Uma aplicação do modelo exponencial de grafos aleatórios para dados sobre representação social de Zika vírus

Marina Alves Amorim

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

30 de Novembro de 2017

- Autor: Marina Alves Amorim
  Instituto de Ciências Exatas - Departamento de Estatística
- Orientador: Prof. Dr. Gilvan Ramalho Guedes
  Faculdade de Ciências Econômicas - Departamento de Demografia
- Co-orientadora:Profa. Dra. Denise Duarte
  Instituto de Ciências Exatas - Departamento de Estatística
- Colaboradores: Wesley Henrique Silva Pereira
  Graduado em Estatística pela Universidade Federal de Minas Gerais
  Dr. Rodrigo Botelho Ribeiro
  Doutor em Matemática pela Universidade Federal de Minas Gerais

Motivação

- Embora as epidemias de Zika já ocorrecem na África, sudeste Asiático e ilhas do Pacífico, foi no final de 2015 que a infecção pelo vírus tomou carater epidêmico nas Américas.
- Apesar das intensas campanhas promovidas pelas autoridades de saúde e pela mídia, ainda pouco se conhece sobre a interpretação que as pessoas dão a doença e se elas estão transformando o conhecimento adiquirido em práticas preventivas.
- Estudos sobre pensamentos coletivos sugerem que vários fatores atuam no sentido de promover diferenças na forma como os grupos pensam sobre epidemias. Nível de exposição, gênero e diferenciais socioeconômicos são alguns fatores responsáveis por essa heterogeneidade.

Objetivo

Objetivo Geral: Identificar e compreender como se representa o pensamento coletivo sobre Zika vírus em Governador Valadares-MG, suas heterogeneidades em subgrupos e quais variáveis explicam os padrões de afinidade cognitiva formados pela rede.

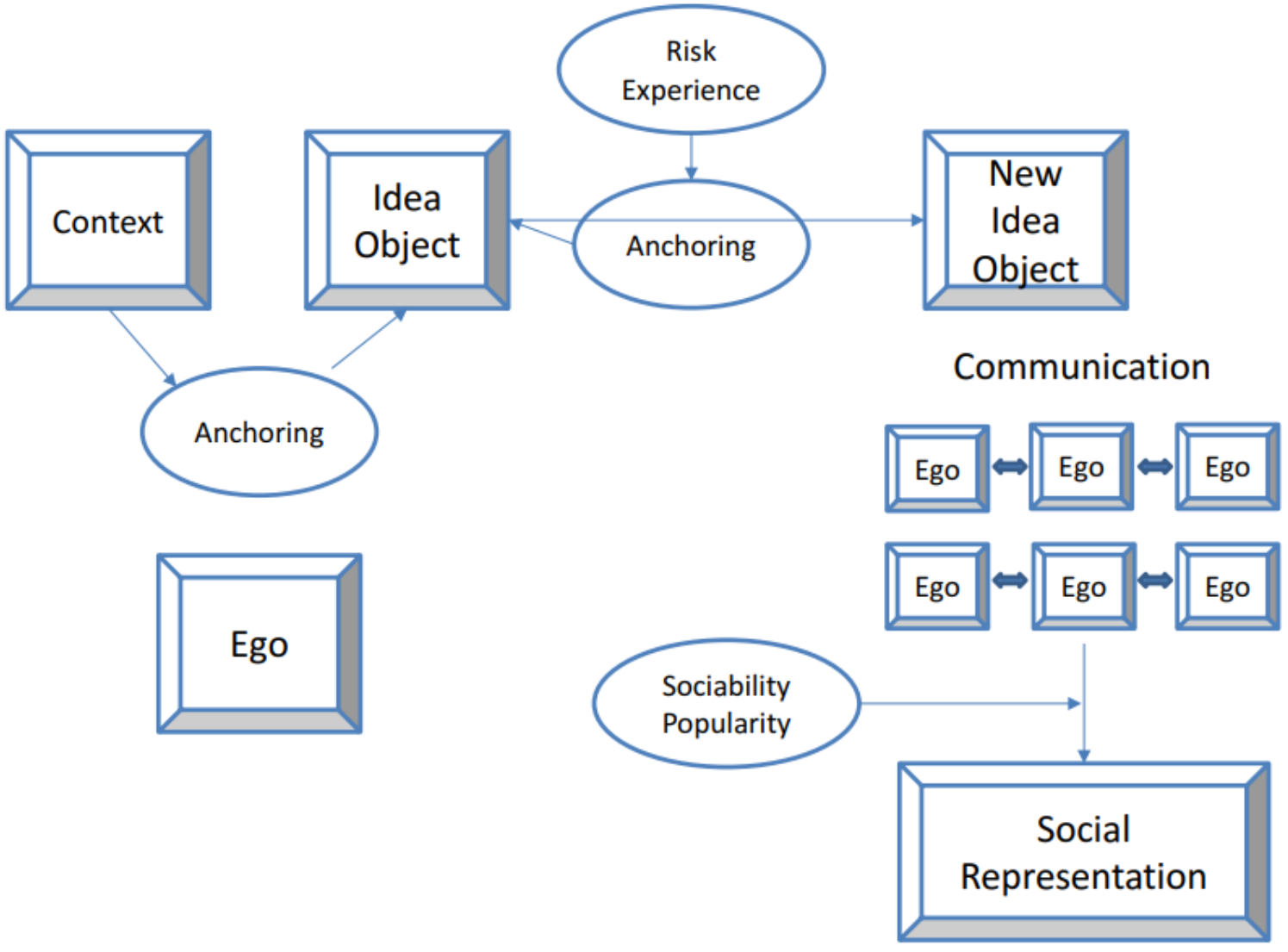
Estratégia empírica

1. Coletar amostras do pensamento individual sobre o Zika vírus;
2. Reconstruir a estrutura do pensamento coletivo combinando informações geradas pela TALP e aplicando a fórmula de afinidade cognitiva de Pereira 2017;
3. Particionar a rede de afinidades gerada com base na modularidade;
4. Identificar a capacidade dessas partições e de atributos individuais de explicar os padrões de ligações cognitivas observados na amostra.

Fundamentação Teórica

Teoria das Representaçãoes Sociais:

- Se dedica à representação de objetos sociais sob uma perspectiva coletiva, embora preserve a individualidade pessoal (Moscovici,1961);
- Considera os fatores: cultura, comunicação e memória coletiva (social/coletivo), além do fator cognitivo (individual);


Risk
Experience
Context
Idea
Object
Anchoring
New
Idea
Object
Anchoring
Communication
Ego
Ego
Ego
Ego
Ego
Ego
Ego
Sociability
Popularity
Social
Representation

Captura e representação do pensamento coletivo

Técnica de Associação Livre de Palavras (TALP):

- Mecanismo capaz de capturar o pensamento leigo;
- Expressões/palavras livremente reportadas e ordenadas mediante a apresentação de um termo indutor;

Abordagem em Grafos

Definição de grafo:

- Conjunto $G = (V, E)$, onde V são vértices, a unidade fundamental do grafo e E são arestas que conectam pares de vértices;

Vantagens da abordagem em grafos:

- Ancoragem em um modelo formal e matemático;
- Particionamento através de suas características;
- Verificação da partição encontrada através de simulação;
- Identificação de indivíduos populares;
- Análise espacial da difusão dos significados.

Pressupostos da estruturação da rede

- Cada indivíduo é representado por um vértice;
- Indivíduos se conectam ao compartilharem pelo menos uma evocação;
- O grafo representativo da rede deve ser não direcionado;
- As ordens de importância das evocações para os indivíduos são conhecidas;
- Logo, as conexões podem ser ponderadas levando em consideração a ponderação das ordens de evocação através do coeficiente de afinidade proposto por Pereira 2017.

Coeficiente de Afinidade

Sejam u e v dois vetores de evocações, sendo n_u e n_v o número de evocações expressadas nesses vetores. Além disso, seja $n = max(n_u, n_v)$ e N o comprimento máximo de um vetor de evocações. O coeficiente de afinidade entre esses vetores é calculado por:

$$\alpha(u,v) = \underbrace{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[\theta(i,j,n) * \rho(i,j,n) * 1_{u_i=v_j}]}{n^2 * (n+1)}}_{\beta(u,v)} * \underbrace{1 - \frac{(N-n)(N-n+1)}{N(N+1)}}_{\omega(n)} \quad (1)$$

Onde:

- $\theta(i,j,n) = 2 * (n+1) - (i+j)$;Ponderação por Ordem
- $\rho(i,j,n) = n - |i-j|$; Ponderação por distância
- $0 \leq \beta(u,v) \leq 1$ e $0 \leq \alpha(u,v) \leq \omega(n)$; Penalizador

Medidas Descritivas

Grau ponderado

- Soma dos pesos das arestas incidentes;
- Nível de atividade dos vértices;
- Medida de popularidade.

Densidade

- Proporção de arestas dentro da rede;
- Coesão dos pensamento coletivo da rede.

