UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA
GRADUAÇÃO EM FÍSICA

BRUNO RODRIGUES DUARTE

ANÁLISE DE MODULARIDADE USANDO MINIMUM SPANNING TREE EM DADOS DE DOENÇAS MATERNO-INFANTIL

FORTALEZA

2018

BRUNO RODRIGUES DUARTE

ANÁLISE DE MODULARIDADE USANDO MINIMUM SPANNING TREE EM DADOS DE DOENÇAS MATERNO-INFANTIL

Monografia de Bacharelado apresentada à Coordenação da Graduação do Curso de Física, da Universidade Federal do Ceará, como requisito parcial para a obtenção do Título de Bacharel em Física.

Orientador: Prof. Dr. Carlos Lenz César.

FORTALEZA
2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Curso de Física

D87a Duarte, Bruno Rodrigues.

Análise de Modularidade usando Minimum Spanning Tree em dados de doenças Materno-Infantil / Bruno Rodrigues Duarte. – Fortaleza, 2018.

38.:il.

Monografia (bacharelado) - Universidade Federal do Ceará, Centro de Ciências, Departamento de Física, Fortaleza, 2018.

.

Orientação: Prof. Dr. Carlos Lenz César.

1. Matriz de correlação. 2. Árvore de expansão mínima. 3. Teoria dos grafos. 4. Rede de dados. I. Título.

CDD 530

BRUNO RODRIGUES DUARTE

ANÁLISE DE MODULARIDADE USANDO MINIMUM SPANNING TREE EM DADOS DE DOENÇAS MATERNO-INFANTIL

Monografia de Bacharelado apresentada à Coordenação da Graduação do Curso de Física, da Universidade Federal do Ceará, como requisito parcial para a obtenção do Título de Bacharel em Física.

Aprovada em 21/06/2018.

BANCA EXAMINADORA

Prof. Dr. Carlos Lenz César (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Humberto de Andrade Carmona
Universidade Federal do Ceará (UFC)

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

A Deus que me dá forças,
a minha família que está sempre comigo
e a quem acreditou e se devotou por mim..

AGRADECIMENTOS

Primeiramente agradeço a Deus que chegou a mim nesta árdua jornada com todo o seu infinito amor e caridade. A Nossa Senhora por ser esta Mãe perfeita que por todos intercede e a Santo Agostinho por tamanho exemplo que é para mim.

Agradeço a minha família inteira, amigos, colegas e a quem orou, se preocupou e fez sinceros sacrifícios por mim.

Também a todas as pessoas que contribuíram, direta ou indiretamente, para a conclusão deste trabalho, como o Prof. Carlos Lenz, os colegas do Laboratório de Sistemas Complexos e os outros membros do departamento de Física pelos ensinamentos e apoio em geral.

RESUMO

Este trabalho acadêmico faz um estudo sobre um método de análise de dados estatísticos. Verifica-se quais são suas limitações e se é possível corrigi-las com o mínimo de alterações nos resultados. Inicialmente são passados os conceitos matemáticos necessários para os procedimentos de análise. Desde definições estatísticas como Coeficiente de Correlação até Teoria de Grafos para estudo de redes e análise de dados. Adiante, explica-se o método e observar-se os resultados buscando possíveis formas de contornar os problemas que surgem durante a análise. As fontes dos dados serão os registros hospitalares de óbitos ocorridos na infância e na maternidade no estado do Ceará.

Palavras-chave: Matriz de correlação. Árvore de expansão mínima. Teoria dos grafos. Rede de dados.

ABSTRACT

This academic work makes a study on a method of statistical analysis. It's verified its limitations and if it's possible to correct them with the minimum of changes in the results. Initially, the necessary mathematical concepts for the analysis procedures are passed. Statistical definitions such as Coefficient of Correlation and Graph Theory for network study and data analysis for example. Later, the method is explained and the results are observed, looking for possible ways to overcome the problems that arise during the analysis. The source of the data will be the hospital records of deaths occurred in childhood and maternity in the brazilian state of Ceará.

Keywords: Correlation matrix. Minimum spanning tree. Graph theory. Data network.

LISTA DE TABELAS

Tabela 1 – Matriz de adjacência 19
Tabela 2 – Amostragem 26
Tabela 3 – Capítulos e suas doenças abrangidas[1] 27
Tabela 4 – “Filtragem”dos dados de duas doenças 30

LISTA DE FIGURAS

Figura 1 – Grafo dos jogos do Campeonato de Vôlei. 18
Figura 2 – Exemplo de caminho em um grafo[2] 20
Figura 3 – Passos do algoritmo de Prim[3] 22
Figura 4 – Exemplo de matriz de distância 23
Figura 5 – MST do grafo dado 24
Figura 6 – Matriz de distância reorganizada 25
Figura 7 – Matriz de correlação 28
Figura 8 – MST do grafo usando Pearson 28
Figura 9 – Correlação (Pearson) e distância rearranjado por MST 29
Figura 10 –Correlação com $N = N_{xy} + 1$ 33
Figura 11 –MST do grafo usando $N = N_{xy} + 1$ 33
Figura 12 –Correlação rearranjado por MST usando $N = N_{xy} + 1$ 34
Figura 13 –Correlação com $N = N_{xy} + 20\%$ 34
Figura 14 –MST do grafo usando $N = N_{xy} + 20\%$ 35
Figura 15 –Correlação e distância rearranjado por MST usando $N = N_{xy} + 20\%$ 36
Figura 16 –Correlação e distância com $N = 2093$ 36
Figura 17 –MST do grafo usando $N = 2093$ 37
Figura 18 –Correlação e distância rearranjado por MST usando $N = 2093$ 37

SUMÁRIO

Introdução 11
1 CONCEITOS ESTATíSTICOS E TEORIA DOS GRAFOS 12
1.1 Momentos, covariância e correlação 12
1.2 Espaço métrico e distância 16
1.3 Teoria dos Grafos 17
1.3.1 Introdução 17
1.3.2 Definição 18
1.3.3 Incidência e adjacência 18
1.3.4 Grau 19
1.3.5 Caminho 19
1.3.6 Árvores 20
1.3.7 MST 20
1.3.8 Algoritmo de Prim 21
2 REDE DE DADOS 23
3 RESULTADOS 26
3.1 Correlação para dados binários 29
3.2 Correlação para diferentes valores de N 32
3.2.1 $N_{xy} + 1$ 32
3.2.2 $N_{xy} + 20\%$ 34
3.2.3 $N_{xy} = 2093$ 36
4 CONCLUSÃO 38
REFERÊNCIAS 39

INTRODUÇÃO

O objetivo deste trabalho é investigar um método de análise de dados que utiliza Teoria dos Grafos. Clustering (ou Agrupamento) é o nome da técnica computacional usada que tem como objetivo separar dados a serem analisados em grupos baseando-se em alguma característica inerente. A técnica de agrupamento está normalmente associada com a análise exploratória, pois envolve problemas em que há pouca informação a priori acerca dos dados, e poucas hipóteses podem ser sustentadas. Por outro lado, é justamente esta ferramenta de agrupamento que, quando aplicada, pode fornecer novas hipóteses a respeito de possíveis inter-relacionamentos dos dados e de sua estrutura intrínseca. Possíveis aplicações estão no reconhecimento de padrões em geral como:

- Informações de vários pacientes com sintomas semelhantes
- Classificação de documentos
- Acessos similares para determinados usuários de internet

O processo de agrupamento envolve as seguintes etapas:

- Representação dos dados
- Definição de uma medida de similaridade
- Agrupamento
- Apresentação do resultado

Bons clusters dependem de alguns fatores como a similaridade entre seus grupos (e os objetos pertencentes ao grupo) e os métodos de cálculo e aplicação dessas similaridades. A similaridade costuma ser expressa como uma medida de distância entre as variáveis. A forma que esta distância é definida depende do tipo de variável (no nosso caso, será definida a partir da correlação estatística).

1 CONCEITOS ESTATÍSTICOS E TEORIA DOS GRAFOS

Neste capítulo, iremos mostrar o conhecimento matemático necessário para o desenvolvimento deste trabalho. Começaremos pelas definições de momento, covariância e coeficiente de correlação. Depois iremos definir uma distância a partir deste coeficiente. Posteriormente, abordaremos a teoria dos grafos e seus conceitos mais utilizados em análise de dados.

