij} \in H$ está um valor numérico que reflita alguma característica da modelagem referente à aresta (ij).

Para completar nossa formulação teórica, falta um último tópico: como utilizar as informações de H para gerar a *Matriz de proporções ajustadas* P. Sabemos que H contém informações sobre as relações entre os vértices do problema, mas não sabemos como estas informações devem impactar os valores π_{ij}, em outras palavras, o quanto um valor $\eta_{ij} = d$ modifica a entrada π_{ij}. Suponha que saibamos como essa influência ocorra, poderíamos então criar uma *Matriz de interferências perturbadas* e utilizá-la para gerar P.

Definição 4.7. Uma *Matriz de interferências perturbadas* M é o resultado de aplicar a função de perturbação aos valores de interferência da matriz H.

Chegamos agora ao ponto crítico desse estudo, o modo como criamos M define grandemente todo o processo de *limiarização pelo método ajustado*. Já que existem diversas formas de realizar essa tarefa, vamos deixar essa

etapa como uma variável do método ajustado. Sendo assim, utilizamos a definição genérica a seguir de uma função que cumpra esse papel, uma *função de perturbação*.

Definição 4.8. Uma *função de perturbação* é em princípio qualquer função crescente.

Mesmo com a matriz M definida, continuamos tendo muitas formas de gerar P. Contudo, consideramos apropriado utilizar o valor de p como valor base das entradas π_{ij}. Mais do que isso, consideramos apropriado que a média aritmética dos valores de P seja o próprio p (*propriedade* 2). Manter essa média igual a p é uma tentativa de manter a densidade de arestas do grafo gerado pelo processo de *limiarização pelo método ajustado* igual a densidade de arestas do grafo gerado pelo processo de *limiarização pelo método controle*. Desta forma as entradas de P contém simplesmente $\pi_{ij} = k \cdot p \cdot \mu_{ij}$, onde k é uma constante normalizadora que garante que a média aritmética de P seja p.

Definição 4.9. Seja $p \in]0, 1[$ uma proporção de corte. Seja $k \in \mathbb{R}$ uma constante normalizadora que garanta que a média das entradas de P seja igual a p. E seja também $M \in \mathbb{R}^{n \times n}$ uma Matriz de interferências perturbadas. Nestas condições, a *função de ajuste das proporções* é definida como:

$$J : \mathbb{R} \text{ x } \mathbb{R} \text{ x } \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}^{n \times n}$$
$$J(k, p, M) = (\pi_{ij})_{n \times n}, \text{ onde } \pi_{ij} = k \cdot p \cdot \mu_{ij}$$

Observe que segundo a definição acima, para manter a média de P igual a p basta que k seja igual ao inverso da média de M. Ou seja, $k = \frac{1}{\bar{\mu}}$. De fato,

$$\bar{P} = p \Leftrightarrow p = \bar{P} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij}}{n^2} \Leftrightarrow p = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} k \cdot p \cdot \mu_{ij}}{n^2} \Leftrightarrow p = p \cdot k \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}}{n^2} \Leftrightarrow$$

$$\Leftrightarrow 1 = k \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}}{n^2} \Leftrightarrow \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}} = k \Leftrightarrow k = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}} = \frac{1}{\bar{\mu}}$$

$$\text{obs: } p \neq 0 \text{ e } \sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij} \neq 0$$

Munidos dessas definições, podemos agora formalizar o processo de *limiarização pelo método ajustado*.

Definição 4.10. Seja A uma matriz de dados referentes a um dado problema. Seja H uma matriz de interferência para esse mesmo problema. Seja p uma proporção de corte para a limiarização. E seja f uma função de perturbação adequada para o problema. Nestas condições o processo de *limiarização pelo método ajustado* tem as seguintes etapas:

- criação da Matriz de interferência perturbada M pela aplicação de f à H:

$$M \leftarrow f(H)$$

- cálculo do valor de normalização k:

$$k \leftarrow \frac{1}{\bar{\mu}}$$

- criação da Matriz de proporções ajustadas P pela aplicação da função de ajuste das proporções J à k, p, M:

$$P \leftarrow J(k, p, M)$$

- criação da Matriz de saída B pela aplicação da função de limiarização l à A e P:

$$B \leftarrow l(A, P)$$

4.3.1 H como distâncias em um espaço métrico:

Depois de definir o processo de *limiarização pelo método ajustado* de forma ampla, vamos agora restringir nosso escopo de estudo e analisar uma classe específica de problemas. Lembre-se que esse estudo trata da modelagem de um problema no campo dos grafos o que implica que as matrizes de adjacência geradas na limiarização são simétricas. Vamos então, nos concentrar em problemas que geram matrizes de entrada A simétricas. Mais do que isso, nos concentraremos naqueles que geram A simétricas e para os quais as entradas $a_{i,i}$ possam ser desconsideradas (não existe loop no grafo gerado).

Podemos confirmar a relevância de estudar essa classe em particular simplesmente recordando de um fato: toda matriz contendo na entrada a_{ij} o valor da correlação entre os itens i e j é simétrica e contém 1's na diagonal principal.

Como nossa matriz A é dessa forma, precisamos encontrar uma matriz de interferência que seja adequada para este caso. Já foi dito que H pode armazenar qualquer valor em suas entradas, desde que esses valores expressem alguma informação do problema. Apesar disso, vamos restringir o modo como ela, a matriz H, é construída. Assim teremos uma matriz de interferência adequada ao escopo de problemas reduzido. Queremos que H seja também simétrica e com a diagonal principal desconsiderável. Uma boa forma de garantir estas propriedades é ter em $\eta_{ij} \in H$ as distâncias entre os elementos i e j de um espaço métrico a nossa escolha.

Nas palavras do professor Elon Lima, 1975:

- Espaço métrico é uma dupla (M, d), onde M é um conjunto e d é uma métrica em M.

- Métrica em um conjunto M é uma função $d : M \times M \longrightarrow \mathbb{R}$, que associa a cada par ordenado de elementos $x, y \in M$ um real $d(x, y)$ chamado a distância de x a y, de modo que sejam satisfeitas as seguintes condições para quaisquer $x, y, z \in M$:

  d1) $d(x, x) = 0$;

  d2) Se $x \neq y$, então $d(x, y) > 0$;

  d3) $d(x, y) = d(y, x)$;

  d4) $d(x, z) \leq d(x, y) + d(y, z)$.

Observe que se utilizarmos uma métrica para gerar H, pela condição d1 teremos $\eta_{i,i} = 0$, ou seja, podemos desprezar a diagonal principal. Além disso, pela condição d3, teremos a garantia de que H é simétrica. As condições d2 e d4 não precisariam ser exigidas, mas em algumas contas utilizamos o fato de que $\eta_{ij} > 0$ para $i \neq j$, condição d2. [1]

Para manter a coerência, vamos supor que a função de perturbação f gere uma Matriz de interferências perturbadas com essas mesmas propriedades caso H as tenha. Ou seja, se H é *simétrica* e com *diagonal principal desconsiderável* $\Rightarrow$ $f(H) = M$ é *simétrica* e com *diagonal principal desconsiderável*.

Temos então que as matrizes A, H e M são próprias dos problemas contidos no escopo reduzido que estudamos. Ao aplicarmos o processo de limiarização em A, tanto a *limiarização pelo método ajustado* quanto a *limiarização pelo método controle* gerarão uma matriz resultante B também simétrica e com diagonal principal desconsiderável. Sendo assim, podemos simplificar o processo de limiarização, independentemente do método adotado.

A simplificação é, em toda etapa do processo, considerar apenas os valores das entradas acima da diagonal principal de cada matriz. Assim a matriz resultante seria *triangular superior*. Ao final do processo seria necessário replicar os valores das entradas superiores a diagonal principal nos valores correspondentes inferiores a diagonal principal. Em outras palavras fazer: para $1 \leq i < j \leq n$, $b_{j,i} = b_{ij}$.

Para efeito das contas seguintes, consideraremos de cada matriz apenas os valores tais que $1 \leq i < j \leq n$. Ou seja apenas $\binom{n}{2}$ valores. De imediato já podemos ajustar o valor de $\bar{\mu}$ pois agora a média será apenas entre os $\binom{n}{2}$ valores.

Definição 4.11. Seja M uma Matriz de interferências perturbadas. Então a Constante normalizadora k é da seguinte forma:

$$k = \frac{1}{\bar{\mu}} = \frac{1}{\frac{\sum\limits_{1 \leq i < j \leq n} \mu_{ij}}{\binom{n}{2}}} = \frac{\binom{n}{2}}{\sum\limits_{1 \leq i < j \leq n} \mu_{ij}}$$

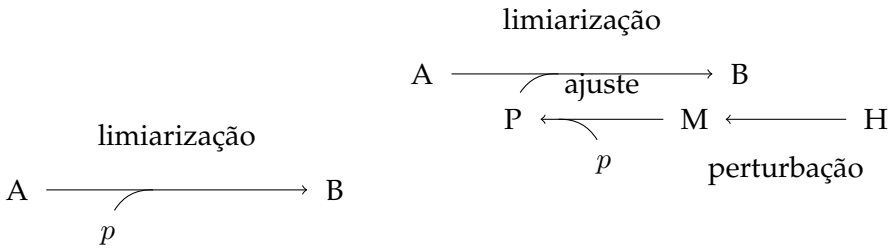


FIGURA 4.1: Esquemas de limiarização utilizando o método controle (esquerda) e utilizando o método ajustado (direita).

Chegamos a um ponto no qual temos a formulação teórica completa do processo que sugerimos (veja os esquemas da imagem 4.1). O que faremos

[1] Apesar de no início deste estudo termos imaginado que a desigualdade triangular, condição d4, fosse muito útil, não chegamos a utilizá-la de fato.

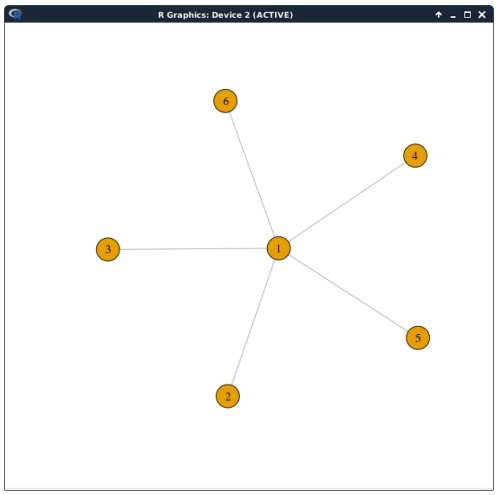


FIGURA 4.2: Grafo $S_6 \in S_n$: Estrela com 6 vértices.
FONTE: Imagem produzida com o módulo para o ambiente de desenvolvimento em R chamado Igraph, 2015.

agora é escolher e utilizar alguns espaços métricos no processo de *limiarização pelo método ajustado* para inferir seus efeitos.