Informações gerais dos dados e da rede

Dados:

- "Demografia da Exceção, intenções reprodutivas e migração em um contexto de Zika vírus e desastres socioambientais"
- O questionário possui 34 itens, dos quais 4 itens relacionados ao instrumento da TALP sobre o tema "Zika vírus" foram explorados;
- Estratificação por sexo, Escolaridade e Nível de Exposição;
- 150 entrevistados.

Rede:

- 106 evocações únicas padronizadas;
- 643 evocações;
- 135 vertices;
- 5476 arestas ponderadas pelo coeficiente de afinidade;

Modularidade e Algorítmo de Louvain

A modularidade é um coeficiente utilizado na detecção de comunidades em uma rede. Assumindo valores entre -1 e 1. A fórmula para calcular a modularidade pode ser parametrizada em (Newman, 2006):

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - P_{i,j}] \delta(c_i, c_j) \tag{2}$$

onde $A_{i,j}$ representa o peso da aresta que liga os vértices i e j; m representa a soma dos pesos de todas as arestas contidas no grafo; $P_{i,j}$ é o valor esperado para o peso da aresta que liga os vértices i e j. $\delta(c_i, c_j)$ refere-se ao delta de Kroneker, que tem funcionamento muito similar ao de uma função indicadora. O coeficiente Consiste em calcular a soma das diferenças entre a proporção observada de arestas dentro de cada comunidade e a proporção esperada de arestas segundo algum modelo aleatório dado que a distribuição dos graus é fixa.

Atividade dos indivíduos na rede

Tabela: **Indivíduos menos ativos na rede**

Indivíduo	Grau P.	PALA01	PALA02	PALA03	PALA04	PALA05
35	2,477	Fear	Caos	Cost	——	——
141	4,749	Cough	Join Pain	Shevery	Weakness	Malaise
21	4,977	Epidemics	Government	Hospital	Public Health	Health Profes.
77	5,257	Standing Water	Open Tanks	Discarted tire	Trash	Liters
37	5,803	Weakness	Join Pain	Something Bad	Itch	Tolerable

Tabela: **Indivíduos mais ativos na rede**

Indivíduo	Grau Ponderado	PALA01	PALA02	PALA03	PALA04	PALA05
128	30,73	Standing Water	Mosquito	Desiase	Pain	Hospital
147	30,97	Pain	Mosquito	Itch	Fever	Medical Leave
92	31,59	Pain	Desiase	Mosquito	Mosquito Bite	Fever
140	31,80	Standing Water	Pain	Mosquito	Fever	Malaise
40	31,91	Fever	Pain	Malaise	Mosquito	Desiase

Partição ótima da rede

Tabela: **Número de indivíduos e densidade das comunidades**

Comunidade	Tamanho	Densidade
1 Tratamento	25	0.7966
2 Prevenção	53	0.9064
3 Sintomas	57	0.9041

Modularidade: 0.24

Verificação da partição encontrada

- Verificação via distribuição amostral de (100.000) réplicas de Monte Carlo;
- Comparar proporções esperadas do número de comunidades em grafos aleatórios com o número de comunidades encontrado na rede original.

Verificação da partição ótima

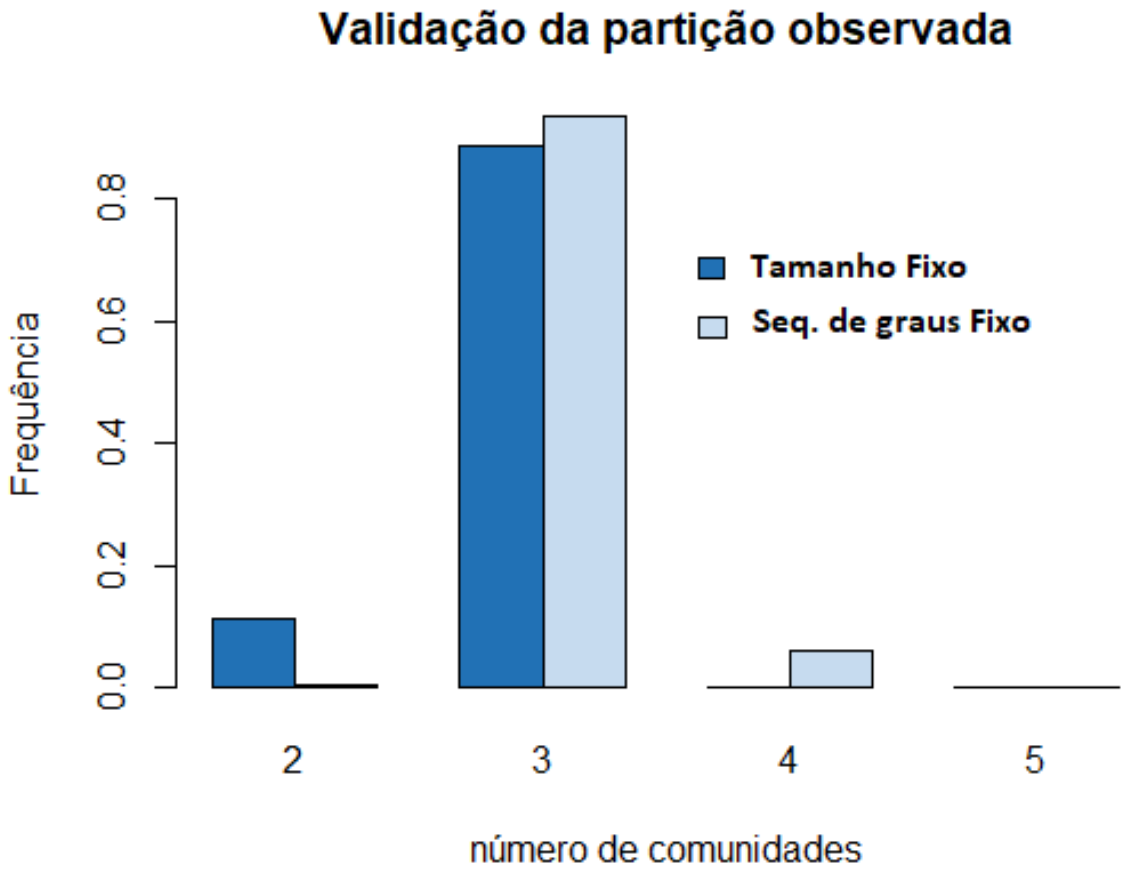
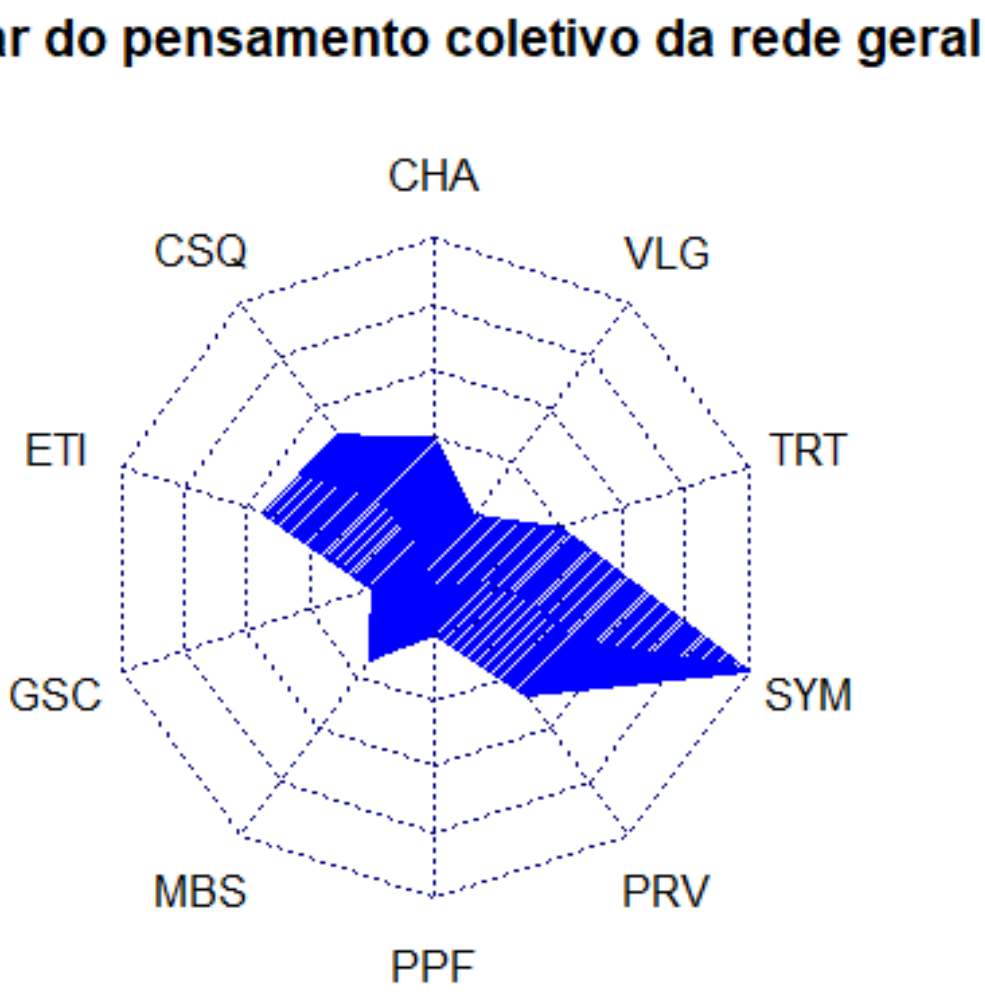
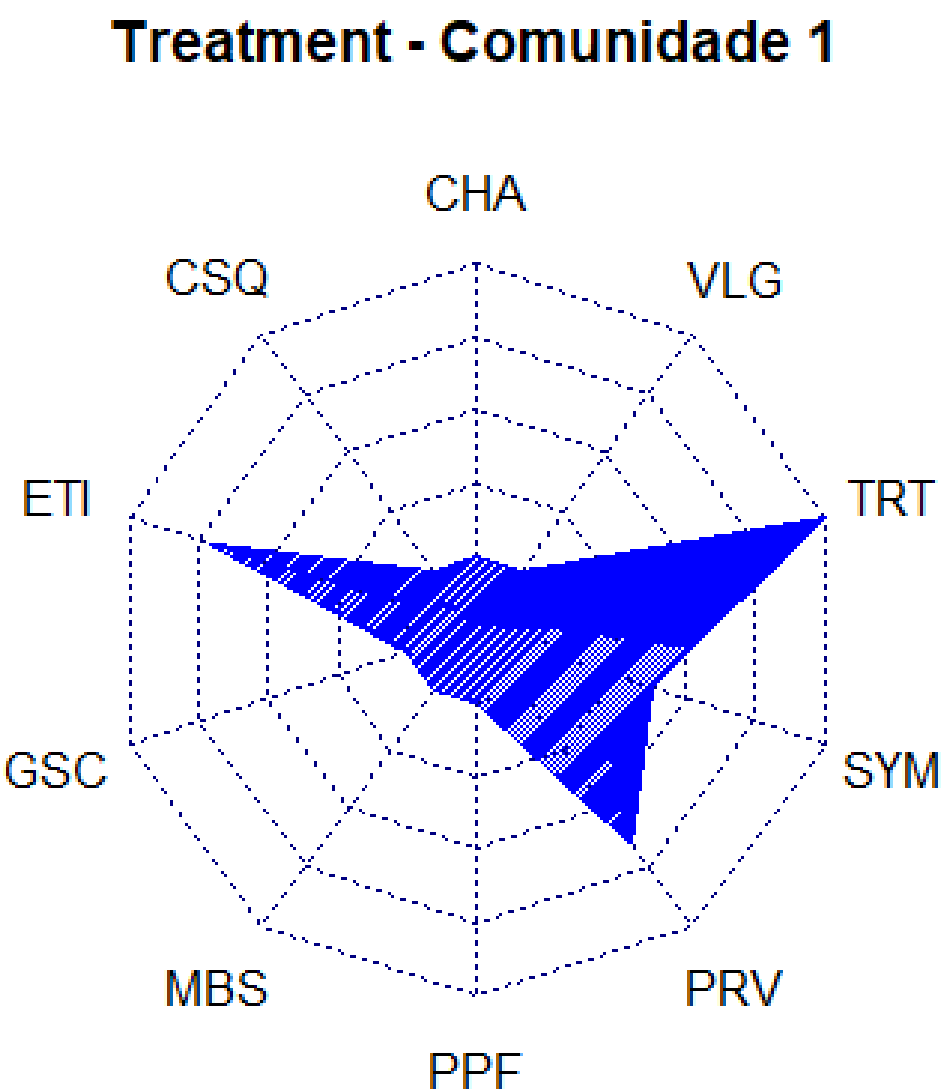