1.1 Momentos, covariância e correlação

Em estatística, as distribuições codificam a informação sobre o comportamento estocástico de um conjunto de variáveis aleatórias $\{V_1, V_2, ..., V_n\} \in \mathbb{R}^n$. Há duas formas de se representar estas distribuições: A função distribuição e a função densidade. Formalmente, a função distribuição $F(V_1, V_2, ..., V_n)$ de um conjunto de variáveis aleatórias é definida como

$$F(V_1, V_2, ..., V_n) = P\{V_1 \leq v_1, V_2 \leq v_2, ..., V_n \leq v_n\} \tag{1.1}$$

onde $V_1 \leq v_1, V_2 \leq v_2, ..., V_n \leq v_n$ são os eventos e $F(V_1, V_2, ..., V_n)$ mede a probabilidade de saída de cada evento. Algumas propriedades da função de distribuição são:

- $F(-\infty, V_2, ..., V_n) = F(V_1, -\infty, ..., V_n) = ... = F(V_1, V_2, ..., -\infty) = 0$;
- $F(+\infty, +\infty, ..., +\infty) = 1$;
- $P\{v_A \leq V_1 \leq v_B, V_2 \leq v_2, ..., V_n \leq v_n\} = F(v_B, V_2, ..., V_n) - F(v_A, V_2, ..., V_n)$

Com a função distribuição, a função densidade pode ser definida como

$$f(V_1, V_2, ..., V_n) = \frac{\partial^n F(V_1, V_2, ..., V_n)}{\partial V_1 \partial V_2 ... \partial V_n} \tag{1.2}$$

$f(V_1, V_2, ..., V_n)$ é uma probabilidade se e somente se satisfazer

$$f(V_1, V_2, ..., V_n) \geq 0$$

e

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(V_1, V_2, ..., V_n) dV_1 dV_2 ... dV_n = 1$$

Estamos interessados em criar medidas usando as funções de distribuição e o conjunto de variáveis aleatórias. Estas medidas são obtidas através do valor esperado de funções

escalares destas variáveis aleatórias. Dessa forma, se $Y = g(V_1, V_2, ..., V_n)$ é uma função do conjunto $\{V_1, V_2, ..., V_n\}$, o valor esperado é definido como

$$\langle Y \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(V_1, V_2, ..., V_n) f(V_1, V_2, ..., V_n) dV_1 dV_2 ... dV_n \tag{1.3}$$

onde $f(V_1, V_2, ..., V_n)$ é a função densidade. Desta expressão, observa-se que o valor esperado é linear:

$$\langle aX + bY \rangle = a\langle X \rangle + b\langle Y \rangle$$

e que $\langle c \rangle = c$ onde c é uma constante

Estamos interessados em duas funções escalares especiais que permitirão definir os momentos centrados e não-centrados. Especificamente, o momento não-centrado é dado por

$$M_{k_1 k_2 ... k_n} = \langle V_1^{k_1} V_2^{k_2} ... V_n^{k_n} \rangle \tag{1.4}$$

onde k_n indica o n-ésimo momento de cada variável V_n. O primeiro-momento é simplesmente o valor esperado, comumente denotado por μ

$$\begin{aligned} M_{10...0} &= \langle V_1 \rangle = \mu_1, \\ M_{01...0} &= \langle V_2 \rangle = \mu_2, \\ &\vdots \\ M_{00...1} &= \langle V_n \rangle = \mu_n. \end{aligned} \tag{1.5}$$

Por outro lado, o momento centrado é definido como

$$m_{k_1 k_2 ... k_n} = \langle (V_1 - \mu_{V_1})^{k_1} (V_2 - \mu_{V_2})^{k_2} ... (V_n - \mu_{V_n})^{k_n} \rangle. \tag{1.6}$$

O segundo momento centrado possibilita medir o desvio dos valores em torno do valor esperado de cava variável aleatória. Essa quantidade chama-se variância e é denotada por σ_n^2 a qual é definida como

$$m_{00...2} = \sigma_n^2 = \langle (V_n - \mu_n)^2 \rangle = \langle V_n^2 \rangle - \mu_n^2. \tag{1.7}$$

de onde tiramos o desvio padrão $\sigma_n = \sqrt{\sigma_n^2}$.

É importante mencionar que em conjuntos de dados reais é impossível obter o valor esperado como uma integral sobre um número infinito de possibilidades. Em vez do valor esperado real nós medimos o que se chama estimador, que é a média dos momentos sobre uma

amostragem finita da distribuição. Um bom estimador $\hat{q}$ é tal que $\langle \hat{q} \rangle = \langle q \rangle$. O estimador do valor esperado é denotado por

$$\bar{q} = \frac{1}{n} \sum_i q_i \tag{1.8}$$

onde q_i são os valores observados na amostra. O estimador da variância é dado por

$$s_q^2 = \frac{1}{n-1} \sum_i (q_i - \bar{q})^2. \tag{1.9}$$

Podemos definir uma variável normalizada a partir de q como:

$$\tilde{q} = \frac{q - \langle q \rangle}{\sigma_q} \tag{1.10}$$

É fácil provar que

$$\langle \tilde{q} \rangle = 0$$
$$\langle \tilde{q}^2 \rangle = 1$$

Os mesmos resultados permanecem para os estimadores. Podemos dizer que a variável normalizada funciona como um vetor unitário.

É importante observar que o primeiro momento centrado de duas variáveis aleatórias nos fornece uma quantidade estatística útil que mede o grau dependência linear entre estas duas variáveis. Esta quantidade chamada covariância é definida como

$$\begin{aligned} m_{01\ldots1} = cov(V_i, V_j) &= \langle (V_i - \mu_{V_i})(V_j - \mu_{V_j}) \rangle \\ &= \langle V_i V_j \rangle - \langle V_i \rangle \langle V_j \rangle \end{aligned} \tag{1.11}$$

A covariância pode ser positiva, negativa ou igual a zero. Vejamos o que significa cada uma dessas situações

- **Dependência direta ou indireta**: Quando a covariância assume valores positivos, indica que o cescimento (ou decréscimo) de uma das variáveis é acompanhado pelo crescimento (ou decréscimo) da outra. Isso mostra que existe uma relação direta entre elas. Ser negativa indica o contrário: que o cescimento de uma das variáveis é acompanhado pelo decréscimo da outra. Ou seja, há uma relação inversa.

- **Independência**: É o caso da covariância igual a zero. Não há nenhuma relação entre variáveis nesse caso.

Algumas propriedades da covariância:

- $cov(V_i, V_j) = cov(V_j, V_i)$
- $cov(V_i + V_j, V_k) = cov(V_i, V_k) + cov(V_j, V_k)$
- $cov(aV_i, bV_j) = abcov(V_i, V_j)$
- $cov(V_i, k) = 0$ quando k é constante

Definimos então o estimador da covariância:

$$cov(q_i, q_j) = \frac{1}{n}\sum_i (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \tag{1.12}$$

Embora a covariância meça o grau de dependência linear entre duas variáveis aleatória, existe uma pequena dificuldade na sua definição relacionada com suas dimensões. Observe que a covariância possui as mesmas dimensões que a variável aleatória. Dessa forma, o mais apropriado é que essa medida seja adimensional. Uma medida adimensional que permite conhecer o grau de dependência linear entre duas variáveis aleatória é o coeficiente de correlação. Matematicamente, ele é definido como

$$\rho(i, j) = \frac{cov(i, j)}{\sigma(i)\sigma(j)} \tag{1.13}$$

onde i e j são as duas variáveis aleatórias e $\sigma(i)$ e $\sigma(j)$ são seus respectivos desvios padrões. As propriedades do coeficiente de correlação são as mesmas da covariância, porém, este possui os valores restritos a $[-1, 1]$.

Podemos dar uma interpretação geométrica ao coeficiente de correlação. Se considerarmos i e j como vetores, o coeficiente de correlação nos dá o cosseno do ângulo entre eles. Existe uma matriz de representação do coeficiente de correlação que é bastante usada quando se trabalha com vários conjuntos de variáveis aleatórias. Por exemplo, dado um conjunto $\mathbf{V}_m = (V_{1m}V_{2m}, ..., V_{nm})$ onde n representa a quantidade de variáveis aleatórias e m é o número de vezes que cada variável aparece. A matriz de correlação deste conjunto é dada por

$$\rho_{nn'} = Diag\{\mathbf{\Gamma}_{nn'}(m)\}^{-1/2}\mathbf{\Gamma}_{nn'}(m)Diag\{\mathbf{\Gamma}_{nn'}(m)\}^{-1/2} \tag{1.14}$$

onde $\mathbf{\Gamma}_{nn'}(m) = \langle(\mathbb{V}_m - \mu)(\mathbb{V} - \mu)'\rangle$ é a matriz de covariância e $Diag\{\mathbf{\Gamma}_{nn'}(m)\}$ são os elementos da diagonal da matriz de covariâcia. É importante notar que a matriz de correlação é simétrica.