Escolhemos algumas famílias de grafos com estruturas bastante distintas e definimos as distâncias nesses grafos como as várias métricas. Assim, cada grafo foi considerado um espaço métrico e, a partir das distâncias entre os vértices i e j, definiu-se o valor da entrada η_{ij}. Vejamos um exemplo:

Exemplo 4.1:

Suponha que nosso problema gerou uma matriz $A \in \mathbb{R}^{6\times 6}$ contendo a relação entre cada par dentre 6 elementos. Suponha ainda que saibamos que uma propriedade desses elementos possa ser representada como as distâncias em uma estrela S_6, veja este grafo na figura 4.2. Desta forma, nossa H conteria na entrada η_{ij} a distância $d(ij)$ e ficaria assim:

$$H = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 & 2 & 2 \\ 1 & 2 & 0 & 2 & 2 & 2 \\ 1 & 2 & 2 & 0 & 2 & 2 \\ 1 & 2 & 2 & 2 & 0 & 2 \\ 1 & 2 & 2 & 2 & 2 & 0 \end{pmatrix} \end{array}$$

Para este estudo, utilizamos cinco famílias de grafos com estruturas bastante diferentes. Uma descrição de cada uma delas pode ser encontrada no apêndice A. As famílias utilizadas foram:

Caminhos (P_n)

Circuitos (C_n)

Estrelas (S_n)

Tabuleiros (Q_{n_1,n_2})

Tabuleiros Toroidais (T_{n_1,n_2})

4.4 Avaliação pela distância de Hamming

Para avaliar o efeito do processo de *limiarização pelo método ajustado* precisamos comparar o grafo gerado por esse processo com algum referencial. O referencial para esse caso claramente é o grafo resultante do processo de *limiarização pelo método controle*, mas o modo de comparação pode se dar por diferentes formas.

Neste estudo escolhemos utilizar um conjunto de estatísticas de grafos para observar os efeitos do processo sugerido. Falaremos mais delas no capítulo sobre o módulo de testes, capítulo 6.

Entretanto, uma dessas estatísticas nos pareceu bastante significativa para comparar os efeitos dos dois processos. Assim, escolhemos a distância de Hamming para ser o indicativo de quão diferente são os grafos formados por cada um deles.

Para esse trabalho, a distância de Hamming entre dois grafos é a distância de Hamming entre as matrizes de adjacência de cada grafo divida por dois. Ou seja, a distância de Hamming entre um grafo G_1 e um outro grafo G_2 é o número de entradas diferentes entre a matriz de adjacência $A(G_1)$ e a matriz de adjacência $A(G_2)$ dividido por dois. I.e. se

$$A(G_1) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array} \text{ e } A(G_2) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

então a distância de Hamming entre G_1 e G_2, denotada por dist_Hamm$_{G_1,G_2}$, tem valor igual a 2, pois quatro das entradas de $A(G_1)$ são diferentes das entradas de $A(G_2)$. A saber, as entradas diferentes são $a_{1,3}$, $a_{2,3}$, $a_{3,1}$ e $a_{3,2}$.

Para poder avaliar as alterações dos grafos devidas à limiarizações com diferentes parâmetros, foram feitas as duas normalizações citadas no apêndice B gerando o índice de Hamming.

O que faremos a seguir é verificar matematicamente o que se espera do valor desse parâmetro quando se compara os grafos gerados por ambos os processos de limiarização, *limiarização pelo método ajustado* e *limiarização pelo método controle*.

Capítulo 5

Esperança do Índice de Hamming

Neste capítulo, verificaremos matematicamente o que se espera do valor da distância de Hamming entre os grafos gerados pelo processo de *limiarização pelo método controle* e os grafos gerados pelo processo de *limiarização pelo método ajustado*.

Para alcançar o valor esperado para o índice de Hamming, basta normalizar o valor da esperança pelas duas maneiras descritas no apêndice B.

Em um primeiro momento faremos cálculos de probabilidade e esperança considerando uma função de perturbação f e uma matriz de interferência H genéricas. Em seguida olharemos para matrizes de interferência criadas a partir das distâncias em exemplares das famílias de grafos escolhidas.

5.1 Modelo genérico

Chamemos a matriz resultante do processo de *limiarização pelo método controle* de B_{ctl}. Sabemos que tal matriz é a matriz de adjacência do grafo G_{ctl}. Analogamente, a matriz gerada no processo de *limiarização pelo método ajustado*, B_{adj}, é a matriz de adjacência do grafo G_{adj}.

O objetivo desta sessão é calcular $\mathbb{E}(\text{dist\_Hamm}_{G_{\text{ctl}},G_{\text{adj}}})$. Ou seja, queremos a distância de Hamming entre B_{ctl} e B_{adj} dividida por dois. No restante desse trabalho, omitiremos os subíndice de dist_Hamm por comodidade e por supor que isso não prejudicará o entendimento do leitor. Denotaremos o valor estudado nessa seção simplesmente por $\mathbb{E}(\text{dist\_Hamm})$.

A dist_Hamm entre G_{ctl} e G_{adj} pode ser contabilizada a partir da soma dos dois seguintes valores parciais:

- arestas que estão em G_{ctl} e não estão em $G_{\text{adj}} \rightarrow$ metade [1] do número de entradas iguais a 1 em B_{ctl} e iguais a 0 em B_{adj}.
- arestas que não estão em G_{ctl} e estão em $G_{\text{adj}} \rightarrow$ metade do número de entradas iguais a 0 em B_{ctl} e iguais a 1 em B_{adj}.

Definição 5.1. Seja B_{ctl} a matriz resultante do processo de *limiarização pelo método controle* e G_{ctl} o respectivo grafo. Seja B_{adj} a matriz resultante do processo de *limiarização pelo método ajustado* e G_{adj} o respectivo grafo. Nestas condições dizemos que uma aresta:

[1] Lembre-se que trabalhamos com grafos e que a matriz de adjacência é simétrica. Assim, cada aresta $\{ij\} \in E(G)$ é representada por duas entradas na matriz de adjacência $A(G)$: $a_{ij} \in A(G)$ e $a_{j,i} \in A(G)$.

(a) *desaparece* se ela estiver em G_{ctl} e não estiver em G_{adj}. Ou seja, a aresta $\{ij\}$ *desaparece* se $b_{\text{ctl}_{ij}} = b_{\text{ctl}_{j,i}} = 1$ e $b_{\text{adj}_{ij}} = b_{\text{adj}_{j,i}} = 0$.

(b) *aparece* se ela não estiver em G_{ctl} e estiver em G_{adj}. Ou seja, a aresta $\{ij\}$ *aparece* se $b_{\text{ctl}_{ij}} = b_{\text{ctl}_{j,i}} = 0$ e $b_{\text{adj}_{ij}} = b_{\text{adj}_{j,i}} = 1$.

Com a definição 5.1 já podemos descrever os eventos e as variáveis aleatórias que serão utilizados para calcular $\mathbb{E}(\text{dist\_Hamm})$:

I seja $\text{des}_d :=$ o evento de uma aresta $\{ij\}$ cujo $\eta_{ij} = \eta_{j,i} = d$ desaparecer

II seja $\text{ap}_d :=$ o evento de uma aresta $\{ij\}$ cujo $\eta_{ij} = \eta_{j,i} = d$ aparecer

III seja $\text{Des}_d :=$ número de arestas $\{ij\}$ cujo $\eta_{ij} = \eta_{j,i} = d$ e que desaparecem

IV seja $\text{Ap}_d :=$ número de arestas $\{ij\}$ cujo $\eta_{ij} = \eta_{j,i} = d$ e que aparecem

V seja Des := número de arestas que desaparecem

VI seja Ap := número de arestas que aparecem

Desta forma é possível calcular dist_Hamm por suas parciais Ap e Des, pois dist_Hamm $=$ Ap $+$ Des. Consequentemente, podemos calcular:

$$\mathbb{E}(\text{dist\_Hamm}) = \mathbb{E}(\text{Ap}) + \mathbb{E}(\text{Des})$$

Para facilitar os próximos cálculos, também utilizaremos as seguintes definições:

Definição 5.2. Seja G_H o grafo gerado a partir da matriz de adjacência H. Tome um grafo G_H com n vértices, m arestas e diâmetro D. Seja $f()$ uma função de perturbação crescente. Defina:

(a) $r_d :=$ número de ocorrências da distância $d(u, v) = d$ onde $u, v \in V(G_H)$, vértices de G_H. Ou seja, metade do número de entradas de H iguais a d.

(b) $F := \sum_{1 \le i < j \le n} \mu_{ij} = \sum_{1 \le i < j \le n} f(\eta_{ij}) = \sum_{d=1}^{D} r_d \cdot f(d)$

(c) $k := \frac{1}{\bar{\mu}} = \frac{\binom{n}{2}}{F}$

(d) $\theta := max\ \{d \in \mathbb{N} \mid f(d) < \frac{F}{\binom{n}{2}}\}$

1 Queremos calcular $\mathbb{E}(\text{Des})$:

Sabemos que a aresta $\{ij\}$ desaparece se:

$$\begin{cases} \text{está em } G_{\text{ctl}} & \Rightarrow a_{ij} < p \\ \text{e} & \\ \text{não está em } G_{\text{adj}} & \Rightarrow a_{ij} > p \cdot k \cdot f(\eta_{ij}) \end{cases}$$

Ou seja, a aresta $\{ij\}$ desaparece se $p \cdot k \cdot f(\eta_{ij}) < a_{ij} < p$. Essa sentença só é válida se:

$$p \cdot k \cdot f(\eta_{ij}) < p \overset{p \neq 0}{\Rightarrow} k \cdot f(\eta_{ij}) < 1 \overset{k \neq 0}{\Rightarrow} f(\eta_{ij}) < \frac{1}{k}$$

Então,

$$\mathbb{P}(\{ij\}\ \textit{desaparecer}) = \begin{cases} p \cdot (1 - k \cdot f(\eta_{ij})) & \text{, se } \eta_{ij} \leq \theta \\ 0 & \text{, c.c} \end{cases}$$

$$\mathbb{P}(\text{des}_d) = \begin{cases} p \cdot (1 - k \cdot f(d)) & \text{, se } d \leq \theta \\ 0 & \text{, c.c} \end{cases}$$

E por fim,

$$\begin{aligned} \mathbb{E}(\text{Des}) &= \sum_{d=1}^{\theta} \mathbb{E}(\text{Des}_d) \\ &= \sum_{d=1}^{\theta} r_d \cdot \mathbb{P}(\text{des}_d) \\ &= \sum_{d=1}^{\theta} r_d \cdot p \cdot (1 - k \cdot f(d)) \\ &= p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=1}^{\theta} k \cdot r_d \cdot f(d) \right] \end{aligned}$$

2 Queremos calcular $\mathbb{E}(\text{Ap})$:

Sabemos que a aresta $\{ij\}$ aparece se: $\begin{cases} \text{não está em } G_{\text{ctl}} & \Rightarrow a_{ij} > p \\ \text{está em } G_{\text{adj}} & \Rightarrow a_{ij} < p \cdot k \cdot f(\eta_{ij}) \end{cases}$

Ou seja, a aresta $\{ij\}$ aparece se $p < a_{ij} < p \cdot k \cdot f(\eta_{ij})$.