Figura: **Proporções esperadas para o número de comunidades em grafos aleatórios - 100.000 réplicas**

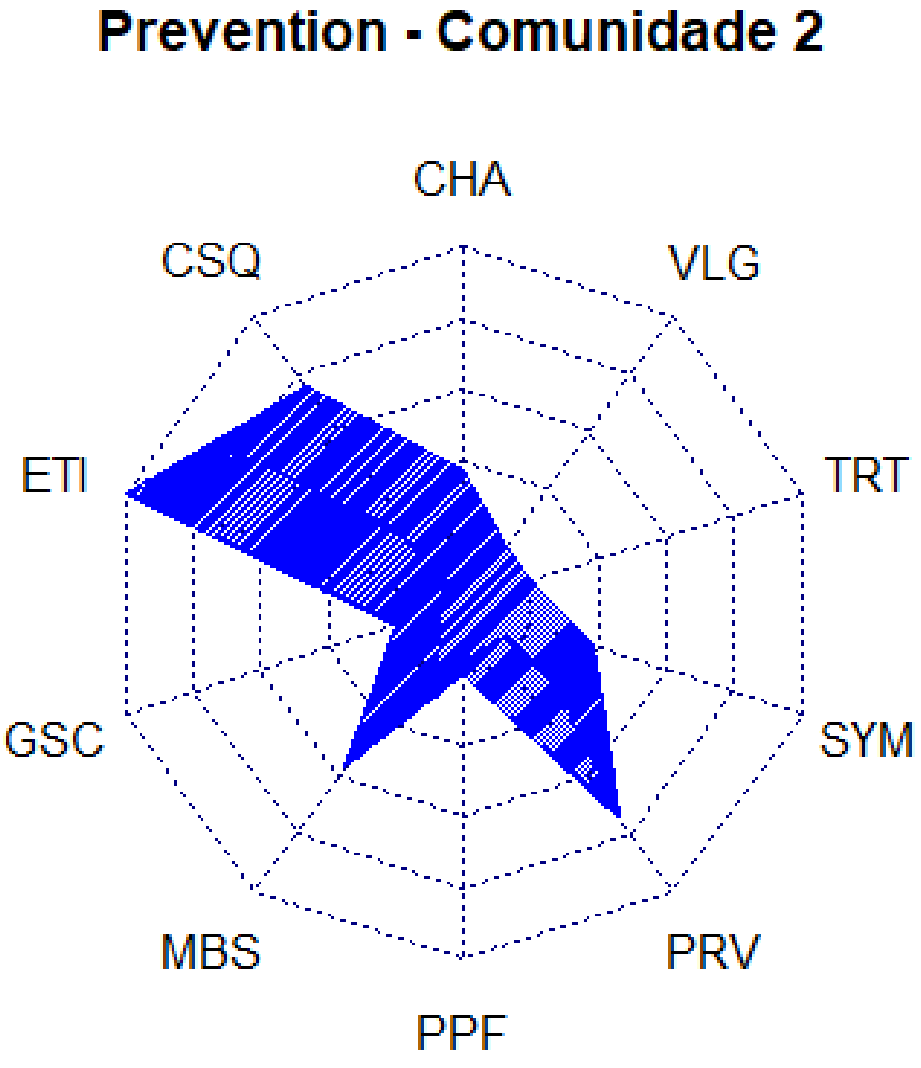
Verificação da partição ótima



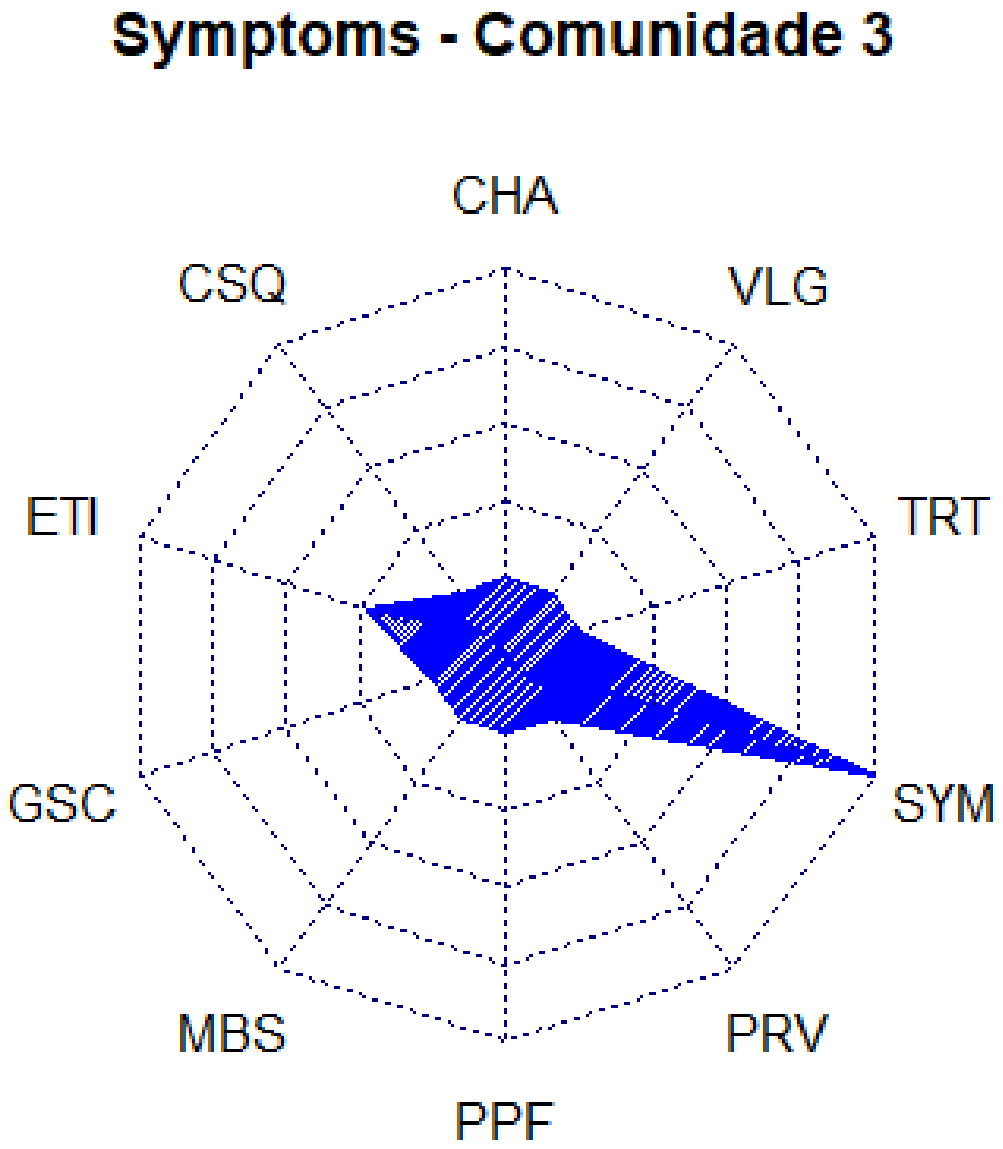
Comparando os pensamentos das comunidades



Comparando os pensamentos das comunidades



Comparando os pensamentos das comunidades



ERGM Bernoulli com cováriaveis

Temos Y sendo uma matriz de adjacências correspondente a rede. A suposição básica dos modelos ERGM é de que a estrutura em um grafo observado y pode ser explicada por uma estatística $s(y)$, dependendo da rede observada. Assim, a probabilidade de observar o grafo y seria dada por

$$P(Y = y|\theta) = \frac{\exp(\theta^T s(y))}{c(\theta)}, \tag{3}$$

onde θ é um vetor de parâmetros do modelo associado a $s(y)$ e $c(\theta)$ é uma constante de normalização. Podemos ainda incluir atributos ao modelo, com objetivo de melhorar a análise, da seguinte forma :

$$P(Y = y|\theta) = \frac{\exp(\theta^T s(y, X))}{c(\theta)}, \tag{4}$$

onde a estatística $s(y, X)$ depende agora de variáveis exógenas à rede, representadas pelo vetor de atributos X.

Um modelo ERGM é construido seguindo a seguinte nomeclatura:
y ~<term1> + <term2> + ...
No qual y é um objeto da rede ou uma matriz, <termo1>, <termo2> são termos escolhidos, como por exemplo arestas e atributos da rede.

- **Nodefactor :** Efeito Principal ou Prevalência - Para variáveis categóricas, compreende a probabilidade de um vértice com determinado atributo fazer ou não conexão com outro vértice de atributo diferente, independentemente das conexões do mesmo. Para variáveis contínuas temos o nodemain.
- **Nodematch :** Efeito de Homofilia - Compreende a probabilidade de que dois vértices com os mesmos atributos formem conexão.