1.2 Espaço métrico e distância

Na matemática, um espaço métrico é um par (M, d) onde $M = \{x_1, x_2, ..., x_j\}$ é um conjunto de elementos e $d(x_i, x_j)$ é uma função que associa a cada par de elementos um número real positivo que representa a distância entre eles. Se o espaço é um espaço métrico, $d(x_i, x_j)$ deve satisfazer os seguintes axiomas

1. A distância é simétrica: $d(x_i, x_j) = d(x_j, x_i)$
2. Se $d(x_i, x_j) = 0$, então $x_i = x_j$
3. A distância deve satisfazer a desigualdade triangular

É possível construir uma medida de distância a partir do coeficiente de correlação. Sendo i e j duas variáveis aleatórias de forma que:

$$\begin{aligned} \mu_i &= \frac{1}{n}\sum_n i_n, \qquad \sigma_i^2 = \frac{1}{n}\sum_n (i_n - \mu_i)^2, \\ \mu_j &= \frac{1}{n}\sum_n j_n, \qquad \sigma_j^2 = \frac{1}{n}\sum_n (j_n - \mu_j)^2, \end{aligned} \tag{1.15}$$

sejam os estimadores do valor médio e da variância de cada variável aleatória. O estimador da covariância entre elas é

$$cov(i, j) = \frac{1}{n}\sum_n (i_n - \mu_i)(j - \mu_j) \tag{1.16}$$

Agora iremos definir duas variáveis aleatórias normalizadas

$$x_n = \frac{i_n - \mu_i}{\sqrt{n}\sigma_i}, \qquad y_n = \frac{j_n - \mu_j}{\sqrt{n}\sigma_j} \tag{1.17}$$

É fácil demonstrar que

$$\sum_n x_n = \sum_n y_n = 0, \qquad \sum_n x_n^2 = \sum_n y_n^2 = 1$$

Dessa forma, podemos tratar x_n e y_n como vetores unitários. O quadrado da distância euclideana entre x_n e y_n é então

$$d^2(x_n, y_n) = \sum_n (x_n - y_n)^2 = 2(1 - \sum_n x_n y_n) \tag{1.18}$$

Por outro lado, de acordo a equação 1.17

$$\sum_n x_n y_n = \sum_n \left(\frac{i_n - \mu_i}{\sqrt{n}\sigma_i}\right)\left(\frac{j_n - \mu_j}{\sqrt{n}\sigma_j}\right) = \frac{1}{n\sigma_i\sigma_j}\sum_n (i_n - \mu_i)(j_n - \mu_j) = \frac{cov(i,j)}{\sigma_i\sigma_j} \quad (1.19)$$

portanto, a distância é dada por

$$d(x_n, y_n) = \sqrt{2(1 - \rho(i,j))} \quad (1.20)$$

que é conhecida como distância de correlação. Ela está de acordo com os três axiomas da distância do espaço métrico[4]. As propriedades da distância de correlação são as mesmas do coeficiente de correlação com uma pequena alteração. $d(i,j) = 2$ significa dependência indireta e $d(i,j) = 0$ significa dependência direta.

Assim como o coeficiente de correlação, também há uma matriz de distância de correlação.

1.3 Teoria dos Grafos

1.3.1 Introdução

Começamos com um simples exemplo[5]. Em uma escola, os professores decidiram realizar um torneio de vôlei entre os alunos. Eles foram separados em times formados por alunos de cada sala de um mesmo ano letivo. Até o momento, se inscreveram as turmas 6A, 6B, 7A, 7B, 8A e 8B. Eis os jogos que já foram realizados:

6A jogou com 7A, 7B e 8B
6B jogou com 7A, 8A e 8B
7A jogou com 6A e 6B
7B jogou com 6A, 8A e 8B
8A jogou com 6B, 7B e 8B
8B jogou com 6A, 6B, 7B e 8A

Iremos desenhar uma figura que representará os jogos realizados. As turmas serão representadas por círculos e os jogos serão representados por linhas.

Esta figura é o que chamamos de **grafo**. Grafos podem ser usados para modelar muitos tipos de relações e processos em física, biologia, sistemas de informações entre outras coisas. Possuem uma grande aplicabilidade especialmente em ciências da computação.

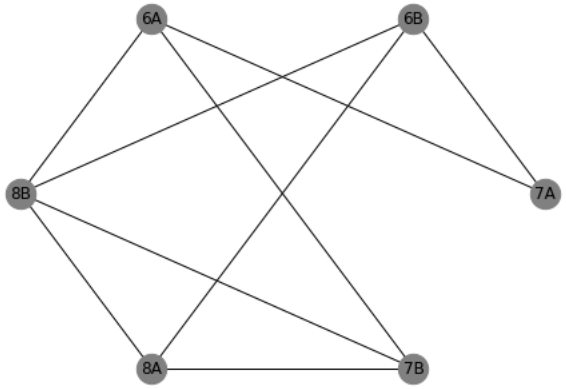

Figura 1 – Grafo dos jogos do Campeonato de Vôlei.

1.3.2 Definição

Um grafo G consiste em um conjunto finito e não-vazio $V(G)$ de objetos chamados vértices, juntamente com um conjunto $E(G)$ de pares não-ordenados de vértices; os elementos de $E(G)$ são chamados de arestas. Podemos representá-lo por $G = (V; E)$, onde $V = V(G)$ e $E = E(G)$. O número de vértices é chamado de **ordem** do grafo, enquanto a quantidade total de pares ordenados pertencentes a $E(G)$ é dita como o **tamanho** do grafo[6].

Um **grafo direcionado** (ou **dígrafo**) é aquele em que suas arestas possuem um direcionamento. Ou seja, dados dois vértices A e B ligados por uma aresta, devemos informar de qual vértice a arestas está "saindo"e qual ela está "chegando". Nesse caso, a relação não é simétrica. As suas arestas chamam-se **arcos**.

Um **grafo não direcionado** (ou simplemente **grafo**) consiste em um grafo que não há direcionamento em suas arestas. Ou seja, dados dois vértices A e B ligados por uma aresta $(A; B)$, temos que $(A; B) = (B; A)$.

Para nossos propósitos, iremos trabalhar apenas com grafos não direcionados. Sempre que se falar em grafo, ficará implícito que será não direcionado.

Podemos criar uma função $w : E \longrightarrow \mathbb{R}$ que associa a cada aresta um valor real. Muitas vezes refere-se à esta função como sendo o *peso* da aresta. O que este peso representa irá depender do tipo de problema abordado.

1.3.3 Incidência e adjacência

Se uma aresta conecta dois vértices, então esses dois vértices são ditos incidentes à aresta. Usando o grafo acima como exemplo temos: 6A é incidente a 7A, 7B e 8B, mas não é incidente a 6B ou 8A; 6B é incidente a 7A, 8A e 8B, mas não a 6A nem a 7B.

Dois vértices são considerados adjacentes se uma aresta existe entre eles. No grafo do exemplo, os vértices 6A e 7A são adjacentes, mas os vértices 6B e 7B não são. O conjunto

de vizinhos de um vértice consiste de todos os vértices adjacentes a ele. No grafo-exemplo, o vértice 7A possui 2 vizinhos: vértice 6A e vértice 6B.

Um grafo finito direcionado de ordem N pode ser representado por um quadro que é chamado de matriz de adjacência: uma matriz $N \times N$ cujo valor na linha i e coluna j fornece o número de arestas do i-ésimo ao j-ésimo vértices. Eis a matriz de adjacência do exemplo dado:

Times	6A	6B	7A	7B	8A	8B
6A	0	0	1	1	0	1
6B	0	0	0	0	1	1
7A	1	0	0	0	0	0
7B	1	0	0	0	1	1
8A	0	1	0	1	0	1
8B	1	1	0	1	1	0

$$\rightarrow \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Tabela 1: Matriz de adjacência

O grafo do exemplo dado é não-valorado, ou seja, as suas arestas não possuem uma função peso associada. Nessas condições, o elemento a_{ij} da matriz de adjacência assume apenas os valores 0 e 1 (indicando ausência de aresta ou presença respectivamente).

1.3.4 Grau

O grau de um vértice é o número de arestas incidentes a ele. Se houver laços (aresta que liga um vértice a ele mesmo), serão contabilizados duas vezes. No grafo do exemplo acima o vértice 7A possui grau igual a 2, os vértices 6A, 6B, 7B e 8A têm grau igual 3 e o vértice 8B tem grau igual a 4. Se E é finito, então grau total dos vértices é o dobro do número de arestas. Em um dígrafo, distingue-se o grau de saída (o número de arestas saindo de um vértice) e o grau de entrada (o número de arestas entrando em um vértice). O grau de um vértice em um dígrafo é igual à soma dos graus de saída e de entrada. O grau de um vértice é definido somente quando o número de arestas incidentes sobre o vértice é finito.

1.3.5 Caminho

Caminho é uma sequência de vértices conectados por uma sequência de arestas. Um caminho é chamado simples se nenhum dos vértices no caminho se repete. O comprimento do caminho é o número de arestas que o caminho usa, contando-se arestas múltiplas vezes. O peso de um caminho num grafo valorado é a soma dos pesos das arestas atravessadas. Dois caminhos são independentes se não tiverem nenhum vértice em comum, exceto o primeiro e o último.

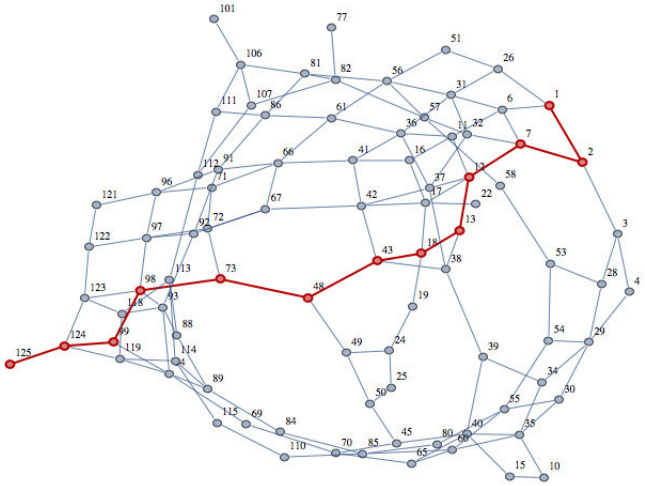

Figura 2 – Exemplo de caminho em um grafo[2]

1.3.6 Árvores

Há vários tipos de grafos. Estamos interessados em um tipo específico chamado *árvore*. Antes de irmos para a definição, devemos checar dois tipos específicos de grafos.