Então,

$$\mathbb{P}(\{ij\}\ \textit{aparecer}) = \begin{cases} p \cdot (k \cdot f(\eta_{ij}) - 1) & \text{, se } \eta_{ij} > \theta \\ 0 & \text{, c.c} \end{cases}$$

$$\mathbb{P}(\text{ap}_d) = \begin{cases} p \cdot (k \cdot f(d) - 1) & \text{, se } d > \theta \\ 0 & \text{, c.c} \end{cases}$$

E por fim,

$$\begin{aligned} \mathbb{E}(\text{Ap}) &= \sum_{d=\theta+1}^{D} \mathbb{E}(\text{Ap}_d) \\ &= \sum_{d=\theta+1}^{D} r_d \cdot \mathbb{P}(\text{ap}_d) \\ &= \sum_{d=\theta+1}^{D} r_d \cdot p \cdot (k \cdot f(d) - 1) \\ &= p \cdot \left[\sum_{d=\theta+1}^{D} k \cdot r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \right] \end{aligned}$$

3 Queremos calcular $\mathbb{E}(\text{dist\_Hamm}) = \mathbb{E}(\text{Des}) + \mathbb{E}(\text{Ap})$

De 1 e 2 temos que:

$$\mathbb{E}(\text{Des}) = p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=1}^{\theta} k \cdot r_d \cdot f(d) \right]$$

$$\mathbb{E}(\text{Ap}) = p \cdot \left[\sum_{d=\theta+1}^{D} k \cdot r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \right]$$

Então,

$$
\begin{aligned}
\mathbb{E}(\text{dist\_Hamm}) &= \mathbb{E}(\text{Des}) + \mathbb{E}(\text{Ap}) \\
&= p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=1}^{\theta} k \cdot r_d \cdot f(d) \right] + p \cdot \left[\sum_{d=\theta+1}^{D} k \cdot r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \right] \\
&= p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=\theta+1}^{D} r_d + k \cdot \left(\sum_{d=\theta+1}^{D} r_d \cdot f(d) - \sum_{d=1}^{\theta} r_d \cdot f(d) \right) \right] \\
&= p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=\theta+1}^{D} r_d + \tfrac{\binom{n}{2}}{F} \cdot \left(\sum_{d=\theta+1}^{D} r_d \cdot f(d) - \sum_{d=1}^{\theta} r_d \cdot f(d) \right) \right] \\
&= p \cdot \left[\sum_{d=1}^{\theta} r_d - \sum_{d=\theta+1}^{D} r_d - \tfrac{\binom{n}{2}}{F} \cdot \left(\sum_{d=1}^{\theta} r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \cdot f(d) \right) \right] \\
&\overset{\text{(b)}}{=} p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tbinom{n}{2} - \tfrac{\binom{n}{2}}{F} \cdot \left(\sum_{d=1}^{\theta} r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \cdot f(d) \right) \right] \\
&\overset{\text{(c)}}{=} p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tbinom{n}{2} - \tfrac{\binom{n}{2}}{F} \cdot \left(2 \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) - F \right) \right] \\
&= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tbinom{n}{2} - \tbinom{n}{2} \cdot \left(\tfrac{2}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) - 1 \right) \right] \\
&= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tbinom{n}{2} - \tbinom{n}{2} \cdot \tfrac{2}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) + \tbinom{n}{2} \right] \\
&= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tbinom{n}{2} \cdot \tfrac{2}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right] \\
&= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tfrac{n \cdot (n-1)}{2} \cdot \tfrac{2}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right] \\
&= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tfrac{n \cdot (n-1)}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right]
\end{aligned}
$$

Assim, a fórmula genérica para a Esperança de Hamming é:

$$
\mathbb{E}(\text{dist\_Hamm}) = p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \frac{n \cdot (n-1)}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right] \tag{5.1}
$$

Observações 5.1.1 :

(a) : Seja $a + b = R$. Então,

$$a - b = a + a - a - b = 2 \cdot a - a - b = 2 \cdot a - (a + b) = 2 \cdot a - R$$

(b) : $\sum_{d=1}^{\theta} r_d + \sum_{d=\theta+1}^{D} r_d = \sum_{d=1}^{D} r_d = \binom{n}{2}$. Então,

$$\sum_{d=1}^{\theta} r_d - \sum_{d=\theta+1}^{D} r_d \overset{(a)}{=} 2 \cdot \sum_{d=1}^{\theta} r_d - \sum_{d=1}^{D} r_d = 2 \cdot \sum_{d=1}^{\theta} r_d - \binom{n}{2}$$

(c) :

$$\begin{aligned} \sum_{d=1}^{\theta} r_d \cdot f(d) - \sum_{d=\theta+1}^{D} r_d \cdot f(d) &\overset{(a)}{=} 2 \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) - \sum_{d=1}^{D} r_d \cdot f(d) \\ &= 2 \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) - F \end{aligned}$$

Já sabemos como é a forma da $\mathbb{E}(\text{dist\_Hamm})$. Mas ela ainda depende das interferências perturbadas. Em outras palavras, para ter uma ideia melhor de como é esse valor, precisamos olhar para alguns espaços métricos e algumas funções de perturbação.

5.2 Estrela

Vamos começar por um espaço métrico simples, distâncias em uma estrela. Nesse tipo de grafo temos apenas dois valores possíveis para as distâncias entre $\{ij\}$. Ou η_{ij} vale 1 ou vale 2. O valor 1 está vinculado com arestas entre o vértice central, no nosso caso o vértice 1, e qualquer outro vértice. O valor 2 está vinculado com arestas que não envolvam o vértice central. Já sabemos daqui que em uma estrela com qualquer número de vértices $D = 2$.

Sabemos também, que em uma estrela de n vértices, o número de arestas que envolvem o vértice 1 é $(n - 1)$. Em outras palavras, $r_1 = (n - 1)$. Todas as outras arestas tem distância 2, ou seja $r_2 = \binom{(n-1)}{2}$.

Como nosso modelo aplica uma normalização que força parte dos limiares aumentarem seu valor e parte deles diminuírem, e essa divisão é baseada nas distâncias da matriz de interferência. Sabemos que, no caso da estrela, as arestas relacionadas com a distância 1 diminuem seu limiar e as arestas com distância 2 o aumentam. Em outras palavras $\theta = 1$.

Agora, podemos substituir esses valores na definição 5.2-(b), obtendo:

$$\begin{aligned} F &:= \sum_{d=1}^{D} r_d \cdot f(d) \\ &= \sum_{d=1}^{2} r_d \cdot f(d) \\ &= r_1 \cdot f(1) + r_2 \cdot f(2) \\ &= (n-1) \cdot f(1) + \tbinom{(n-1)}{2} \cdot f(2) \\ &= (n-1) \cdot f(1) + \tfrac{(n-1)\cdot(n-2)}{2} \cdot f(2) \\ &= (n-1) \cdot \left[f(1) + \tfrac{(n-2)}{2} \cdot f(2) \right] \end{aligned}$$

E com isso já podemos refinar nosso valor esperado de dist_Hamm para o caso de utilizar uma estrela como espaço métrico.

$$\begin{aligned} \mathbb{E}(\text{dist\_Hamm}) &= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tfrac{n\cdot(n-1)}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right] \\ &= p \cdot \left[2 \cdot \sum_{d=1}^{1} r_d - \tfrac{n\cdot(n-1)}{F} \cdot \sum_{d=1}^{1} r_d \cdot f(d) \right] \\ &= p \cdot \left[2 \cdot r_1 - \tfrac{n\cdot(n-1)}{F} \cdot r_1 \cdot f(1) \right] \\ &= p \cdot \left[2 \cdot (n-1) - \tfrac{n\cdot(n-1)}{F} \cdot (n-1) \cdot f(1) \right] \\ &= p \cdot \left[2 \cdot (n-1) - \tfrac{n\cdot(n-1)}{(n-1)\cdot\left[f(1) + \frac{(n-2)}{2} \cdot f(2) \right]} \cdot (n-1) \cdot f(1) \right] \\ &= p \cdot \left[2 \cdot (n-1) - \tfrac{n}{f(1) + \frac{(n-2)}{2} \cdot f(2)} \cdot (n-1) \cdot f(1) \right] \\ &= p \cdot \left[2 \cdot (n-1) - \tfrac{n \cdot f(1)}{f(1) + \frac{(n-2)}{2} \cdot f(2)} \cdot (n-1) \right] \\ &= p \cdot 2 \cdot (n-1) \cdot \left[1 - \tfrac{n \cdot f(1)}{2 \cdot f(1) + (n-2) \cdot f(2)} \right] \end{aligned}$$

Aqui é o máximo que podemos chegar sem saber qual perturbação estamos utilizando.

5.2.1 Utilizando perturbação $f(d) = d^{\alpha}$

Vamos agora utilizar uma perturbação bastante trivial. Vamos utilizar $f(d) = d^{\alpha}$.

Sendo assim, $f(1) = 1^{\alpha} = 1$ e $f(2) = 2^{\alpha}$. A $\mathbb{E}$(dist_Hamm) pode ser modificada para:

$$\begin{aligned} \mathbb{E}(\text{dist\_Hamm}) &= p \cdot 2 \cdot (n-1) \cdot \left[1 - \tfrac{n \cdot f(1)}{2 \cdot f(1) + (n-2) \cdot f(2)} \right] \\ &= p \cdot 2 \cdot (n-1) \cdot \left[1 - \tfrac{n \cdot 1}{2 \cdot 1 + (n-2) \cdot 2^{\alpha}} \right] \\ &= p \cdot 2 \cdot (n-1) \cdot \left[1 - \tfrac{n}{2 + (n-2) \cdot 2^{\alpha}} \right] \end{aligned}$$

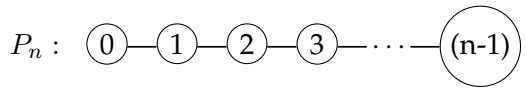
5.3 Caminho

O próximo espaço métrico que vamos investigar é um pouco mais complicado. Vamos utilizar um caminho P_n para gerar nossa matriz de interferências H. Suponha um caminho de tamanho n arbitrário. Sabemos então

que:

$$\begin{aligned} P_n: \quad & V(P_n) = \{0, 1, 2, 3, \cdots, (n-1)\} \\ & E(P_n) = \{\{ij\} \mid ij \in V(P_n) \text{ e } |i-j| = 1\} \end{aligned}$$

Uma representação para este caminho pode ser:



Para este grafo é fácil perceber que a distância entre dois vértices é dada pela formula $d(ij) = |i-j|$. Assim o diâmetro de P_n é a distância do vértice 0 ao vértice $(n-1)$, ou seja $D = (n-1)$.

Para saber quantas distâncias de tamanho d existem, calcular r_d, vamos definir a seguinte relação de equivalência para esse espaço métrico:

Definição 5.3. Seja G um grafo, $V(G)$ os vértices de G e $E(G)$ as arestas de G. Sejam $ij, r, s \in V(G)$ e $\{ij\}, \{r, s\} \in E(G)$. Seja $d : V(G) \times V(G) \to \mathbb{R}$ uma métrica. Nessas condições definimos a relação de equivalência $\sim$ da seguinte maneira:

$$\{ij\} \sim \{r, s\} \Leftrightarrow d(ij) = d(r, s)$$

Sobre esta definição de $\sim$ sabemos que a classe de equivalência dos elementos cujo $d(ij) = k$ para um k arbitrário, pode ser expressa por:

$$\begin{aligned} &[\{0, k\}] = \{\{ij\} \in E(G) \text{ t.q. } |i-j| = k\} \\ &[\{0, k\}] = \{\{0, k\}, \{1, k+1\}, \{2, k+2\}, \cdots, \{n-1-k, k+n-1-k = n-1\}\} \\ &|[\{0, k\}]| = n-1-k+1 = n-k \end{aligned}$$

Assim, sabemos que o valor de r_k nada mais é do que a cardinalidade da classe $[\{0, k\}]$, ou seja $r_k = |[\{0, k\}]| = n-k$. E para todo $d \in \{1, 2, \cdots n-1\}$, $r_d = n - d$.