Resultados

Tabela: **Coeficientes do modelo ERGM**

Coeficiente	Estimativa	Razão de Chance
Edge	-0.4114 ***	1.51
Comunidade Prevention	-0.3216 ***	1.38
Comunidade Symptons	-0.1154*	1.12
Escolaridade Média	0.6162 ***	1.8508
Escolaridade Alta	0.38938 ***	1.476
Sexo	0.1481 ***	1.1596
Alter(Nenhuma Doença)	0.5219***	1.6852
Alter(Dengue ou Chikungunya)	-0.39756*	1.48
Alter(Zika)	0.5354***	1.7081
Ego	0.39742***	1.4879

Os coeficientes nos ajudam a explicar a ocorrência de conexões da rede e como varia de acordo com os atributos.

[0] 0 '***' , 0.001 '**' , 0.01 '*'

Bondade de Ajuste

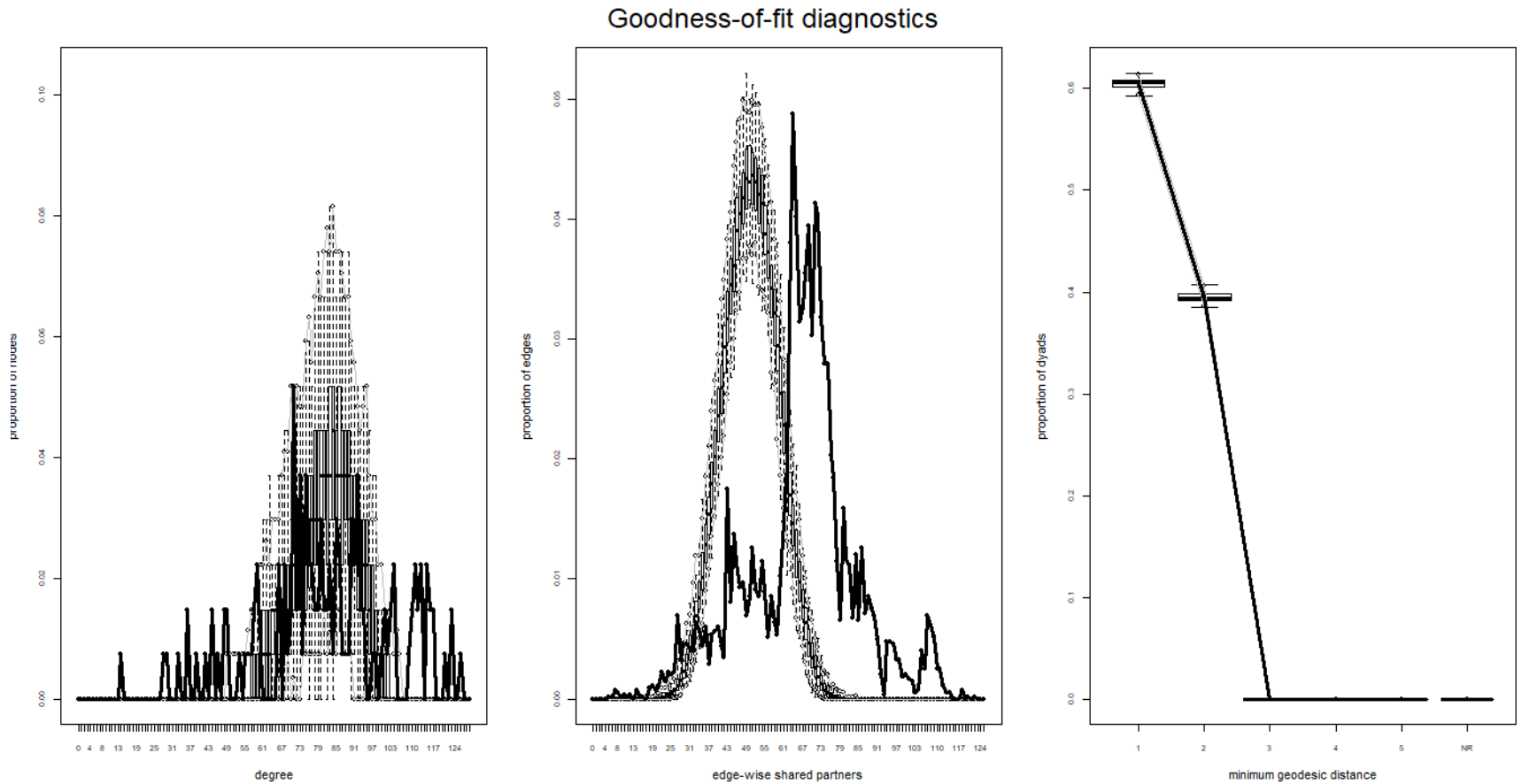


Figura: **Bondade de Ajuste**

Conclusão

- Foi possível perceber algumas diferenças nos focos de pensamentos entre as comunidades, embora a modularidade tenha se mostrado baixa;
- Podemos perceber que o modelo ERGM nos ajuda a quantificar o quanto as cováriaveis influenciam a densidade da afinidade cognitiva.

- J. C. Abric. Pratiques sociales et representations, chapter Las representations sociales: aspects theoriques. Presses Universitaires de France, Paris, 1994.
- Gábor Csárdi and Tamas Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006. URL http://igraph.org.
- S. Moscovici. La psychanalyse, son image et son public. Presses Universitaires de France, Paris, 1961.
- Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical review E, 69(2):026113, 2004.
- Eric D. Kolaczyk and Gábor Csárdi. Statistical Analysis of Network Data with R. Springer. 2014.
- David Hunter et. al. ERGM: A package of fit, Simulate and dignose Exponential Family Models for Networks, Jornal of statistical software, vol.24, 2008.