- **Grafo simples** é um grafo não direcionado, sem laços e existe no máximo uma aresta entre quaisquer dois vértices (sem arestas paralelas). Para um grafo simples, o número de aresta incidentes a um vértice é igual ao seu grau.

- **Grafo conexo** é aquele em que é possível estabelecer um caminho de qualquer vértice para qualquer outro vértice de um grafo. Se for sempre possível estabelecer um caminho de qualquer vértice para qualquer outro vértice mesmo depois de remover $k-1$ vértices, então diz-se que o grafo está k-conexo. Em um grafo genérico G, o corte associado a um conjunto X de vértices é o conjunto de todas as arestas que têm uma ponta em X e outra em $V(G) - X$, onde $V(G)$ é o conjunto de todos os vértices pertencentes ao grafo G.

Dessa forma podemos dizer de maneira bem suscinta que uma árvore é um grafo simples acíclico e conexo.

1.3.7 MST

Do inglês *Minimum Spanning Tree*, a Árvore de Extensão Mínima é um subgrafo que possui aplicações no processo de otimização de dados. Trata-se de um grafo onde se preserva a ordem original do grafo (mantém-se todos os vértices), mas a soma de todos os pesos individuais é a mínima possível.

Podemos formular um problema que nos mostraria sua aplicabilidade. Imagine que existe um bairro da cidade de Fortaleza em que todas as suas vias são de mão dupla (ou seja, as arestas não são orientadas). A prefeitura encarregou que uma determinada equipe de

profissionais fizesse a coleta de lixo neste bairro. Eles dispõe de um único caminhão para este trabalho. Supomos que cada domicílio pode ser representado por um vértice no grafo e que o peso de cada aresta seria a distância entre cada domicílio. Nesse caso, encontrar a MST nos forneceria um modo direto de "varrer"todas as casa gastando o mínimo de combustível.

Agora que definimos uma MST, observamos duas de suas propriedades.

- Em uma MST, há apenas um único caminho entre qualquer par de vértices;
- Uma MST com n vértices, possui $n - 1$ arestas.

Dado um grafo valorado (ou seja, com uma função peso associada a suas arestas) há diversos algoritmos que podem ser utilizados para se encontrar a MST. Para os nossos propósitos, o algoritmo ideal é o de Prim que será explicado na próxima subseção.

1.3.8 Algoritmo de Prim

A premissa do algoritmo de Prim é que a árvore "cresça"a partir de um vértice determinado. Geralmente, o algoritmo do Prim desenvolve uma MST selecionando arestas com o menor peso a cada passo. Funciona da seguinte forma: primeiro, seleciona aleatoriamente um vértice inicial no grafo e, em seguida, procura o vértice adjacente ao vértice inicial com o menor peso. Posteriormente, o algoritmo busca o vértice adjacente ao vértice anterior com o menor peso até a segunda condição da MST ser cumprida.

Vamos ilustrá-lo com o seguinte exemplo: É dado o grafo do item (a) da figura 3. Supomos que o algoritmo comece os passos pelo vértice a. O próximo passo é verificar qual aresta incidente tem o menor peso. No caso, trata-se da aresta (a, b). Agora, o próximo passo é verificar o peso das próximas arestas incidentes a qualquer um dos vértices já conectados. Observe que há duas a se considerar. A aresta (b, c) e a aresta (a, h). Ambas com peso igual a 8. Nesse caso, o algoritmo escolhe uma das duas aleatoriamente (Isso não prejudica o resultado. Mais adiante, este caso será explicado). Agora que a árvore em construção possui três vértices, novamente se aplica o algoritmo e vemos que a próxima aresta é a (i, c). Esta sucessão de passos vai sendo aplicada sucessivamente até que todos os vértices sejam computados.

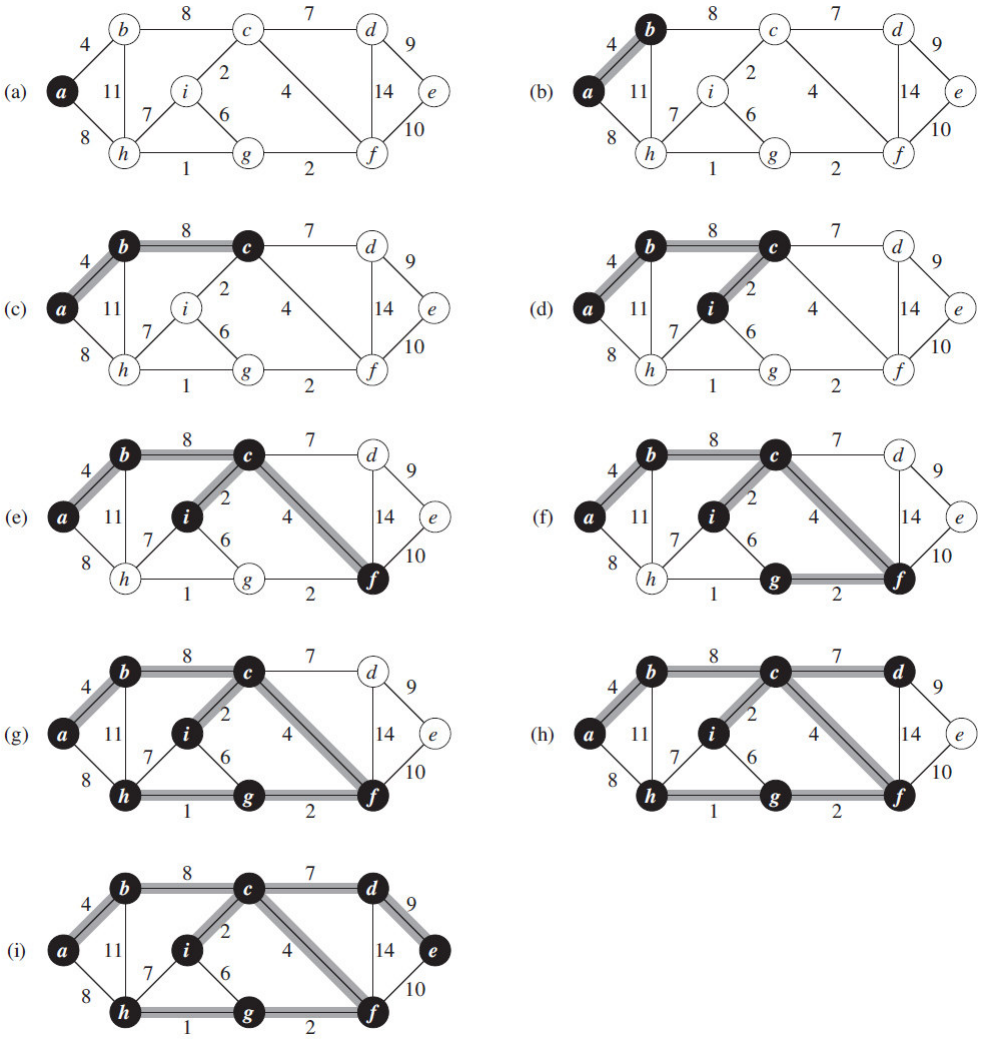

Figura 3 – Passos do algoritmo de Prim[3]

O peso total desta MST é igual a 37. Mas o que aconteceria se o algoritmo tivesse escolhido (a, h) em vez de (b, c)? Nesse caso, a árvore permaneceria quase igual, mudando apenas essas duas arestas. O peso total ainda seria 37. Este é um caso especial em que um grafo possui mais de uma MST.

2 REDE DE DADOS

Sabemos que todo grafo possui uma matriz de adjacência. Se um grafo é valorado, então cada elemento a_{ij} dessa matriz representa o peso da aresta entre o vértice i e o vértice j. A matriz de distância de correlação pode ser vista como uma matriz de adjacência de um determinado grafo.

Para exemplificar o método de análise de dados, iremos começar com um exemplo em que a matriz de distância já nos é dada: As eleições da AdUnicamp em 2018. Temos uma tabela com os votos e percentuais de cada departamento. A matriz surge a partir de uma distância definida (que satisfaz os axiomas de distância) a partir da porcentagem de votos para a Chapa 1 ou Chapa 2 de cada departamento em relação ao outro. O que se obtém é a matriz simétrica da figura 4.