Com isso já podemor melhoras um pouco a expressão que define F e a $\mathbb{E}$(dist_Hamm):

$$\begin{aligned} F &:= \sum_{d=1}^{D} r_d \cdot f(d) \\ &= \sum_{d=1}^{n-1} r_d \cdot f(d) \\ &= \sum_{d=1}^{n-1} (n-d) \cdot f(d) \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\text{dist\_Hamm}) &= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} r_d - \tfrac{n \cdot (n-1)}{F} \cdot \sum_{d=1}^{\theta} r_d \cdot f(d) \right] \\ &= p \cdot \left[2 \cdot \sum_{d=1}^{\theta} (n-d) - \tfrac{n \cdot (n-1)}{F} \cdot \sum_{d=1}^{\theta} (n-d) \cdot f(d) \right] \end{aligned}$$

Aqui é o máximo que podemos chegar sem saber qual perturbação estamos utilizando.

5.3.1 Utilizando perturbação $f(d) = d^\alpha$

Voltamos a utilizar novamente a perturbação $f(d) = d^\alpha$. Agora, com a perturbação definida, o primeiro passo é melhorar a expressão de F, assim

temos:

$$\begin{aligned} F &= \sum_{d=1}^{n-1} (n-d) \cdot f(d) \\ &= \sum_{d=1}^{n-1} (n-d) \cdot d^{\alpha} \end{aligned}$$

Como d^{α} é contínua, crescente e positiva, é possível chegar a limitantes para F da seguinte maneira:

$$\begin{gathered} \int_1^{n-2} (n-x)x^{\alpha} dx \leq \sum_{d=1}^{n-1} (n-d) \cdot d^{\alpha} \leq \int_1^{n-1} (n-x)x^{\alpha} dx \\ n \cdot \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_1^{n-2} - \left[\frac{x^{\alpha+2}}{\alpha+2}\right]_1^{n-2} \leq \sum_{d=1}^{n-1} (n-d) \cdot d^{\alpha} \leq n \cdot \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_1^{n-1} - \left[\frac{x^{\alpha+2}}{\alpha+2}\right]_1^{n-1} \\ n \cdot \left(\frac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\frac{(n-2)^{\alpha+2}-1}{\alpha+2}\right) \leq \sum_{d=1}^{n-1} (n-d) \cdot d^{\alpha} \leq n \cdot \left(\frac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \left(\frac{(n-1)^{\alpha+2}-1}{\alpha+2}\right) \end{gathered}$$

Ou seja:

$$n \cdot \left(\frac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\frac{(n-2)^{\alpha+2}-1}{\alpha+2}\right) \leq F \leq n \cdot \left(\frac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \left(\frac{(n-1)^{\alpha+2}-1}{\alpha+2}\right) \tag{5.2}$$

Para poder avaliar dist_Hamm, precisamos estimar o valor de θ para descobrir o limite das somatórias utilizadas na fórmula de $\mathbb{E}$(dist_Hamm). Sabemos pela definição 5.2-(d) que:

$$\theta := max\ \{d \in \mathbb{N} \mid f(d) < \frac{F}{\binom{n}{2}}\}$$

Já sabemos da inequação 5.2 que:

$$
\begin{aligned}
n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-2)^{\alpha+2}-1}{\alpha+2}\right) &\leq F \\
&\Rightarrow \\
F &\geq n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-2)^{\alpha+2}-1}{\alpha+2}\right) \\
&\geq n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-2)\cdot(n-2)^{\alpha+1}-1}{\alpha+2}\right) \\
&\geq n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-1)\cdot(n-2)^{\alpha+1}}{\alpha+2} - \tfrac{(n-2)^{\alpha+1}}{\alpha+2} - \tfrac{1}{\alpha+2}\right) \\
&\geq n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-1)\cdot(n-2)^{\alpha+1}}{\alpha+2} - \tfrac{(n-2)^{\alpha+1}+1}{\alpha+2}\right) \\
&\overset{\text{(d)}}{\geq} n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-1)\cdot(n-2)^{\alpha+1}}{\alpha+2} - \tfrac{(n-1)}{\alpha+2}\right) \\
&\geq n \cdot \left(\tfrac{(n-2)^{\alpha+1}-1}{\alpha+1}\right) - \left((n-1)\cdot\tfrac{(n-2)^{\alpha+1}-1}{\alpha+2}\right) \\
&\geq \left((n-2)^{\alpha+1}-1\right)\cdot\left(\tfrac{n}{\alpha+1} - \tfrac{(n-1)}{\alpha+2}\right) \\
&\geq \left((n-2)^{\alpha+1}-1\right)\cdot\left(\tfrac{n\cdot(\alpha+2)-(n-1)\cdot(\alpha+1)}{(\alpha+1)\cdot(\alpha+2)}\right) \\
&\geq \left((n-2)^{\alpha+1}-1\right)\cdot\left(\tfrac{n\cdot\alpha+2\cdot n-n\cdot\alpha-n+\alpha+1}{(\alpha+1)\cdot(\alpha+2)}\right) \\
&\geq \left((n-2)^{\alpha+1}-1\right)\cdot\left(\tfrac{n+\alpha+1}{(\alpha+1)\cdot(\alpha+2)}\right) \\
&\Rightarrow \\
\tfrac{F}{\binom{n}{2}} &\geq \tfrac{\left((n-2)^{\alpha+1}-1\right)\cdot\left(\frac{n+\alpha+1}{(\alpha+1)\cdot(\alpha+2)}\right)}{\binom{n}{2}} \\
&\geq \tfrac{2\cdot\left((n-2)^{\alpha+1}-1\right)\cdot\left(\frac{n+\alpha+1}{(\alpha+1)\cdot(\alpha+2)}\right)}{n\cdot(n-1)} \\
&\geq \left((n-2)^{\alpha+1}-1\right)\cdot\tfrac{2}{(\alpha+1)\cdot(\alpha+2)}\cdot\tfrac{n+\alpha+1}{n\cdot(n-1)} \\
&\geq \left((n-2)^{\alpha} - \tfrac{1}{(n-2)}\right)\cdot\tfrac{2}{(\alpha+1)\cdot(\alpha+2)}\cdot\tfrac{(n+\alpha+1)\cdot(n-2)}{n\cdot(n-1)} \\
&\Rightarrow \\
\theta^d &\geq \left((n-2)^{\alpha} - \tfrac{1}{(n-2)}\right)\cdot\underbrace{\frac{2}{(\alpha+1)\cdot(\alpha+2)}}_{\beta}\cdot\underbrace{\frac{(n+\alpha+1)\cdot(n-2)}{n\cdot(n-1)}}_{\gamma} \\
&\Rightarrow \\
\theta &\geq \left((n-2)^{\alpha} - \tfrac{1}{(n-2)}\right)^{\frac{1}{\alpha}}\cdot\beta^{\frac{1}{\alpha}}\cdot\gamma^{\frac{1}{\alpha}} \\
&\overset{\text{(b)}}{\geq} \left((n-2) - \tfrac{1}{\alpha}\cdot\tfrac{1}{(n-2)}\cdot(n-2)^{1-\alpha}\right)\cdot\beta^{\frac{1}{\alpha}}\cdot\gamma^{\frac{1}{\alpha}} \\
&\geq \left((n-2) - \tfrac{1}{\alpha\cdot(n-2)^{\alpha}}\right)\cdot\beta^{\frac{1}{\alpha}}\cdot\gamma^{\frac{1}{\alpha}} \\
&\overset{\text{(e)}}{\geq} \left((n-2) - \tfrac{1}{\alpha\cdot(n-2)^{\alpha}}\right)\cdot\beta^{\frac{1}{\alpha}} \\
&\geq \left((n-2)-1\right)\cdot\beta^{\frac{1}{\alpha}} \\
&\geq \left((n-3)\right)\cdot\beta^{\frac{1}{\alpha}} \\
&\overset{\text{(f)}}{\geq} \cdot(n-3)^{\frac{2}{9}}
\end{aligned}
$$

Por outro lado temos que:

$$
\begin{aligned}
F \quad & \leq n \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \left(\tfrac{(n-1)^{\alpha+2}-1}{\alpha+2}\right) \\
& \leq n \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \left((n-1) \cdot \tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{\alpha+2}\right) \\
& \Rightarrow \\
\tfrac{F}{\binom{n}{2}} \quad & \leq \frac{\left(n \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \left((n-1) \cdot \tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{\alpha+2}\right)\right)}{\binom{n}{2}} \\
& \leq \frac{2 \cdot n \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right)}{n \cdot (n-1)} - \frac{2 \cdot (n-1) \cdot \left(\tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{\alpha+2}\right)}{n \cdot (n-1)} \\
& \leq \frac{2 \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right)}{(n-1)} - \frac{2 \cdot \left(\tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{\alpha+2}\right)}{n} \\
& \leq \tfrac{2}{(n-1)} \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{\alpha+1}\right) - \tfrac{2}{n} \cdot \left(\tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{\alpha+2}\right) \\
& \leq \tfrac{2}{\alpha+1} \cdot \left(\tfrac{(n-1)^{\alpha+1}-1}{(n-1)}\right) - \tfrac{2}{\alpha+2} \cdot \left(\tfrac{(n-1)^{\alpha+1}-\frac{1}{(n-1)}}{n}\right) \\
& \leq \tfrac{2}{\alpha+1} \cdot \tfrac{(n-1)^{\alpha+1}}{(n-1)} - \tfrac{2}{\alpha+1} \cdot \tfrac{1}{(n-1)} - \tfrac{2}{\alpha+2} \cdot \tfrac{(n-1)^{\alpha+1}}{n} + \tfrac{2}{\alpha+2} \cdot \tfrac{1}{n \cdot (n-1)} \\
& \leq \tfrac{2}{\alpha+1} \cdot (n-1)^{\alpha} - \tfrac{2}{(\alpha+1) \cdot (n-1)} - \tfrac{2 \cdot (n-1) \cdot (n-1)^{\alpha}}{(\alpha+2) \cdot n} + \tfrac{2}{(\alpha+2) \cdot n \cdot (n-1)} \\
& \leq 2 \cdot (n-1)^{\alpha} \cdot \left(\tfrac{1}{\alpha+1} - \tfrac{(n-1)}{(\alpha+2) \cdot n}\right) + \tfrac{2}{(n-1)} \cdot \left(\tfrac{1}{(\alpha+2) \cdot n} - \tfrac{1}{(n-1)}\right) \\
& \leq 2 \cdot (n-1)^{\alpha} \cdot \left(\tfrac{(\alpha+2) \cdot n - (\alpha+1) \cdot (n-1)}{(\alpha+1) \cdot (\alpha+2) \cdot n}\right) + \tfrac{2}{(n-1)} \cdot \left(\tfrac{1}{(\alpha+2) \cdot n} - \tfrac{1}{(n-1)}\right) \\
& \leq 2 \cdot (n-1)^{\alpha} \cdot \left(\tfrac{\alpha \cdot n + 2 \cdot n - \alpha \cdot n + \alpha - n + 1}{(\alpha+1) \cdot (\alpha+2) \cdot n}\right) + \tfrac{2}{(n-1)} \cdot \left(\tfrac{1}{(\alpha+2) \cdot n} - \tfrac{1}{(n-1)}\right) \\
& \leq 2 \cdot (n-1)^{\alpha} \cdot \left(\tfrac{n+\alpha+1}{(\alpha+1) \cdot (\alpha+2) \cdot n}\right) + \tfrac{2}{(n-1)} \cdot \left(\tfrac{1}{(\alpha+2) \cdot n} - \tfrac{1}{(n-1)}\right) \\
& \leq (n-1)^{\alpha} \cdot \tfrac{2}{(\alpha+2) \cdot (\alpha+1)} \cdot \left(\tfrac{n+\alpha+1}{n}\right) + \tfrac{2}{(n-1)} \cdot \left(\tfrac{1}{(\alpha+2) \cdot n} - \tfrac{1}{(n-1)}\right) \\
& \Rightarrow \\
\theta^{\alpha} \quad & \leq (n-1)^{\alpha} \cdot \tfrac{2}{(\alpha+2) \cdot (\alpha+1)} \cdot \left(\tfrac{n+\alpha+1}{n}\right) + o(\tfrac{1}{n^2}) \\
& \Rightarrow \\
\theta \quad & \leq (n-1) \cdot \tfrac{2}{(\alpha+2) \cdot (\alpha+1)}^{\frac{1}{\alpha}} \cdot \left(\tfrac{n+\alpha+1}{n}\right)^{\frac{1}{\alpha}} + o(\tfrac{1}{n^2}) \\
& \leq (n-1) \cdot \tfrac{2}{(\alpha+2) \cdot (\alpha+1)}^{\frac{1}{\alpha}} \cdot \left(1 + \tfrac{\alpha+1}{n}\right)^{\frac{1}{\alpha}} + o(\tfrac{1}{n^2}) \\
& \overset{\text{(g)}}{\leq} (n-1) \cdot \tfrac{2}{(\alpha+2) \cdot (\alpha+1)}^{\frac{1}{\alpha}} \cdot 1.03 + o(\tfrac{1}{n^2}) \\
& \overset{\text{(h)}}{\leq} (n-1) \cdot 0.34 \cdot 1.03 + o(\tfrac{1}{n^2}) \\
& \leq (n-1) \cdot 0.35 + o(\tfrac{1}{n^2}) \\
& \leq (n-1) \cdot \tfrac{100}{285} + o(\tfrac{1}{n^2})
\end{aligned}
$$