Obrigada !

Padronização das evocações

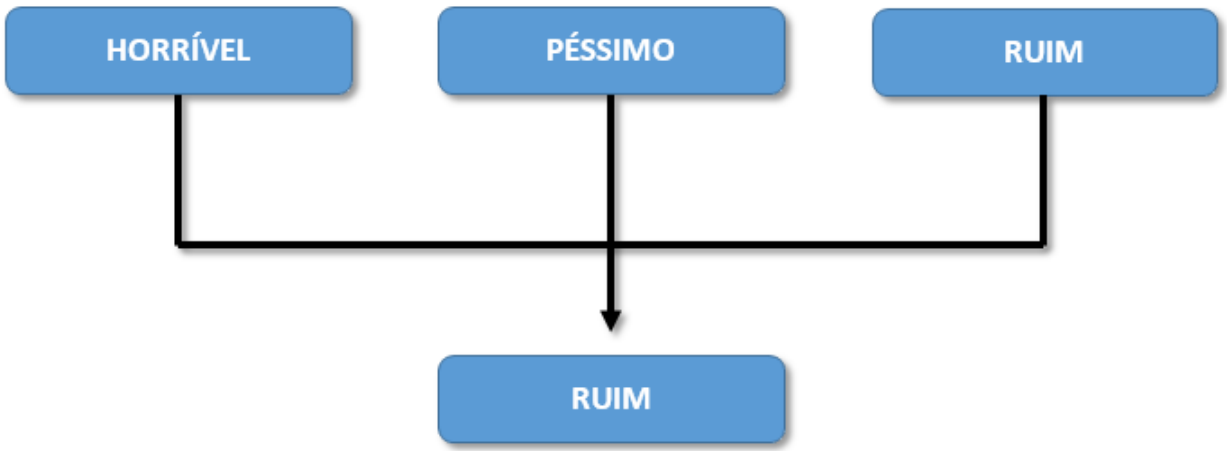


Figura: **Padronização**

Ligação entre 2 indivíduos

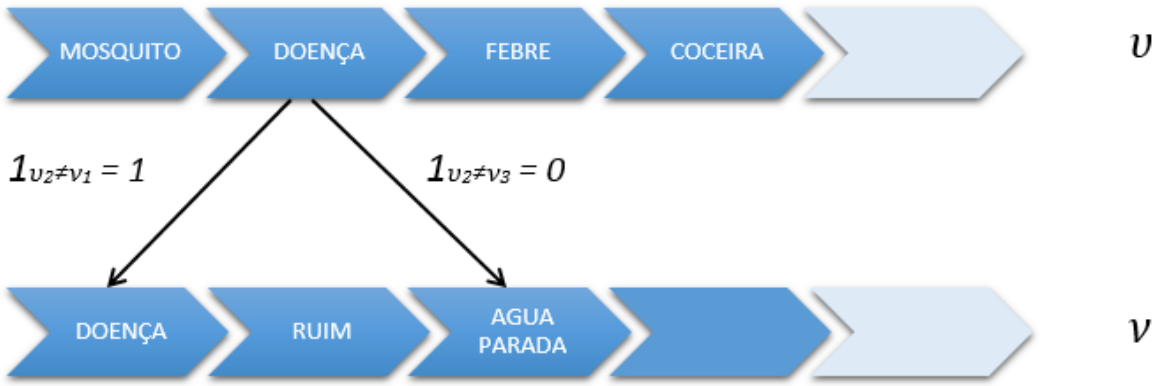
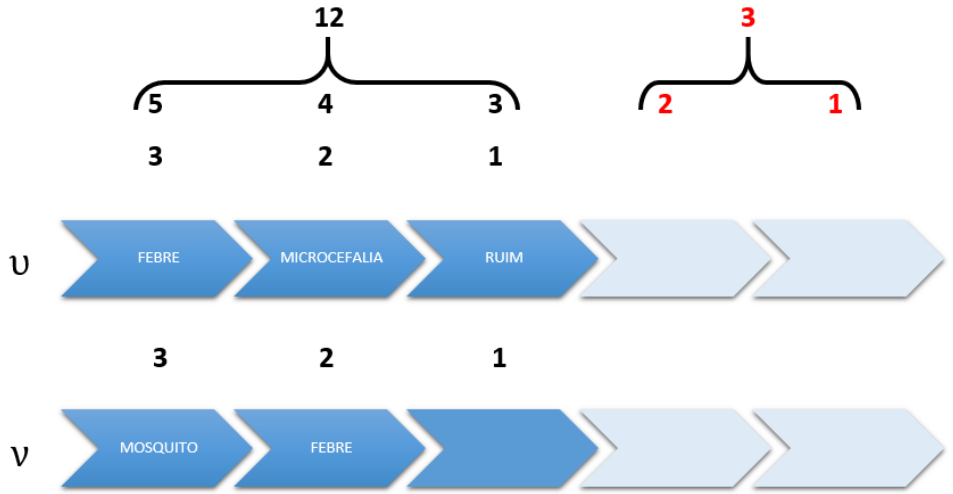


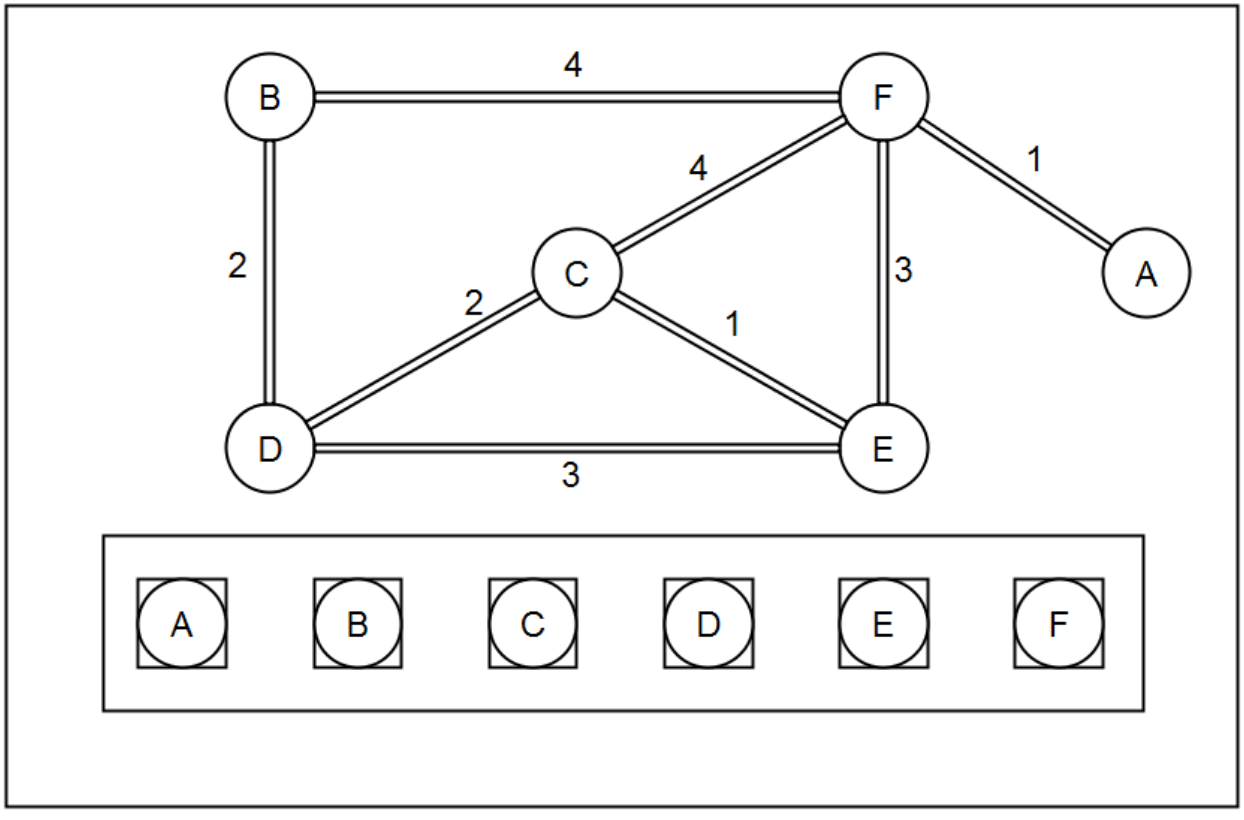
Figura: **Matriz de Adjacência, que recebe 1 se existe ligação e 0 caso contrário**