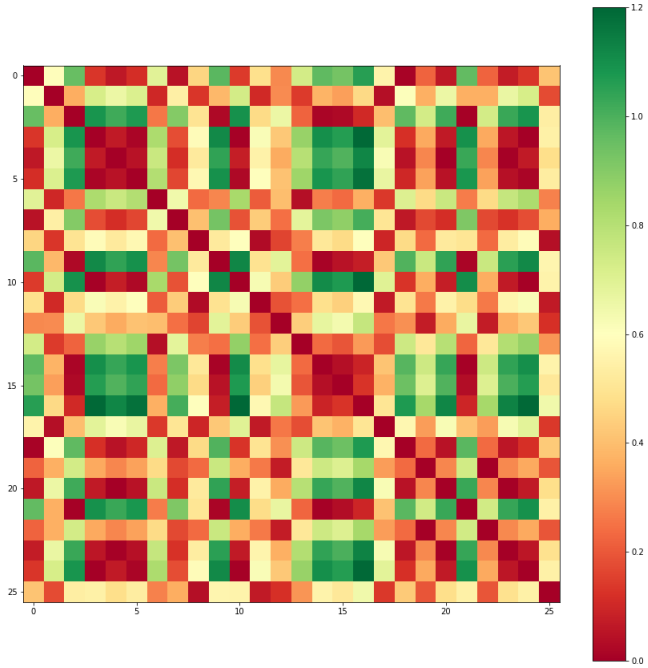

Figura 4 – Exemplo de matriz de distância

As cores representam os valores indo de um mínimo 0 (vermelho mais escuro) a um máximo 2 (azul mais escuro). O que isso significa? É posssível tirar alguma conclusão ou padrão? Devemos lembrar que as linhas e colunas não estão ordenadas seguindo algum critério pré-estabelecido. Que grafo esta matriz representa? Usando-a como matriz de adjacência,

obtemos um grafo completo. Com essa informação, podemos usar o software Python que possui bibliotecas que calculam a MST a partir de uma matriz de adjacência qualquer[7]. Usando esta matriz de distância e aplicando os valores obtidos no software PAJEK para o desenho da rede, obtemos a seguinte MST do grafo:

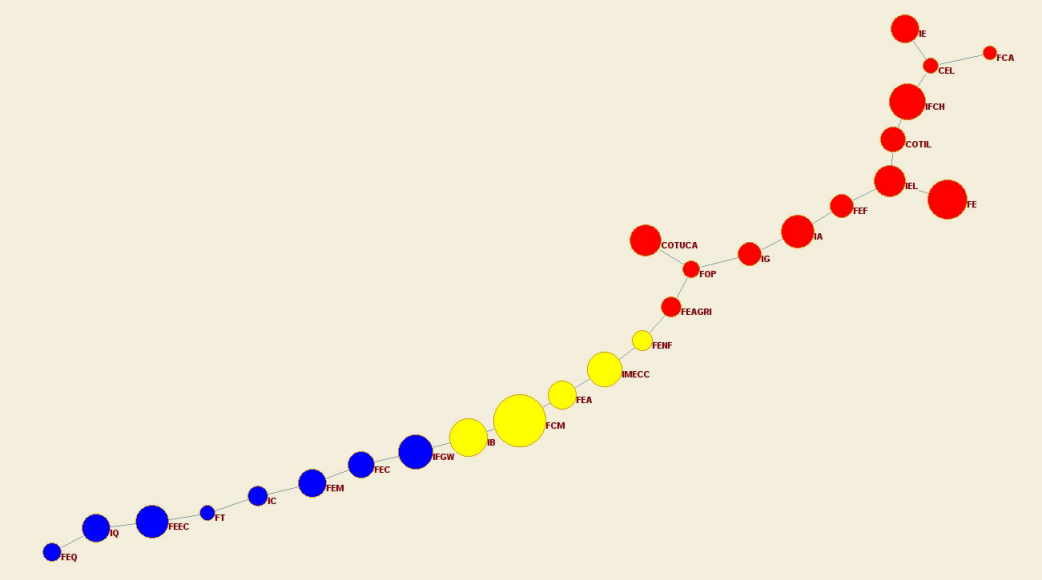

Figura 5 – MST do grafo dado

As cores na árvore representam as maiorias votantes de uma determinada chapa. A idéia é que usando a MST iremos enxergar os grupos de variáveis que se correlacionam mais a partir do grafo. O MST se organiza em alguns agrupamentos, logo, espera-se que o mesmo ocorra com a matriz. A ordem dos passos se segue:

- Os vértices (doenças) da ligação de menor distância são os dois primeiros.
- Depois se descobre a aresta incidente à essa última com a menor distância
- Repete-se o item anterior com a última aresta

Estas instruções equivalem ao algoritmo de Prim com uma pequena modificação que é iniciar a árvore pela aresta mais curta. Quando as variáveis (equivalente as arestas) na matriz são organizadas seguindo o algoritmo de Prim, as mais correlacionadas ficam mais próximas seguindo uma ordem de distância. Ao aplicarmos no exemplo da matriz dada:

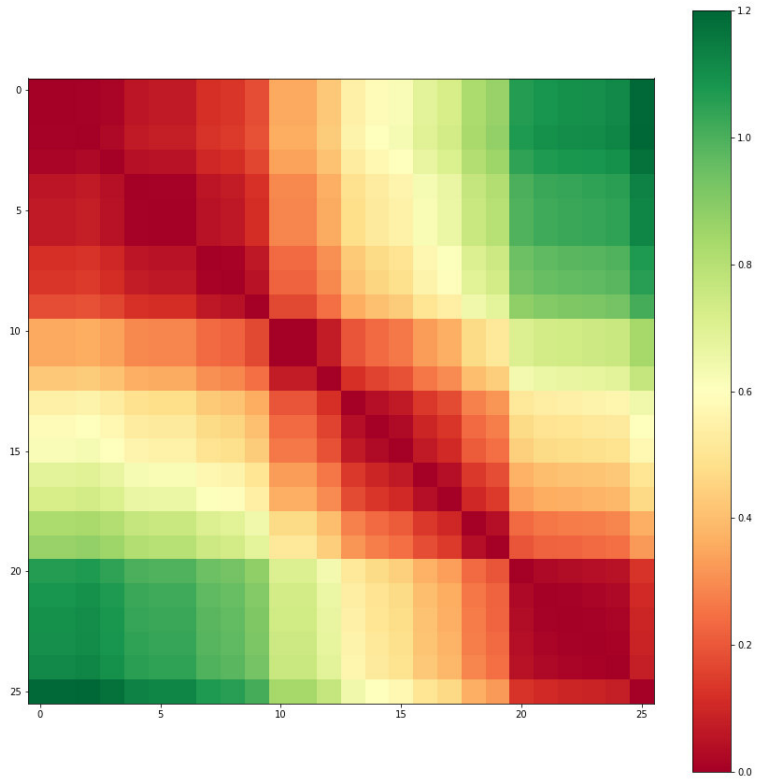

Figura 6 – Matriz de distância reorganizada

Vemos que a separação em grupos é notória. A distância é menor nas diagonais e em suas proximidades e maior nos extremos opostos. Há três clusters formados. Dois nos extremos da diagonal indicando a proximidade entre os que tem maioria votando na mesma chapa, e um mais difuso no meio mostrando justamente os que são mais divididos entre votos. Agora passemos a aplicar estes métodos na análise de dados das doenças.

3 RESULTADOS

Os dados a serem analisados são de doenças (ou qualquer outra enfermidade) que resultaram em óbito ocorridas em mulheres em ínício de maternidade e em pessoas hospitalizadas de até a idade de 19 anos no estado do Ceará. Cada doença tem um código específico que é dado pelo Código Internacional de Doenças (CID-10). As várias formas de sarampo por exemplo, são codificadas como B05.

Acontece que em nossos dados há várias doenças "semelhantes"que foram registradas. Por exemplo, há códigos diferentes para tuberculose se ela se manifestar no pulmão direito ou esquerdo. Por isso, além dos códigos (que já agrupam um certo número de doenças), decidimos que nossos dados-base a serem analisados serão Grupos de Doenças. Esses grupos foram separados com base em alguma semelhança arbitrária em suas doenças de forma que um caso de ocorrência nesse grupo significa que alguma dessas doenças se manifestou na amostragem.

Pacientes	A00 - A09	A15 - A19	...	Y10 - Y34	Y83 - Y84
1	0	0	...	0	0
2	0	0	...	0	0
⋮	⋮	⋮	...	⋮	⋮
6235	0	0	...	0	0
6236	0	0	...	0	0

Tabela 2: Amostragem

O primeiro vetor coluna representa os diferentes pacientes dos quais se obtiveram os dados de doenças. Os dados são binários, ou seja, 1 significa ocorrência e 0 significa não ocorrência. Há um total de 117 grupos de doenças. O que buscamos é a medida de correlação entre essas doenças. Cada vetor-doença é uma variável. Logo, iremos obter uma matriz de 117 linhas e 117 colunas.

Na MST, os grupos estarão representados em 22 cores que representam os capítulos. Estes por sua vez, representam um grupo ainda mais abrangente de doenças. Esta classificação vem diretamente do Ministério da Saúde.