Dessa forma, temos que:

$$(n-3) \cdot \frac{2}{9} \leq \theta \leq (n-1) \cdot \frac{100}{285} + o(\frac{1}{n^2})$$

Infelizmente esses limitantes para o valor de θ não favorecem as contas para estimar a esperança de Hamming. Sendo assim, o trabalho de avaliar essa estatística será deixado para o módulo de testes.

Observações 5.3.1 :

(a) :

$$\begin{aligned}\binom{1/\alpha}{m} &= \frac{\frac{1}{\alpha}!}{m!\cdot(\frac{1}{\alpha}-m)!}\\ &= \frac{\frac{1}{\alpha}\cdot(\frac{1}{\alpha}-1)-(\frac{1}{\alpha}-m+1)}{m!}\\ &= (-1)^{m-1}\cdot\binom{m-1-1/\alpha}{m}\end{aligned}$$

(b) :

$$\begin{aligned}(x^\alpha-b)^{\frac{1}{\alpha}} &= x\cdot(1-\tfrac{b}{x^\alpha})^{\frac{1}{\alpha}}\\ &= x\cdot\left(1-\binom{1/\alpha}{1}\cdot\left(\tfrac{b}{x^\alpha}\right)+\binom{1/\alpha}{2}\cdot\left(\tfrac{b}{x^\alpha}\right)^2-\binom{1/\alpha}{3}\cdot\left(\tfrac{b}{x^\alpha}\right)^3\cdots\right.\\ &\overset{(a)}{=} x\cdot\left(1-\sum_{m\geq 1}\binom{m-1-1/\alpha}{m}\cdot\left(\tfrac{b}{x^\alpha}\right)^m\right)\\ &\geq x\cdot(1-\tfrac{1}{\alpha}\cdot\tfrac{b}{x^\alpha})\\ &\geq x-\tfrac{1}{\alpha}\cdot b\cdot x^{1-\alpha}\end{aligned}$$

$$\begin{aligned}(x^\alpha+b)^{\frac{1}{\alpha}} &= x\cdot(1+\tfrac{b}{x^\alpha})^{\frac{1}{\alpha}}\\ &= x\cdot\left(1+\binom{1/\alpha}{1}\cdot\left(\tfrac{b}{x^\alpha}\right)+\binom{1/\alpha}{2}\cdot\left(\tfrac{b}{x^\alpha}\right)^2+\binom{1/\alpha}{3}\cdot\left(\tfrac{b}{x^\alpha}\right)^3\cdots\right.\\ &\overset{(a)}{=} x\cdot\left(1-\sum_{m\geq 1}(-1)^{m-1}\cdot\binom{m-1-1/\alpha}{m}\cdot\left(\tfrac{b}{x^\alpha}\right)^m\right)\\ &\leq x\cdot(1+\tfrac{1}{\alpha}\cdot\tfrac{b}{x^\alpha})\\ &\leq x+\tfrac{1}{\alpha}\cdot b\cdot x^{1-\alpha}\end{aligned}$$

(c) : para $b>0$ e $\alpha>0$,

$$x-\frac{1}{\alpha}\cdot b\cdot x^{1-\alpha}\overset{(b)}{\leq}(x^\alpha-b)^{\frac{1}{\alpha}}\leq x$$

(d) : para $\alpha>0$,

$$(n-2)^{\alpha+1}+1\geq(n-1)$$

Observações 5.3.2 :

ⓔ :

$$\begin{aligned}
\gamma &= \tfrac{(n+\alpha+1)\cdot(n-2)}{n\cdot(n-1)} \\
&= \tfrac{n^2-2\cdot n+n\cdot\alpha-2\cdot\alpha+n-2}{n\cdot(n-1)} \\
&= \tfrac{n^2+(\alpha-1)\cdot n-2\cdot(\alpha+1)}{n\cdot(n-1)} \\
&= \tfrac{n^2+(\alpha-1)\cdot n-2\cdot(\alpha+1)}{n^2-n} \\
&= \tfrac{n^2\cdot\left(1+\frac{(\alpha-1)}{n}-\frac{2\cdot(\alpha+1)}{n^2}\right)}{n^2\cdot(1-\frac{1}{n})} \\
&= \tfrac{1+\frac{(\alpha-1)}{n}-\frac{2\cdot(\alpha+1)}{n^2}}{1-\frac{1}{n}} \\
&= \tfrac{1+\frac{\alpha}{n}-\frac{1}{n}-\frac{2\cdot(\alpha+1)}{n^2}}{1-\frac{1}{n}} \\
&= \tfrac{\left(1-\frac{1}{n}\right)+\frac{\alpha}{n}-\frac{2\cdot(\alpha+1)}{n^2}}{1-\frac{1}{n}} \\
&= 1+\tfrac{\frac{\alpha}{n}-\frac{2\cdot(\alpha+1)}{n^2}}{1-\frac{1}{n}} \\
&\leq 1+\tfrac{\frac{\alpha}{n}}{1-\frac{1}{n}} \\
&\leq 1+\tfrac{\alpha}{n} \\
&\Rightarrow \\
\gamma^{\frac{1}{\alpha}} &\leq \left(1+\tfrac{\alpha}{n}\right)^{\frac{1}{\alpha}} \\
&\leq \left(1^{\alpha}+\tfrac{\alpha}{n}\right)^{\frac{1}{\alpha}} \\
&\overset{\text{ⓑ}}{\leq} 1+\tfrac{1}{\alpha}\cdot\tfrac{\alpha}{n}\cdot 1 \\
&\leq 1+\tfrac{1}{n} \\
&\Rightarrow \\
1\leq\gamma^{\frac{1}{\alpha}} &\leq 1+\tfrac{1}{n}
\end{aligned}$$

ⓕ para α pequeno:

$$\begin{aligned}
\left(\tfrac{2}{(\alpha+1)\cdot(\alpha+2)}\right)^{\frac{1}{\alpha}} &= \left(\tfrac{2}{(\alpha+1)\cdot 2\cdot(\frac{\alpha}{2}+1)}\right)^{\frac{1}{\alpha}} \\
&= \left(\tfrac{1}{(\alpha+1)\cdot(\frac{\alpha}{2}+1)}\right)^{\frac{1}{\alpha}} \\
&= \left(\tfrac{1}{\alpha+1}\right)^{\frac{1}{\alpha}}\cdot\left(\tfrac{1}{\frac{\alpha}{2}+1}\right)^{\frac{1}{\alpha}} \\
&\overset{x=\frac{1}{\alpha}}{=} \left(\tfrac{1}{\frac{1}{x}+1}\right)^{x}\cdot\left(\tfrac{1}{\frac{1}{2\cdot x}+1}\right)^{x} \\
&\Rightarrow \\
\lim_{\alpha\to 0}\left(\tfrac{2}{(\alpha+1)\cdot(\alpha+2)}\right)^{\frac{1}{\alpha}} &= \lim_{x\to\infty}\left(\left(\tfrac{1}{\frac{1}{x}+1}\right)^{x}\cdot\left(\tfrac{1}{\frac{1}{2\cdot x}+1}\right)^{x}\right) \\
&= \lim_{x\to\infty}\left(\left(\tfrac{1}{1+\frac{1}{x}}\right)^{x}\cdot\left(\tfrac{1}{1+\frac{1}{2\cdot x}}\right)^{x}\right) \\
&= \tfrac{1}{e}\cdot\tfrac{1}{\sqrt{e}} \\
&= e^{-\frac{3}{2}} \\
&\sim \tfrac{2}{9}
\end{aligned}$$

Observações 5.3.3 :

(g) : Sabemos que a função $\left(1+\frac{\alpha+1}{n}\right)^{\frac{1}{\alpha}}$ é decrescente tanto em n como em α, derivada negativa. Assim, quando utilizamos valores de $\alpha \geq 0.1$ e valores de $n \geq 225$ temos que:

$$\left(1+\frac{\alpha+1}{n}\right)^{\frac{1}{\alpha}} \leq \left(1+\frac{0.1+1}{255}\right)^{\frac{1}{0.1}} \leq 1.03$$

(h) : Sabemos que a função $\frac{2}{(\alpha+2)\cdot(\alpha+1)}^{\frac{1}{\alpha}}$ é crescente no domínio $\alpha \geq 0$, derivada positiva. Assim, quando utilizamos $0.1 \leq \alpha \leq 1$ temos que:

$$\frac{2}{(\alpha+2)\cdot(\alpha+1)}^{\frac{1}{\alpha}} \leq \frac{2}{(1+2)\cdot(1+1)}^{\frac{1}{1}} \leq 0.34$$

Capítulo 6

Módulo para testes

Este capítulo trata do código escrito na linguagem R responsável por aplicar os processos de *limiarização pelo método controle* e *limiarização pelo método ajustado*. Este código será chamado aqui de módulo por sua estrutura modular que facilita sua modificação e reutilização. Além de aplicar os processos citados, este módulo ainda calcula algumas estatísticas com o objetivo de investigar os efeitos de cada processo.