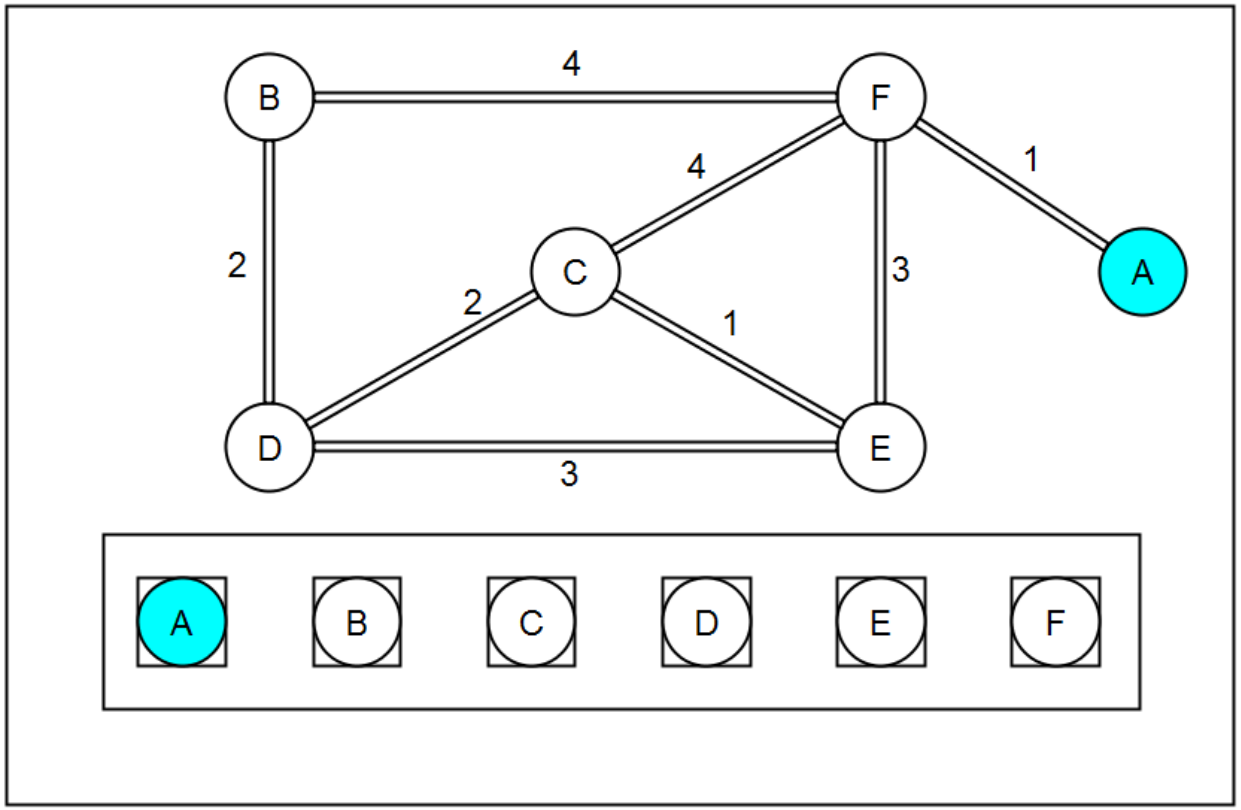
Exemplo de ponderação por Afinidade

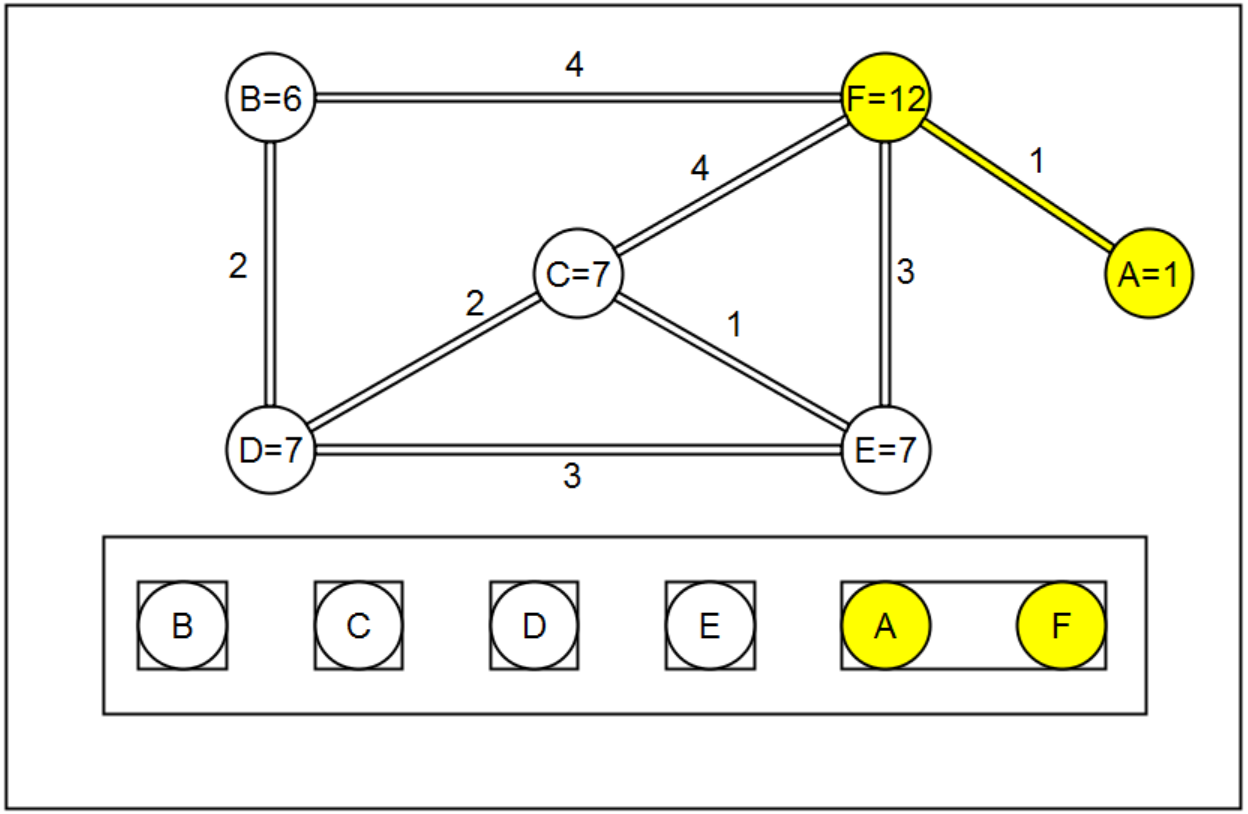


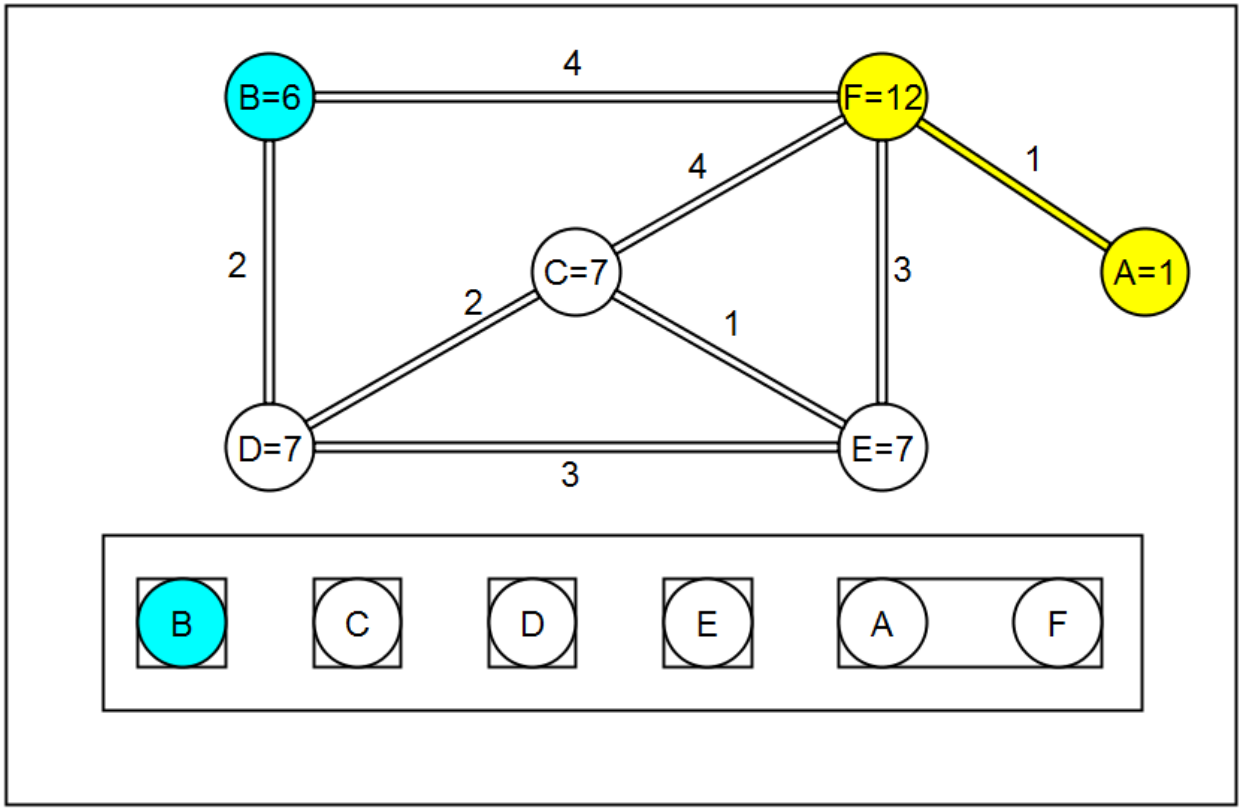
$$I_{u_i v_j} = 0, \forall (i,j) \neq (1,2)$$