Capítulo	Tipo de doença
I	Algumas doenças infecciosas e parasitárias
II	Neoplastias (tumores)
III	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
IV	Doenças endócrinas nutricionais e metabólicas
V	Transtornos mentais e comportamentais
VI	Doenças do sistema nervoso
VII	Doenças dos olhos e anexos
VIII	Doenças do ouvido e da apófise mastóide
IX	Doenças do aparelho circulatório
X	Doenças do aparelho respiratório
XI	Doenças do aparelho digestivo
XII	Doenças da pele e do tecido subcutâneo
XIII	Doenças do sistema osteomuscular e do tecido conjuntivo
XIV	Doenças do aparelho geniturinário
XV	Gravidez parto e puerpério
XVI	Algumas afecções originadas no período perinatal
XVII	Malformações congênitas, deformidades e anomalias cromossômicas
XVIII	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte
XIX	Lesões, envenenamento e algumas outras consequências de causas externas
XX	Causas externas de morbidade e mortalidade
XXI	Fatores que influenciam no estado de saúde e o contato com os serviços de saúde
XXII	Códigos para propósitos especiais

Tabela 3: Capítulos e suas doenças abrangidas[1]

Aplicando diretamente a fórmula de correlação de Pearson (que foi definida no primeiro capítulo), obtemos a seguinte matriz de correlação:

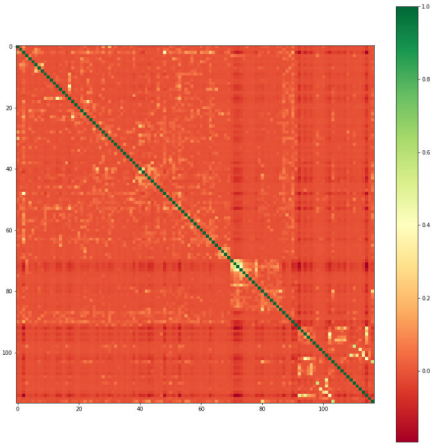

Figura 7 – Matriz de correlação

Como esperado, ele não nos dá respostas de imediato. Apenas que a correlação é máxima nas diagonais o que é uma obviedade matemática. Agora iremos desenhar a MST. Os grupos de doenças serão separados em cores e os vértices serão proporcionais a raiz quadrada do número de ocorrências. Cada cor representa um capítulo definido pelo próprio Ministério da Saúde. Plotando a MST, vemos cerca de seis aglomerados de grupos de doenças:

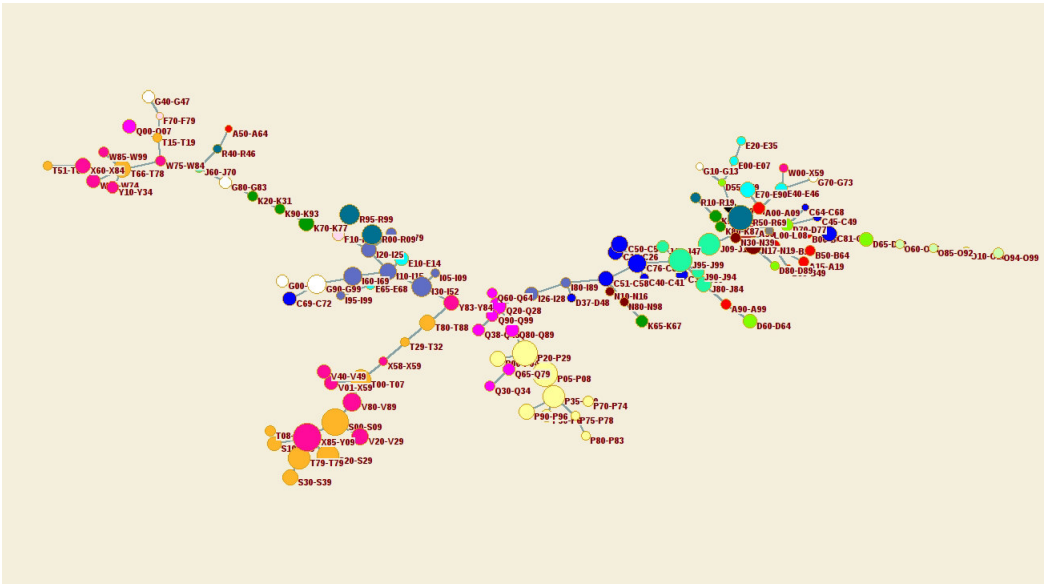

Figura 8 – MST do grafo usando Pearson

Agora a matriz rearranjada de acordo com a MST:

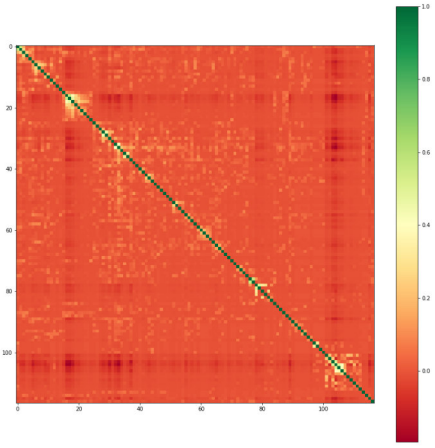

Figura 9 – Correlação (Pearson) e distância rearranjado por MST

Há alguns grupos com correlações razoavelmente maiores se formando nas diagonais, mas são bem pequenos e pouco perceptíveis. Bem longe do nítido exemplo da AdUnicamp. Isso mostra que as doenças tendem a ser bem pouco correlacionadas. É possível melhorar esta análise com base na nitidez destes clusters? Podemos fazer uma espécie de "maquiagem"nos dados de forma a realçar o que está mais correlacionado? Um simples clareamento das cores não surtiria efeito desejado.

3.1 Correlação para dados binários

O problema dos clusters anteriores serem pouco conclusivos se deve ao fato de que a grande maioria dos elementos da sua matriz original (Pacientes x Grupos de Doenças) são iguais zeros. A correlação de Pearson trabalha com médias e elas tendem a ir ao seu valor mínimo em casos como esse. É necessário "forçar"um contraste maior entre as variáveis se quisermos enxergar melhor os possíveis clusters que se formam.

Uma forma de fazer isso é eliminando o excesso de zeros da média calculada nos valores esperados da correlação. Imaginemos que o programa está calculando a média entre duas variáveis D_i e D_j. O valor da média subiria se eliminássemos da contagem de N (o numerador do estimador) as vezes em que nenhuma das doenças aparecem, ou seja, quando $D_i = D_j = 0$. De imediato, percebemos que os valores de correlação irão crescer em medidas diferentes para cada elemento da matriz. Mas prossigamos com o método.

-	D_i	D_j
1	1	0
2	0	0
3	0	0
4	0	0
5	0	1
6	0	0
7	1	0
8	0	0
9	0	1
⋮	⋮	⋮
6319	0	0
6320	0	0
6321	0	0
6322	1	0
6323	1	0
6324	0	1
6325	0	0
6326	0	0

$\longrightarrow$

-	D_i	D_j
1	1	0
5	0	1
7	1	0
9	0	1
⋮	⋮	⋮
6322	1	0
6323	1	0
6324	0	1

Tabela 4: “Filtragem”dos dados de duas doenças

É possível fazer isso usando o fato dos dados serem binários. Relembrando a fórmula de correlação:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} \tag{3.1}$$

Onde

$$\begin{aligned} cov(x,y) &= \frac{1}{N}\sum_n (x_n - \mu_x)(y_n - \mu_y) \\ &= \frac{1}{N}\left[\sum_n x_n y_n - \frac{1}{N}\left(\sum_n x_n\right)\left(\sum_n y_n\right)\right] \\ &= \frac{1}{N}\sum_n x_n y_n - \mu_x\mu_y \end{aligned}$$

e o produto dos desvios padrões pode ser escrito como

$$\sigma(x)\sigma(y) = \sqrt{\left[\frac{1}{N}\sum_n (x_n - \mu_x)^2\right]\left[\frac{1}{N}\sum_n (y_n - \mu_y)^2\right]} \tag{3.2}$$

Se observarmo bem, vemos que a covariância entre dois conjuntos de variáveis iguais é igual ao seu desvio padrão ao quadrado.

$$cov(x,x) = (\sigma(x))^2 = \frac{1}{N}\sum(x_n - \mu_x)^2 = \frac{1}{N}\sum_n x_n^2 - \mu_x^2$$

Acontece que os valores de x_n e y_n são binários. Logo, $x_n^2 = x_n$ e $y_n^2 = y_n$.

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_n x_n - \mu_x^2} = \sqrt{\mu_x - \mu_x^2} = \sqrt{\mu_x(1-\mu_x)}$$

De onde chagamos a uma nova fórmula para o produto dos desvios padrões

$$\sigma(x)\sigma(y) = \sqrt{\mu_x\mu_y\big(1-\mu_x\big)\big(1-\mu_y\big)} \tag{3.3}$$

Nessas equações, $N = 6236$ para todos os cálculos de média. Há uma forma de eliminar os pares $(0,0)$ da contagem e consequentemente fazer N diminuir (e elevar a média calculada). Levando em consideração que tratam-se de dados binários, existem duas coisas a se observar

Soma entre variáveis. Existem quatro possíveis resultados:

1. $(x_n, y_n) = (0,0) \rightarrow x_n + y_n = 0$
2. $(x_n, y_n) = (1,0) \rightarrow x_n + y_n = 1$
3. $(x_n, y_n) = (0,1) \rightarrow x_n + y_n = 1$
4. $(x_n, y_n) = (1,1) \rightarrow x_n + y_n = 2$

Definimos agora o número de vezes que cada par aparece:

- $N_{(0,0)} \rightarrow$ Número de vezes que $(0,0)$ aparece
- $N_{(1,0)} \rightarrow$ Número de vezes que $(1,0)$ aparece
- $N_{(0,1)} \rightarrow$ Número de vezes que $(0,1)$ aparece
- $N_{(1,1)} \rightarrow$ Número de vezes que $(1,1)$ aparece

Observamos que

$$\sum_i^N (x_n + y_n) = N_{(1,0)} + N_{(0,1)} + 2N_{(1,1)} \tag{3.4}$$

Produto entre variáveis. Existem quatro possíveis resultados:

1. $(x_n, y_n) = (0, 0) \rightarrow x_n y_n = 0$
2. $(x_n, y_n) = (1, 0) \rightarrow x_n y_n = 0$
3. $(x_n, y_n) = (0, 1) \rightarrow x_n y_n = 0$
4. $(x_n, y_n) = (1, 1) \rightarrow x_n y_n = 1$

Agora isso nos dá o número de vezes que o par $(1, 1)$ aparece:

$$\sum_{i}^{N} x_n y_n = N_{(1,1)} \tag{3.5}$$

Com isso, podemos obter o número N_{xy} de pares ordenados diferentes de $(0, 0)$:

$$\begin{aligned} N_{xy} &= N_{(i,j)\neq(0,0)} = N_{(1,0)} + N_{(0,1)} + N_{(1,1)} && (3.6) \\ &= \sum_{n}(x_n + y_n) - \sum_{n} x_n y_n && (3.7) \end{aligned}$$

Esse é o valor que irá substituir N em todas essas expressões anteriores para o cálculo dos estimadores.