Na tentativa de facilitar o entendimento e a localização de partes específicas, o código foi dividido em quatro partes, são elas:

- *principal*
- *estágio 1*
- *estágio 2*
- *estágio 3*

Cada uma das partes será descrita em uma das seções a seguir.

6.1 Principal

A parte principal, main, é o ponto de entrada do programa. Ela é a responsável por orquestrar as chamadas de cada estágio e apresentar os resultados ao final do processamento.

Essa parte encontra-se no arquivo main.r e tem em seu preâmbulo a definição de algumas variáveis globais responsáveis por configurar o modo como módulo irá se comportar. Dentre essas variáveis encontram-se as variáveis responsáveis por controlar a ordem dos grafos limiarizados (| V(G) |), os valores de proporções de corte utilizadas no processo, quais espaços métricos serão utilizados na limiarização ajustada e quais estatísticas serão calculadas.

Com estas informações o módulo executa a seguinte rotina:

- Chama rotinas do estágio 1 para criar uma Matriz A.
- Chama rotinas do estágio 2 para limiarizar a matriz A pelo processo *limiarização pelo método controle*. Uma chamada é feita para cada um dos valores de proporção de corte informados no preâmbulo.
- Chama rotinas do estágio 3 para calcular as estatísticas informadas no preâmbulo.
- Chama rotinas do estágio 1 para cria a matriz H de cada um dos espaços métricos informados no preâmbulo.

- Chama rotinas do estágio 2 para limiarizar a matriz A pelo processo *limiarização pelo método ajustado*. Uma chamada é feita para cada um dos valores de proporção de corte informados no preâmbulo.
- Chama rotinas do estágio 3 para calcular as estatísticas informadas no preâmbulo.

Esse processo se repete um número de vezes igual ao número contido na variável *tries* também definida no preâmbulo.

Ao final dessa rotina, os resultados das estatísticas são exibidos ao usuário.

6.2 Estágio 1

O primeiro estágio se responsabiliza por gerar as matrizes A e Raw. A matriz A é a matriz de entrada a ser limiarizada. As matrizes Raw são matrizes de adjacência de grafos das famílias em estudo, estas matrizes são utilizadas para gerar a matriz H de cada caso.

Essa parte encontra-se no arquivo stage1.r e se sustenta principalmente nas seguintes funções:

- *make_raw_data*
- *make_star_adj*
- *make_path_adj*
- *make_cycle_adj*
- *make_board_adj*
- *make_torBoard_adj*

Todas as funções acima recebem as dimensões da matriz a ser gerada e uma função que retorne uma lista de valores para as entradas da matriz. No caso da matriz A, *make_raw_data*, a função passada foi a *runif* configurada para sortear valores em $]0, 1[$. Para todos os outros casos os valores eram iguais a 1.

6.3 Estágio 2

O segundo estágio é responsável pelas funções necessárias para a limiarização da matriz A.

Essa parte encontra-se no arquivo stage2.r e se sustenta principalmente nas seguintes funções:

- *limiarize_ctl*
- *limiarize_adj*
- *adjust*
- *perturbation*
- *limiarization*

As funções acima realizam todo o processo descrito na formulação teórica, capítulo 4. As operações com matrizes foram cuidadosamente projetadas para utilizar as otimizações de cálculos com vetores próprios da linguagem.

A função *limiarization* tem um gatilho que dispara um aviso caso algum valor de π_{ij} calculado exceda o valor máximo 1. Isso pode ocorrer durante a aplicação da função de ajuste quando o valor de p é bastante elevado e a quantidade de entradas que tem seu limiar aumentado é pequena quando comparada com a quantidade de entradas que reduzem o limiar.

Uma forma de calcular o p máximo que se pode utilizar no processo ajustado é a seguinte:

$$\begin{aligned} \pi_{ij} &\leq 1 \\ &\Rightarrow \\ p \cdot k \cdot f(\eta_{ij}) &\leq 1 \\ &\Rightarrow \\ p &\leq \tfrac{1}{k \cdot f(d)} \\ &\Rightarrow \\ p &\leq \tfrac{F}{\binom{n}{2} \cdot f(d)} \\ &\Rightarrow \\ p\_\text{max} &= \tfrac{F}{\binom{n}{2} \cdot f(D)} \end{aligned}$$

6.4 Estágio 3

O terceiro estágio se responsabiliza por calcular as estatísticas dos grafos gerados pelos processos.

Essa parte encontra-se no arquivo stage3.r. Como utilizou-se o pacote igraph para calcular a maior parte das estatísticas estudadas nesse trabalho, essa parte contém apenas duas funções:

- *clustering_coeficient*
- *hamming_dist*

O cálculo dessas estatísticas foi feito como descrito no apêndice B.

Capítulo 7

Resultados

Os resultados que serão apresentados neste capítulo foram obtidos da execução do módulo de testes configurado com os seguintes parâmetros:

- $tries = 100$
- $n1 = 15$
- $n2 = 15$
- $n = n1 * n2 = 225$
- $\alpha = 0.2$
- $p\_list = \{0.25,\ 0.5,\ 0.75\}$

Todas as estatísticas apresentadas aqui são como descritas no apêndice B e foram calculadas como descrito no capítulo 6.

Os valores apresentados nas tabelas 7.1, 7.2 e 7.3 têm a forma *média* $\pm$ *desvio_padrão*. Com exceção da estatística *distância de Hamming*, todos os valores das tabelas mencionadas correspondem a relação entre o resultado obtido no processo ajustado e o resultado obtido no processo controle.

As tabelas foram criadas de modo que se o valor para uma estatística for 1 é porque o processo controle e o processo ajustado apresentam o mesmo resultado para aquela estatística.

Pode-se observar das tabelas que o grau médio se manteve, assim como requisitado pela *propriedade* 2. Também é possível ver que os valores da *distância de Hamming* foram expressivos mas dos outros parâmetros não.

$p = 0,25$	**Estrela**	**Caminho**	**Ciclo**	**Tabuleiro**	**Tabuleiro Toroidal**
Grau_médio	1,0000 ± 0,0005	0,9999 ± 0,0050	0,9997 ± 0,0044	0,9995 ± 0,0039	0,9996 ± 0,0036
Hamming	1.0022 ± 0.0005	1.1514 ± 0.0046	1.1300 ± 0.0050	1.0930 ± 0.0037	1.0754 ± 0.0037
Fechamento	1,0000 ± 0,0008	1,0000 ± 0,0007	0,9999 ± 0,0006	0,9999 ± 0,0005	0,9999 ± 0,0005
Centralidade	0,9998 ± 0,0079	0,9921 ± 0,0157	1,0004 ± 0,0146	0,9953 ± 0,0147	1,0010 ± 0,0135
Transitividade	1,0002 ± 0,0007	0,9988 ± 0,0062	0,9970 ± 0,0053	1,0008 ± 0,0050	0,9992 ± 0,0044
Clustering	1,0002 ± 0,0007	0,9993 ± 0,0062	0,9969 ± 0,0054	1,0009 ± 0,0051	0,9992 ± 0,0045
Modularidade	1,0063 ± 0,0429	1,0063 ± 0,0479	1,0056 ± 0,0429	1,0089 ± 0,0475	1,0132 ± 0,0393
Tamanho_cluster	0,9978 ± 0,0369	0,9988 ± 0,0443	0,9983 ± 0,0450	0,9966 ± 0,0455	0,9972 ± 0,0431

TABELA 7.1: Estatísticas para p=0,25

$p = 0,5$	**Estrela**	**Caminho**	**Ciclo**	**Tabuleiro**	**Tabuleiro Toroidal**
Grau_médio	1,0000 ± 0,0004	0,9997 ± 0,0030	1,0004 ± 0,0028	1,0002 ± 0,0026	0,9999 ± 0,0024
Hamming	1.0022 ± 0.0004	1.1512 ± 0.0035	1.1303 ± 0.0032	1.0929 ± 0.0025	1.0750 ± 0.0022
Fechamento	1,0000 ± 0,0001	1,0000 ± 0,0010	1,0001 ± 0,0009	1,0001 ± 0,0008	0,9999 ± 0,0008
Centralidade	0,9986 ± 0,0052	0,9862 ± 0,0110	1,0007 ± 0,0097	0,9891 ± 0,0104	1,0000 ± 0,0091
Transitividade	1,0002 ± 0,0004	0,9988 ± 0,0033	0,9978 ± 0,0029	1,0012 ± 0,0027	0,9994 ± 0,0025
Clustering	1,0002 ± 0,0004	0,9993 ± 0,0033	0,9978 ± 0,0029	1,0013 ± 0,0028	0,9994 ± 0,0025
Modularidade	1,0024 ± 0,0524	0,9880 ± 0,0512	0,9999 ± 0,0510	0,9947 ± 0,0535	0,9981 ± 0,0542
Tamanho_cluster	0,9903 ± 0,0586	1,0148 ± 0,0615	1,0002 ± 0,0599	1,0115 ± 0,0557	1,0083 ± 0,0663

TABELA 7.2: Estatísticas para p=0,5

$p = 0,75$	**Estrela**	**Caminho**	**Ciclo**	**Tabuleiro**	**Tabuleiro Toroidal**
Grau_médio	0,9999 ± 0,0003	1,0000 ± 0,0024	0,9999 ± 0,0025	1,0001 ± 0,0018	1,0000 ± 0,0017
Hamming	1.0022 ± 0.0003	1.1517 ± 0.0027	1.1307 ± 0.0024	1.0928 ± 0.0021	1.0752 ± 0.0020
Fechamento	1,0000 ± 0,0001	1,0005 ± 0,0014	0,9999 ± 0,0015	1,0004 ± 0,0011	1,0000 ± 0,0010
Centralidade	0,9976 ± 0,0036	0,9794 ± 0,0098	1,0034 ± 0,0090	0,9824 ± 0,0103	1,0018 ± 0,0086
Transitividade	1,0001 ± 0,0003	0,9990 ± 0,0025	0,9972 ± 0,0025	1,0011 ± 0,0018	0,9995 ± 0,0017
Clustering	1,0001 ± 0,0003	0,9995 ± 0,0025	0,9972 ± 0,0025	1,0013 ± 0,0018	0,9995 ± 0,0017
Modularidade	0,9992 ± 0,0394	0,9605 ± 0,0502	0,9789 ± 0,0537	0,9813 ± 0,0505	0,9995 ± 0,0492
Tamanho_cluster	0,9999 ± 0,0444	1,0019 ± 0,0501	0,9940 ± 0,0530	0,9977 ± 0,0471	1,0012 ± 0,0513

TABELA 7.3: Estatísticas para p=0,75

Capítulo 8

Conclusão

Os resultados obtidos pelos testes realizados e descritos no capítulo 7 mostram que os grafos gerados pelo processo de *limiarização pelo método controle* e os grafos gerados pelo processo de *limiarização pelo método ajustado* diferiram em cerca de 10% das arestas. Por outro lado os resultados das estatísticas observadas não apontaram mudança significativa quando comparado os dois métodos.