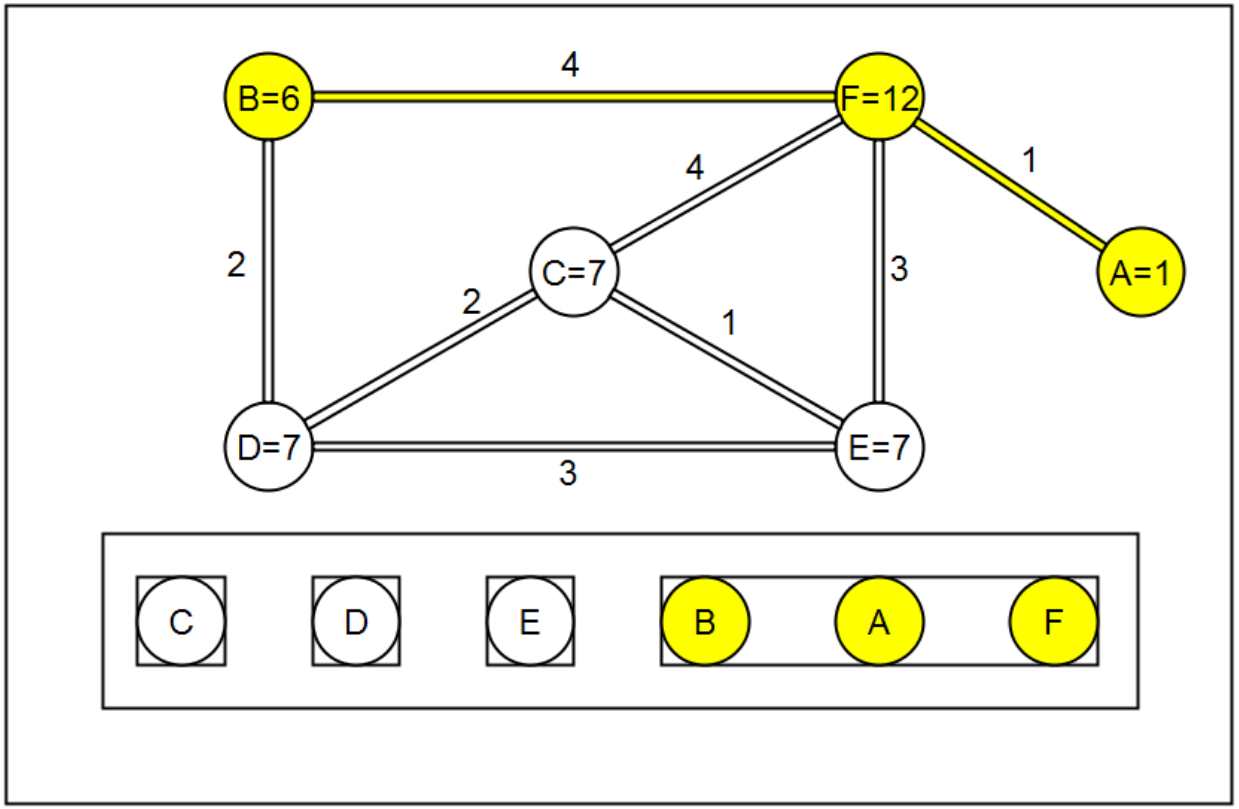
- $\theta(i,j,n) = 2(n+1) - (i+j) \rightarrow \theta(1,2,3) = 2(3+1) - (1+2) = 5$
- $\rho(i,j,n) = n - |i-j| \rightarrow \rho(1,2,3) = 3 - |1-2| = 2$
- $n^2(n+1) \rightarrow 3^2 + (3+1) = 36$
- $\alpha(u,v) = 0.2778[1 - \frac{(5-3)(5-3+1)}{5(5+1)}] = 0.2778 * \frac{12}{15} = 0.222$