3.2 Correlação para diferentes valores de N

Ao aplicarmos diretamente a fórmula anterior, iremos nos deparar com um problema. Haverão casos onde μ_x ou μ_y serão iguais a um (por exemplo, duas doenças que se manifestaram em um único paciente) e isso acarretará em divisões por zero. Uma maneira simples de contornar esse problema é introduzindo um pequeno "erro"no cálculo de N_{xy}. Seguiremos com alguns exemplos e suas possíveis implicações.

3.2.1 $N_{xy} + 1$

Inicialmente faremos o mais próximo possível da nova fórmula para N. Somando com 1, seria equivalente a considerar apenas um par ordenado $(0, 0)$ a mais na soma N_{xy}. As novas matrizes de correlação e distância se seguem:

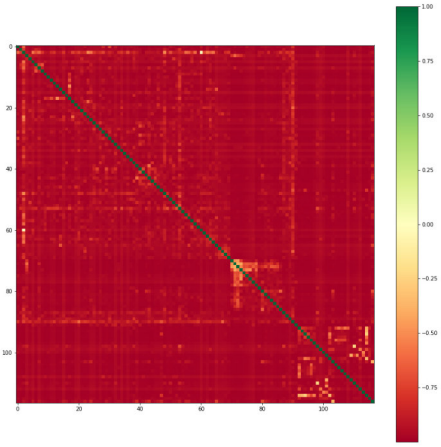

Figura 10 – Correlação com $N = N_{xy} + 1$

São bem uniformes como as primeiras matrizes. Talvez até um pouco mais. Isso indica que os resultados serão novamente pouco visíveis. Mesmo assim, prossigamos calculando a MST e ver se ocorrem mudanças.

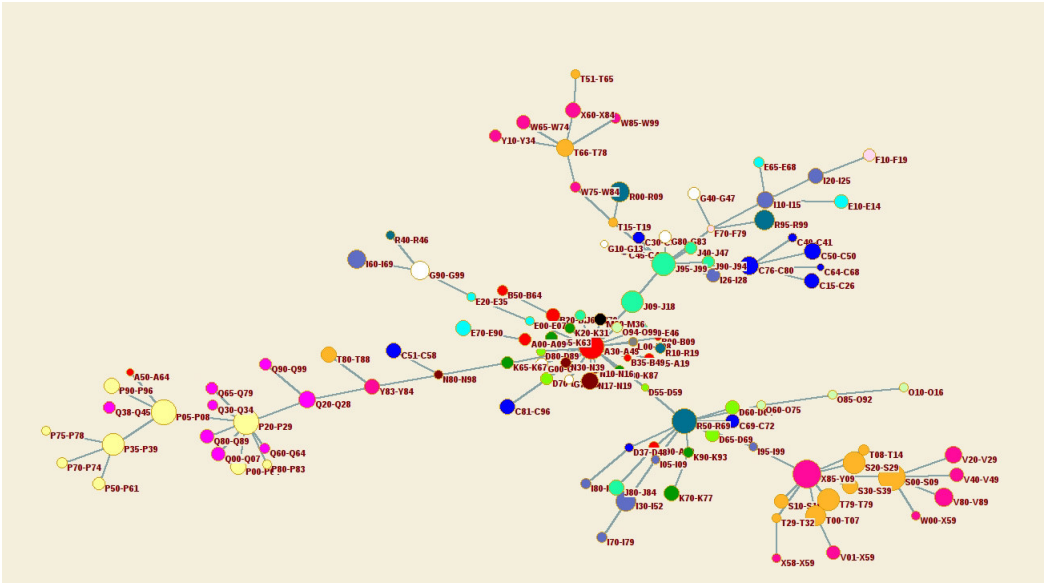

Figura 11 – MST do grafo usando $N = N_{xy} + 1$

Há dessa vez cinco aglomerações em torno de nós visíveis. Houve notórias mudanças em relação a MST original. Por exemplo, K20-K31 que estava ligado a K90-K93 e G80-G83 foi ligar-se a A30-A49. Reordenando as matrizes, ficamos com:

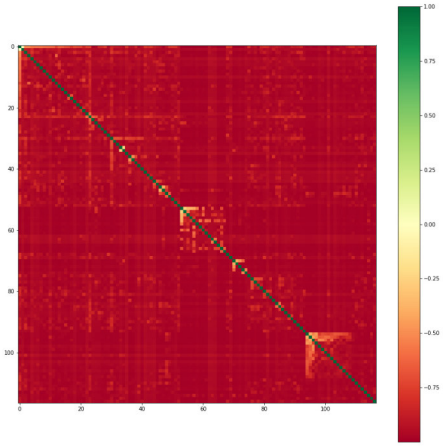

Figura 12 – Correlação rearranjado por MST usando $N = N_{xy} + 1$

Como esperado, ainda pouco conclusivo, embora possamos ver dois pequenos cluster com maior correlação. Apenas a MST nos dá respostas mais conclusivas nesse caso. Passemos para o próximo.

3.2.2 $N_{xy} + 20\%$

Aqui se faz um simples acréscimo de 20% ao valor calculado em N_{xy}. Eis as novas matrizes:

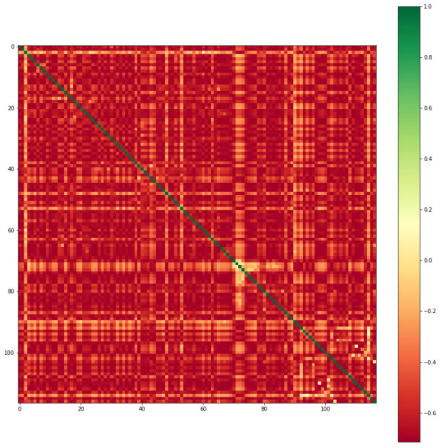

Figura 13 – Correlação com $N = N_{xy} + 20\%$

Finalmente contrastes bem maiores. Novamente, não será supresa ver que estas nova configuração altera a MST novamente.

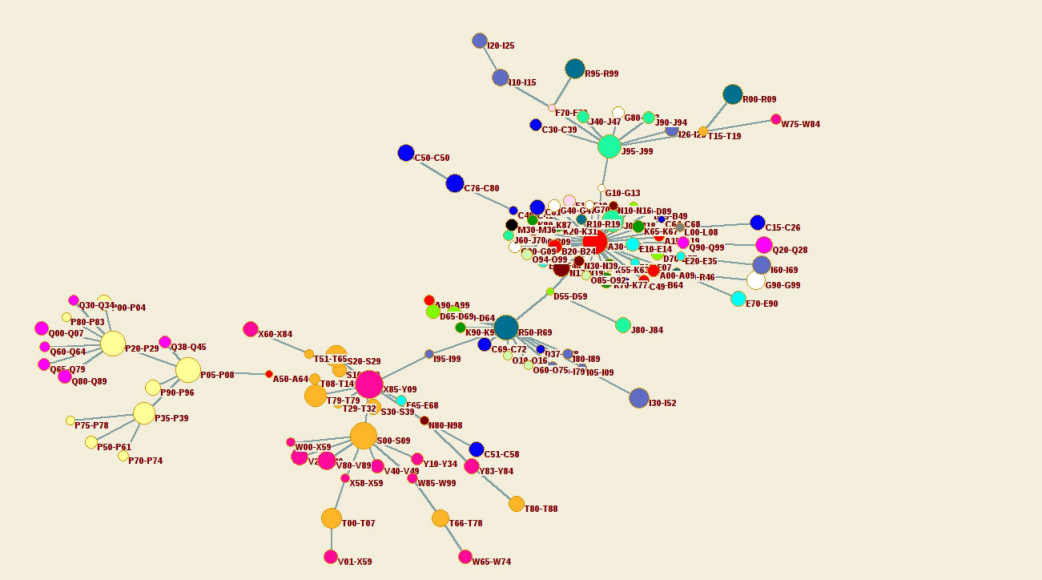

Figura 14 – MST do grafo usando $N = N_{xy} + 20\%$

A primeira vista, a separação em agrupamentos em torno dos nós se intensificou ainda mais. Como no item anterior, formam-se cinco clusters. Nota-se o surgimento de um padrão. A medida que modificamos a média, os nós citados parecem se correlacionar mais com outros vértices. É possível ver que vários outros vértices passaram a estar ligados com outros a medida que mudamos o cálculo de N, mas boa parte deles foram se acumular em alguns desses nós (que são os mesmos da média anterior). É interessante notar o nó A30-A49. Este abrange algumas doenças infecciosas. Há muitas outras doenças formando um cluster com ele e indicando alta correlação. Uma boa explicação está no fato de que entre as doenças infecciosas está a septicemia que é uma infecção comum em ambientes hospitalares. Ela costuma atacar muitas pessoas que já estão com outras enfermidades, daí podemos ter uma ideia do porquê dessa alta correlação de várias doenças não-correlacionadas com um único nó. Reordenando as matrizes:

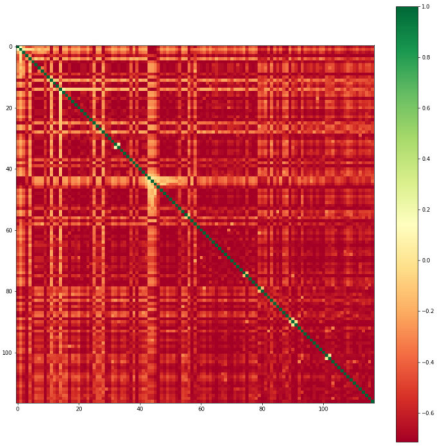

Figura 15 – Correlação e distância rearranjado por MST usando $N = N_{xy} + 20\%$

Não são clusters tão organizados (pelo menos em uma observação direta) quanto os da figura 6, mas não são inconclusivos. Esta "cruz"no segundo quadrante (quase ao meio) seria justamente A30-A49. Ele está rodeado de variáveis não-correlacionadas, que é justamente o que se vê ao se analisar a MST.

3.2.3 $N_{xy} = 2093$

2093 é o valor máximo obitido com a fórmula (3.9) para N_{xy} nos dados das doenças.

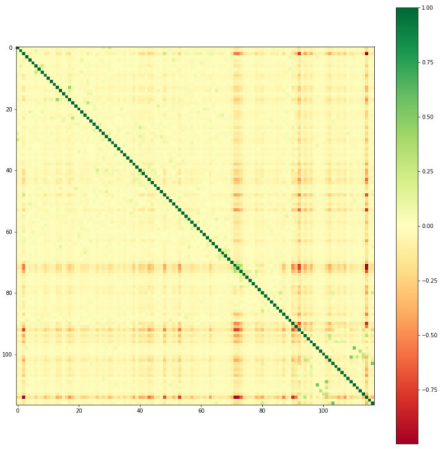

Figura 16 – Correlação e distância com $N = 2093$

A correlação ficou bem menos negativa, mas a nitidez voltou a ficar ruim. Vejamos o que ocorre com a MST.

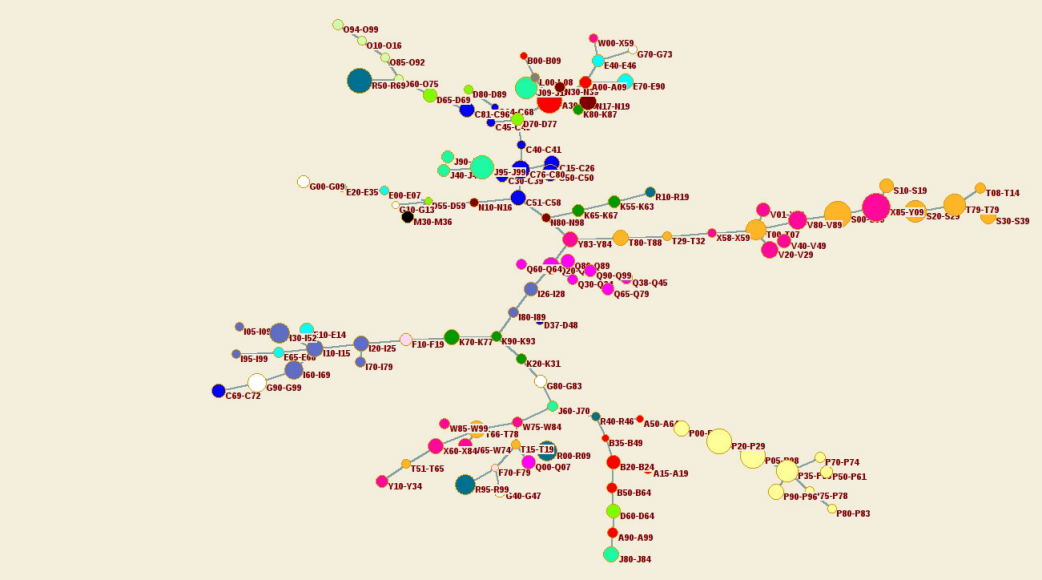

Figura 17 – MST do grafo usando $N = 2093$

Esta se asemelha mais a primeira MST, mas ainda ocorrem muitas diferenças. Por outro lado, este resultado está de acordo com o padrão que se tem visto nas MST's. Todas as médias calculadas acima são maiores do que a média calculada usando $N = 2093$. Ao se aproximar do valor original de N, mais a árvore tende a se parecer com a primeira MST. Reordenando as matrizes:

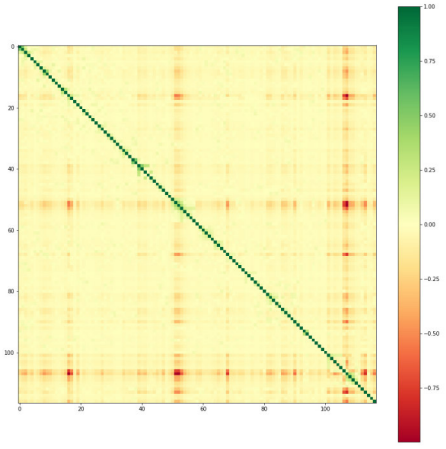

Figura 18 – Correlação e distância rearranjado por MST usando $N = 2093$

Pouco conclusivo como esperado. Parece que nesse estágio está havendo regressão em se tratando de nitidez. Isso mostra que há um limite para esta técnica utilizada no cálculo de N_{xy}

4 CONCLUSÃO

A intenção inicial deste trabalho era apenas propor uma forma de melhorar a visibilidade dos clusters em matrizes de correlação reordenadas cujos elementos são pouco correlacionados. No entanto, logo ficou evidente que se paga um preço por determinadas alterações nas fórmulas matemáticas: A confiabilidade dos resultados é posta em dúvida. Cada vez que se altera sensivelmente os valores das médias calculadas as MST's mudavam, e com isso os agrupamentos se distanciavam dos originais. Seria interessante saber se há uma "modificação"limite tal que a MST não mude (Ou pelo menos, as mudanças sejam ignoráveis). Com certeza os procedimentos utilizados podem ser aprimorados para tornar isso possível. Restaria saber se a nitidez ainda seria razoável.

No entanto, mesmo com essas dúvidas, obteve-se resultados muito interessantes. O melhor exemplo descritivo foi o do grupo A30-A49 e sua ligação com várias doenças não-correlacionadas. As separações por cores nas MST's também ficaram razoáveis. Os grupos P (Amarelo claro) e Q (Rosa choque) formam um cluster bem visível mesmo na correlação de Pearson. Eles se referem a enfermidades ocorridas no período perinatal e a mal-formações congênitas respectivamente. A relação entre os casos é óbvia. Mais um exemplo de cluster ocorre entre os grupos S e T, ambos em amarelo escuro. Tratam-se de casos de envenenamento, agressão, acidentes e etc. Formam dois clusters com X (Rosa choque), disparos com armas de fogo, e V que engloba os acidentes de trânsito ou outros veículos.

Isso mostra que mesmo posta estas dúvidas, os resultados ainda podem se manter fiéis a realidade sem grandes alterações.

É evidente que usar MST é uma técnica apropriada para análise de dados. O software Python, por exemplo, é uma ferramenta que dispõe de uma série de bibliotecas que facilitam o trabalho com grafos e cálculos estatísticos. Com isso, é razoável supor que há um grande potencial nesta técnica e que pode ser utilizada em redes ainda maiores e mais complexas.

REFERÊNCIAS

[1] MORBIDADE Hospitalar do SUS. Website. Disponível em: <http://tabnet.datasus.gov.br/cgi/sih/mxcid10.htm>.

[2] C., D. *Finding a "not-shortest" path between two vertices*. April 2012.

[3] CORMEN, T. H. *Introduction to Algorithms*. [S.l.: s.n.], 1994. Lectures.

[4] CAMACHO, L. F. M. *Brazilian House of Representatives Analysis from Network Theory Perspective*. Tese (Mestrado) — UNICAMP, 2017.

[5] JURKIEWICZ, S. *Grafos - Uma introdução*. June 2009.

[6] MELO, G. S. de. *Introdução à Teoria dos Grafos*. Tese (Mestrado) — UFPB, 2014.

[7] FERREIRA, A. dos S. *Análise de uma rede de mortalidade utilizando teoria dos grafos*. Tese (Doutorado) — UFC, 2017.