Ao observar esses resultados, algumas questões surgem naturalmente. Considerando que os grafos gerados eram diferentes mas as estatísticas não indicaram isso, é provável que as questões sejam:

1. será que os espaços escolhidos interferiram no resultado?
2. será que o processo proposto tem alguma serventia?
3. será que escolhemos corretamente a forma de aferir as diferenças?

Sobre a primeira pergunta, deve-se dizer que de fato é possível que os resultados tenham sido influenciados pelas matrizes H utilizadas. Por outro lado, é pouco provável que isso tenha realmente acontecido uma vez que as famílias de grafos escolhidas apresentam características bastante distintas umas das outras.

Com relação a segunda pergunta, considerando as diferença encontradas nos grafos gerados, os 10% já mencionados, pode-se dizer que o processo ajustado apresentou um refinamento na forma de aplicar a limiarização ao adicionar informações ao processo. Desta forma a modelagem pelo método proposto parece ser mais fiel ao problema inicial.

Finalmente, sobre a última pergunta, os resultados obtidos mostram que as estatísticas selecionadas são robustas e não se alteram diante das modificações observadas. Então, talvez seja necessário pensar outras estatísticas mais sensíveis quando esse tipo de alteração no grafo modelado for relevante para o problema.

Apêndice A

Grafos conhecidos

Neste apêndice definiremos formalmente um grafo e apresentaremos algumas famílias utilizadas nesse trabalho.

Definição A.1. Enunciaremos duas definições equivalentes para grafos:

(a) Um *grafo* G é um conjunto não vazio V de elementos chamados vértices juntamente com um conjunto possivelmente vazio E de subconjuntos de tamanho 2 de V chamados arestas. (Chartrand, Lesniak e Zhang, 2010)

(b) Um *grafo* $G = (V, E)$ é um objeto abstrato formado por um conjunto V de vértices e um conjunto E de arestas conectando pares de vértices. (Brandes, 2009)

Denotaremos por $V(G)$ o conjunto V de um grafo G. E denotaremos por $E(G)$ o conjunto E de um grafo G.

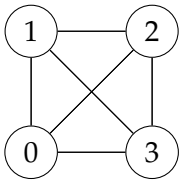
Para descrever as mencionadas famílias, precisamos de mais algumas definições.

Definição A.2. O *grau* de um vértice $v \in V(G)$, denotado por $\deg(v)$ é o número de arestas em $E(G)$ que contém v. (Brandes, 2009)

Definição A.3. Um grafo é dito *completo* se todo subconjunto de vértices $v \in V(G)$ de tamanho 2 está em $E(G)$. Ou seja, $E(G) = \{\{u, v\} | \forall u, v \in V(G)\}$. Denotamos um grafo completo por $K_{|V|}$. (Chartrand, Lesniak e Zhang, 2010)

I.e.:

K_4:



Definição A.4. Um grafo é dito *regular* se todo vértice $v \in V(G)$ tem o mesmo grau. Um grafo regular cujo grau dos vértices é igual a k é chamado de *k-regular*. (Chartrand, Lesniak e Zhang, 2010)

Definição A.5. Um grafo G é chamado *bipartido* se $V(G)$ pode ser particionado em dois conjuntos U e W tal que toda aresta de G contenha um vértice de U e um vértice de W. (Chartrand, Lesniak e Zhang, 2010)

Definição A.6. Um grafo G é chamado *bipartido completo* se $V(G)$ pode ser particionado em dois conjuntos U e W tal que $\{u, w\}$ é uma aresta de G se e somente se $u \in U$ e $w \in W$. Se $|U| = s$ e $|W| = t$, então o grafo completo bipartido tem s+t arestas e é denotado por $K_{s,t}$. (Chartrand, Lesniak e Zhang, 2010)

Definição A.7. O *produto cartesiano* de dois grafos $G = (V, E)$ e $G' = (V', E')$ denotado por $G \times G'$ é um grafo $H = (V_H, E_H)$ tal que:

1. $V_H = \{\ (s,t) \mid s \in V\ \wedge\ t \in V'\ \}$
2. $E_H = \{\ ((u,v),(s,t))\ \mid \begin{cases} u = s & \wedge (v,t) \in E' \\ \text{ou} & \\ (u,s) \in E & \wedge v = t \end{cases} \}$

A.1 Estrela (*Star*)

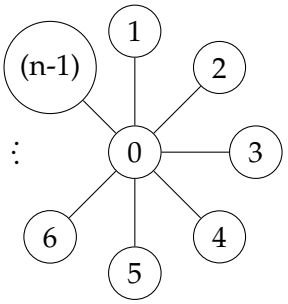
Um grafo bipartido completo $K_{1,t}$ é chamado de *estrela*. (Chartrand, Lesniak e Zhang, 2010)

A família de grafos S_n é o conjunto de todos os grafos bipartidos completos $K_{1,t}$ para $\forall t \in \mathbb{N}$. Um representante dessa família tem exatamente um vértice de grau $n-1$ e $n-1$ vértices de grau 1. Deste modo, todo grafo $G \in S_n$ tem seu conjunto $E(G)$ da seguinte forma:

* Seja $G(V, E)$ um grafo com $V = \{0, 1, 2, \cdots, n-1\}$. Escolha $v \in V$ arbitrariamente. Então,

$$G \in S_n \Leftrightarrow E = \{\{v, u\}, \forall u \neq v \in V\}$$

Para os nosso exemplos, o vértice central de uma estrela será sempre o vértice 0. Segue um exemplo de um grafo $G \in S_n$ com $|V| = n$:



A.2 Caminho (*Path*)

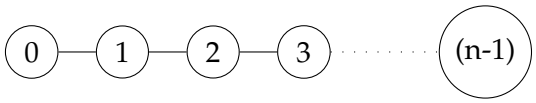
Para um inteiro $n \geq 1$, um caminho P_n é um grafo G tq $|V(G)| = n$ e com $|E(G)| = n-1$, cujos vértices possam ser rotulados por $0, 1, 2, \cdots, n-1$, e as arestas são $\{i, i+1\}$ para $i = 0, 1, 2, \cdots, n-2$. (Chartrand, Lesniak e Zhang, 2010)

Em um caminho, 2 vértices tem grau exatamente 1 e os outros $n-2$ vértices tem grau exatamente 2

* Seja $G(V, E)$ um grafo com $V = \{0, 1, 2, \cdots, n-1\}$. Então,

$$G \in P_n \Leftrightarrow E(G) = \{\{i, i+1\}, \text{para } i = 0, 1, 2, \cdots, n-2\}$$

Segue um exemplo de um grafo $G \in P_n$ com $|V| = n$:



A.3 Ciclo (*Cycle*)

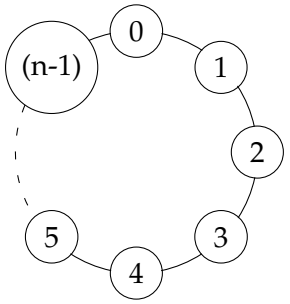
Para um inteiro $n \geq 3$, o ciclo C_n é um grafo tal que $|V(G)| = n$ e $|E(G)| = n$, e cujos vértices possam ser rotulado por $0, 1, 2, \cdots, n-1$ e cujas arestas são $\{0, n-1\}$ e $\{i, i+1\}$ para $i = 0, 1, 2, \cdots, n-2$. (Chartrand, Lesniak e Zhang, 2010)

Todo ciclo é um grafo 2-regular, ou seja, todos os vértices tem grau 2.

* Seja $G(V, E)$ um grafo com $V = \{0, 1, 2, \cdots, n-1\}$. Então,

$$G \in C_n \Leftrightarrow E(G) = \{\{i, i+1\}, \text{para } i = 0, 1, 2, \cdots, n-2\} \cup \{\{0, n-1\}\}$$

Segue um exemplo de um grafo $G \in C_n$ com $|V| = n$:



A.4 Tabuleiro (*Board*)

Para um inteiro $n = s \cdot t$, o tabuleiro $Q_{s,t}$ é um grafo tal que $|V(G)| = s \cdot t$ e $|E(G)| = 2 \cdot s \cdot t - s - t$. Em um tabuleiro os vértices podem ser rotulados por $0, 1, 2, \cdots, (s \cdot t - 1)$ e nesse caso, as arestas são $E(G) = \{\{i, i+1\}, \text{ se } i \neq s-1 \bmod s\} \cup \{\{i, i+s\} \text{ se } i < (t-1) \cdot s\}$.

Todo vértice em um tabuleiro tem grau 2 se for um dos quatro cantos; 3 se estiver em uma das quatro bordas e não for canto; ou 4 se estiver no meio do tabuleiro.

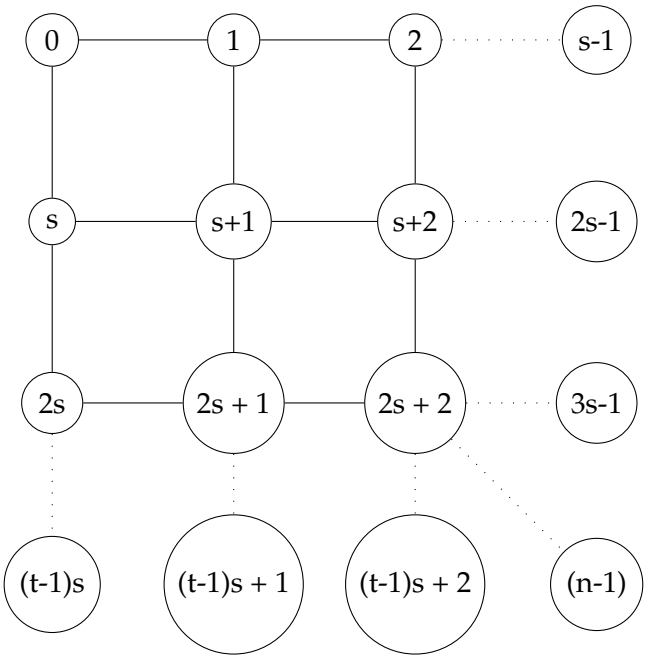
O nome tabuleiro vem do fato de que os vértices desse grafo podem ser arranjados em um tabuleiro de xadrez de forma que as arestas sejam somente entre casas adjacentes.

* Seja n o inteiro dado por $n = s \cdot t$. Seja $G(V, E)$ um grafo com $V = \{0, 1, 2, \cdots, n-1$. Então,

$$G \in Q_n \Leftrightarrow E(G) = \begin{array}{l} \{\{i, i+1\} \mid \forall i \in V(G) \text{ t.q. } i \neq s-1 \bmod s\} \\ \cup \{\{i, i+s\} \mid i < (t-1) \cdot s\} \end{array}$$

Observe que um tabuleiro é o produto cartesiano entre dois caminhos, ou seja $Q_{n1 \cdot n2} = P_{n1} \times P_{n2}$

Segue um exemplo de um grafo $G \in Q_n$ com $|V| = n = s \cdot t$:



A.5 Tabuleiro Toroidal (*Toroidal Board*)

Para um inteiro $n = s \cdot t$, o tabuleiro toroidal $T_{s,t}$ é um grafo tal que $|V(G)| = s \cdot t$ e $|E(G)| = 2 \cdot s \cdot t$. Em um tabuleiro toroidal os vértices podem ser rotulados por $0, 1, 2, \cdots, (s \cdot t - 1)$ e nesse caso, as arestas são $E(G) = \{\{i, (i+1) \text{mod } s\}\} \cup \{\{i, (i+s) \text{mod } n\}\}$.

Todo vértice em um tabuleiro toroidal tem grau exatamente 4.

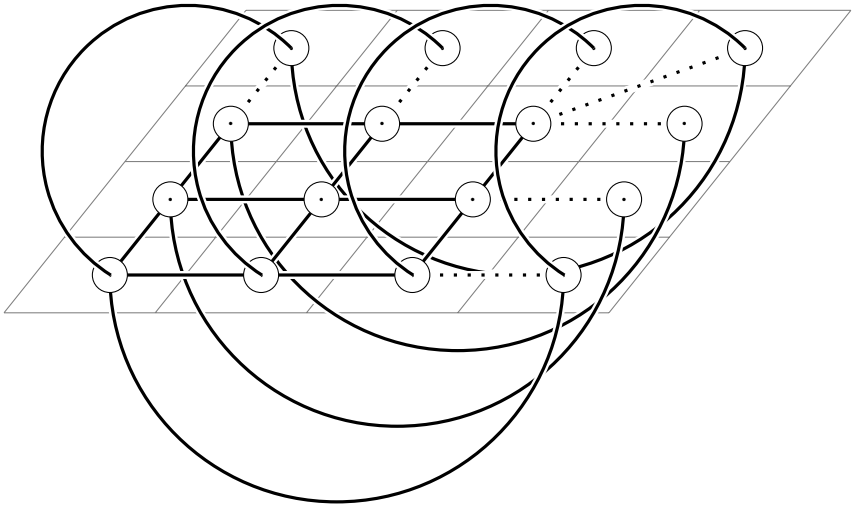
O nome tabuleiro toroidal vem do fato desses grafos serem uma expansão dos grafos tabuleiros adicionando arestas para formar um toro.

* Seja n o inteiro dado por $n = s \cdot t$. Seja $G(V, E)$ um grafo com $V = \{0, 1, 2, \cdots, n-1$. Então,

$$G \in T_n \Leftrightarrow E(G) = \begin{array}{l} \{\{i, (i+1) \text{mod } s\}\} \\ \cup \{\{i, (i+s) \text{mod } n\}\} \end{array}$$

Observe que um tabuleiro toroidal é o produto cartesiano entre dois ciclos, ou seja $T_{n1 \cdot n2} = C_{n1} \times C_{n2}$

Segue um exemplo de um grafo $G \in T_n$ com $|V| = n = s \cdot t$:



Apêndice B

Estatísticas dos grafos

Para comparar os resultados do processo de *limiarização pelo método ajustado* com os resultados do processo de *limiarização pelo método controle*, foi selecionado um conjunto de estatísticas descritas a seguir:

B.1 Grau médio (*Degree Average*)

O grau médio representa o número de arestas das quais, em média, um vértice participa(incide). Esse índice é calculado pela seguinte fórmula:

$$\frac{1}{n} \cdot \sum_{v \in V(G)} \deg(v)$$

B.2 Índice de Hamming (*Hamming Index*)

A distância de Hamming entre dois grafos com o mesmo conjunto de vértices é o tamanho da diferênça simétrica dos dois conjuntos de arestas. Convenientemente, pode-se computar esse valor utilizando as matrizes de adjacência. Para esse trabalho, a distância de Hamming entre dois grafos é a distância de Hamming entre as matrizes de adjacência de cada grafo divida por dois. Ou seja, a distância de Hamming entre um grafo G_1 e um outro grafo G_2 é o número de entradas diferentes entre a matriz de adjacência $A(G_1)$ e a matriz de adjacência $A(G_2)$ dividido por dois. I.e. se

$$A(G_1) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array} \text{ e } A(G_2) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

então a distância de Hamming entre G_1 e G_2, denotada por dist_Hamm$_{G_1,G_2}$, tem valor igual a 2, pois quatro das entradas de $A(G_1)$ são diferentes das entradas de $A(G_2)$. A saber as entradas diferentes são $a_{1,3}$, $a_{2,3}$, $a_{3,1}$ e $a_{3,2}$.

A distância de Hamming mede o número de arestas destoantes entre os grafos. Para calcular esse índice utilizou-se a seguinte fórmula:

$$\sum_{0 \leq i < j \leq n} |b_{ctl_{ij}} - b_{adj_{ij}}|$$

Para poder comparar as distâncias entre grafos gerados pelo processo de limiarização utilizando diferentes parâmetros, foram feitas duas normalizações:

- Divisão pelo número de arestas esperados na normalização de um grafo gerado na *limiarização pelo método controle* de uma matriz com entradas *i.i.d.* em $]0, 1[$. Esse valor normalizado é chamado de distância de Hamming relativa e pode ser expressa pela formula a seguir.

$$\frac{\sum\limits_{0 \leq i < j \leq n} |b_{ctl_{ij}} - b_{adj_{ij}}|}{p \cdot \binom{n}{2}}$$

- Deslocamento de uma unidade no valor da distância de Hamming relativa. Isso se faz para dar chance de comparar essa estatística com as demais uma vez que todas as outras estatísticas têm valor 1 se o grafo é igual ao controle enquanto ela tem valor 0. Esse valor deslocado pode é chamado de índice de Hamming e pode ser expresso pela formula a seguir.

$$1 + \frac{\sum\limits_{0 \leq i < j \leq n} |b_{ctl_{ij}} - b_{adj_{ij}}|}{p \cdot \binom{n}{2}}$$

B.3 Fechamento (*Closeness*)

O fechamento é um índice estrutural que reflete a centralidade do vértice por meio das distâncias dele para os demais vértices. (Brandes, 2009)

Para o grafo todo a medida é calculada da seguinte forma:

$$\frac{1}{n} \cdot \sum_{v \in V} \frac{1}{\sum\limits_{i \neq v} d(v, i)}$$

B.4 Centralidade (*Coreness*)

Centralidade é uma medida de densidade local. Para obtê-la, o *k-core* máximo que cada vértice participa é calculado. A média entre os valores k de cada vértice representa a centralidade do grafo. Um *k-core* é um subgrafo maximal no qual todos os vértices tem grau ao menos k. (Igraph, 2015)

A fórmula para calcular esse parâmetro é:

$$\frac{1}{n} \cdot \sum_{v \in V} \text{k-core}(v)$$

B.5 Transitividade (*Transitivity*)

Transitividade mede a probabilidade de vértices adjacentes a um terceiro estarem conectados entre si, ou seja, é a probabilidade de ter-se um triangulo quando sorteado um vértice e dois vértices adjacentes a ele. (Brandes, 2009)

A fórmula para calcular esse parâmetro é:

$$\frac{3 \cdot \alpha(G)}{\tau(G)} = \frac{3 \cdot \frac{1}{3} \sum\limits_{v \in V} \alpha(v)}{\sum\limits_{v \in V} \binom{\deg(v)}{2}} = \frac{\sum\limits_{v \in V} |\{\triangle \mid v \in V_{\triangle}\}|}{\sum\limits_{v \in V} \binom{\deg(v)}{2}}$$

B.6 Agrupamento (*Clustering*)

Clustering também mede a probabilidade de vértices adjacentes a um terceiro estarem conectados assim como a transitividade. A diferença é que no caso do clustering temos o valor normalizado pelo número de vértices que podem ter triângulos (com grau maior ou igual a 2). (Brandes, 2009)

A fórmula para calcular esse parâmetro é:

$$\frac{1}{|V'|} \cdot \sum_{v \in V'} \frac{\alpha(v)}{\binom{\deg(v)}{2}} \text{, onde } V' = \{v \in V \mid deg(v) \geq 2\}$$

B.7 Modularidade do cluster (*Modularity*)

Após executar um algoritmo para identificar grupos de vértices, *clusters*, o cálculo da modularidade mede quão boa foi a divisão atingida.(Igraph, 2015)

Utilizamos o algoritmo *fastandgreed* que identifica as comunidades via otimização direta da modularidade. (Igraph, 2015)

A fórmula para calcular a modularidade é:

$$\frac{1}{2 \cdot |A(G)|} \cdot \sum_{0 \leq i, < j \leq n} \left(a_{ij} - \frac{\deg(i) \cdot \deg(j)}{2 \cdot |A(G)|} \right) \cdot \delta(c_i, c_j)$$

onde c_i representa o cluster ao qual o vértice i pertence

e $\delta(s,t) = \begin{cases} 1 & \text{se } s = t \\ 0 & \text{c.c.} \end{cases}$

B.8 Variação dos tamanhos das comunidades

A variação dos tamanhos dos cluster é uma medida próxima do desvio padrão dividido pela soma dos tamanhos. Basicamente utiliza-se a raiz quadrada da soma dos quadrados normalizados pela soma dos valores.

A fórmula para calcular a modularidade é:

$$\frac{1}{\sum\limits_{c \in C} \text{size}(c)} \cdot \sqrt{\sum_{c \in C} \text{size}(c)^2}$$

onde C é o conjunto dos *clusters* calculados pelo algoritmo *fastandgreed* como no caso da estatística modularidade.

Bibliografia

Bertucci, J. L. O. (2013). *Metodologia básica para elaboração de trabalhos de conclusão de curso (TCC): ênfase na elaboração de TCC de pós-graduação Lato Sensu*. Edição 1, reimpressão 5. Editora Atlas S.A. ISBN: 978-85-224-5080-0.

Brandes, Ulrik (2009). *Network Analysis: Methodological Foundations*. Springer. ISBN: 978-81-848-9364-9.

Chartrand, Gary, Linda Lesniak e Ping Zhang (2010). *Graphs & Digraphs, Fifth Edition (Textbooks in Mathematics)*. Edição 5. Chapman e Hall/CRC. ISBN: 978-14-398-2627-0.

Garcia-Ramos, C. et al. (2016). «Graph theory and cognition: A complementary avenue for examining neuropsychological status in epilepsy». Em: *Epilepsy & Behavior - article in press* XXX.XXX, p. XXX. URL: http://dx.doi.org/10.1016/j.yebeh.2016.02.032.

Igraph (2015). *The network analysis package*. URL: http://igraph.org.

Lima, E. L. (1975). *Espaços Métricos*. Edição ? Editora ? ISBN: ?

R-project (2016). *The R Project for Statistical Computing*. URL: https://www.r-project.org.

Takahashi, D. Y. et al. (2012). «Discriminating Different Classes of Biological Networks by Analyzing the Graphs Spectra Distribution». Em: *PLoS ONE* 7.12, e49949. URL: http://journals.plos.org/plosone/article/asset?id=10.1371%2Fjournal.pone.0049949.PDF.

Van Wijk, Bernadette CM, Cornelis J Stam e Andreas Daffertshofer (2010). «Comparing brain networks of different size and connectivity density using graph theory». Em: *PloS one* 5.10, e13701.

Zhou, D., W. K. Thompson e G. Siegle (2009). «MATLAB toolbox for functional connectivity». Em: *NeuroImage* 47.4, 1590–1607. URL: http://dx.doi.org/10.1016/j.neuroimage.2009.05.089.