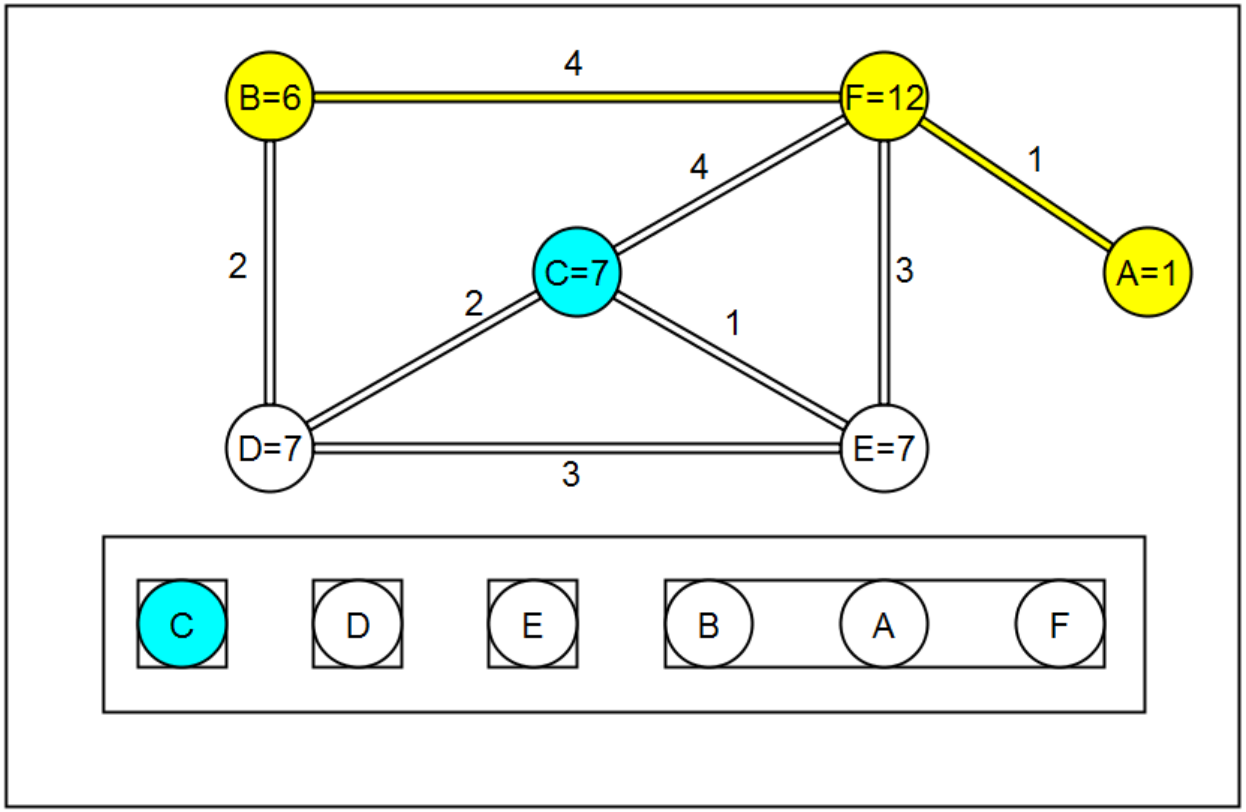

B
F
C
A
D
E
4
4
1
2
3
2
1
3
A
B
C
D
E
F

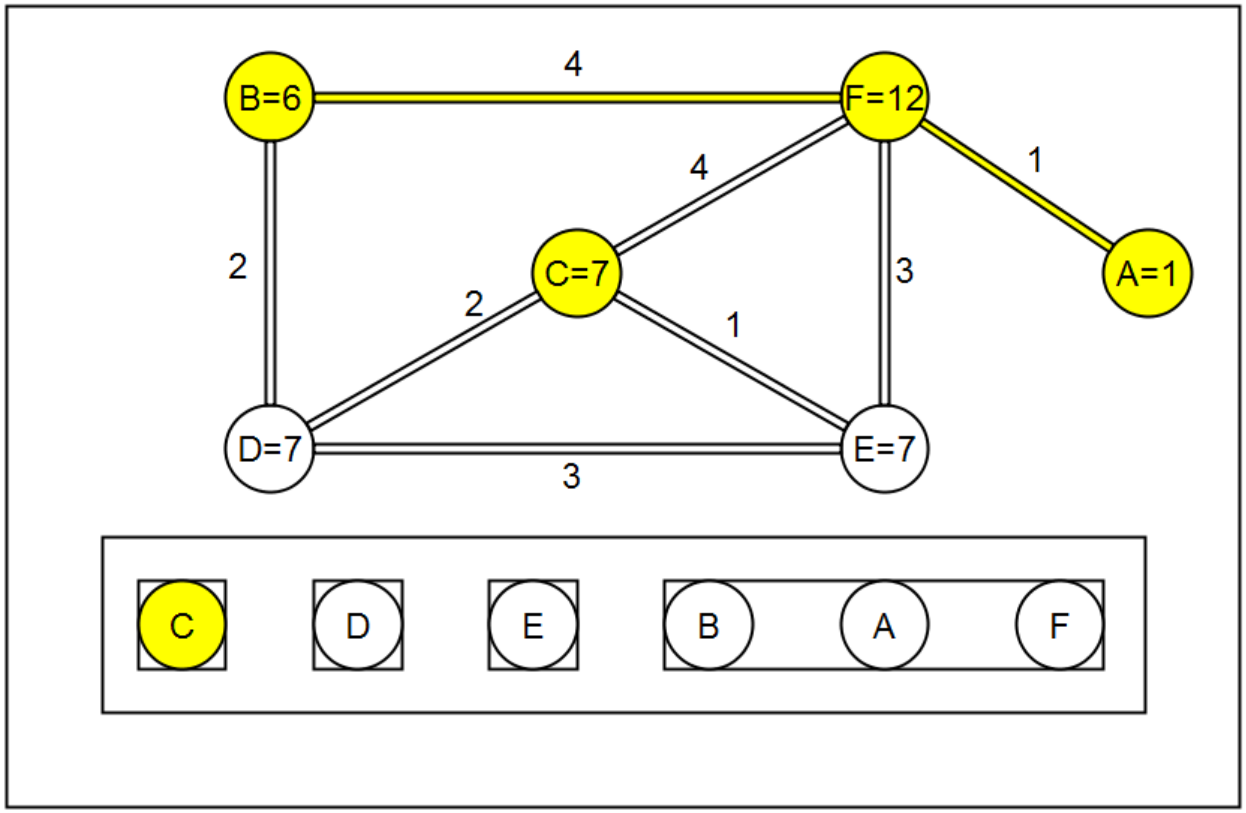

B
F
A
C
D
E
4
4
1
2
2
3
1
3
A
B
C
D
E
F

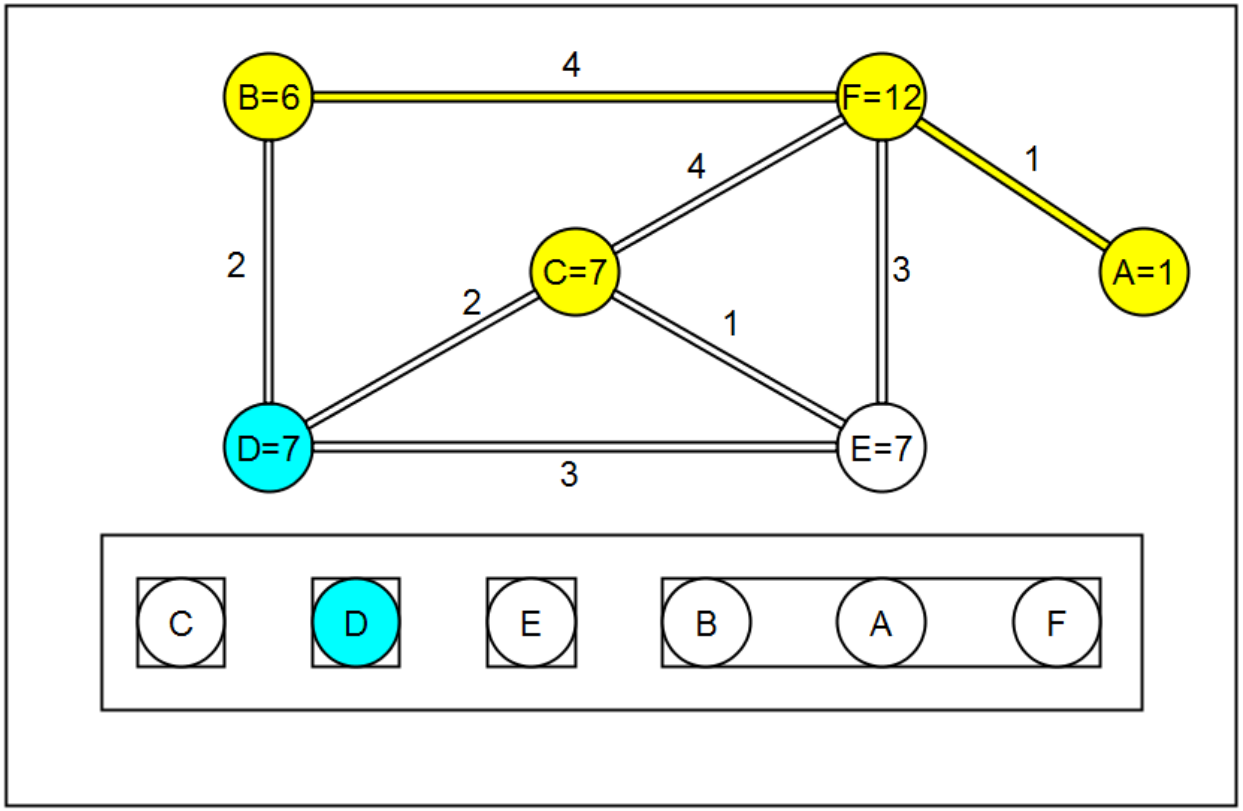

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
B
C
D
E
A
F

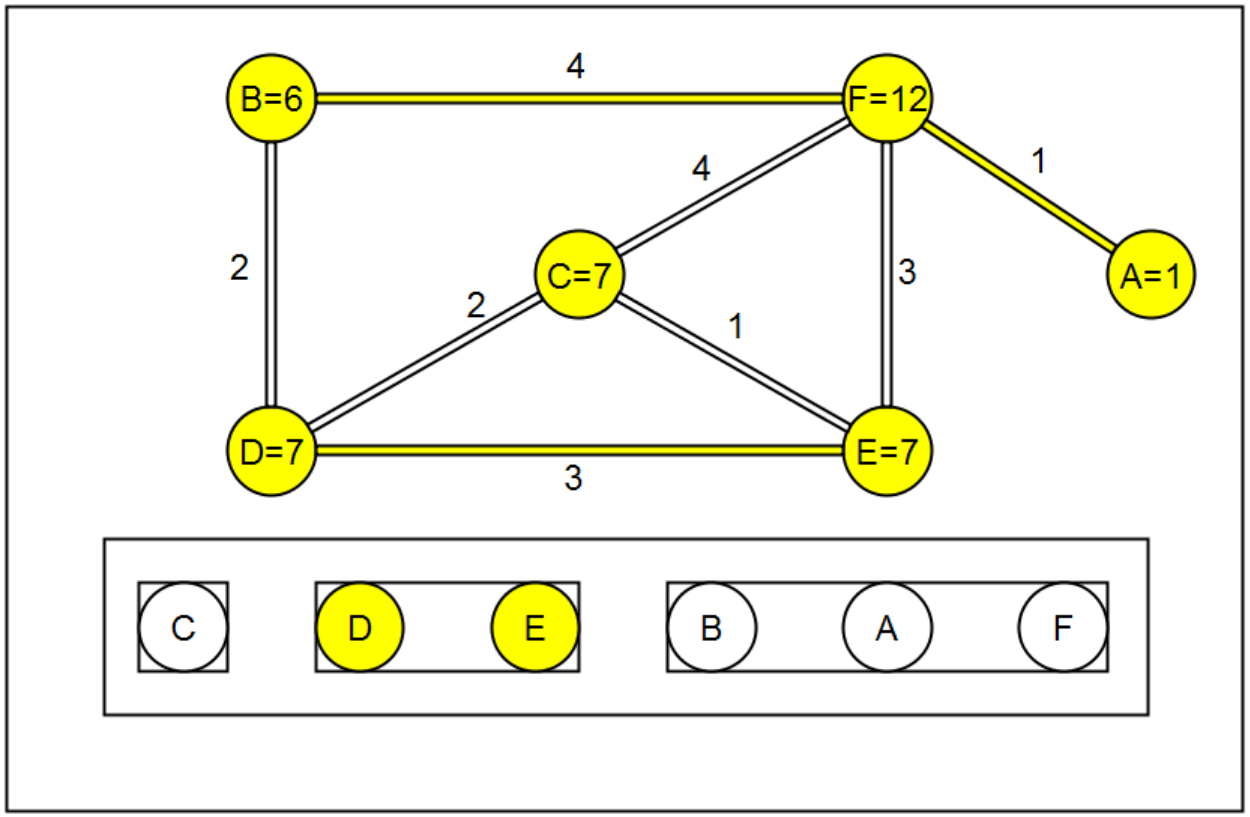

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
B
C
D
E
A
F

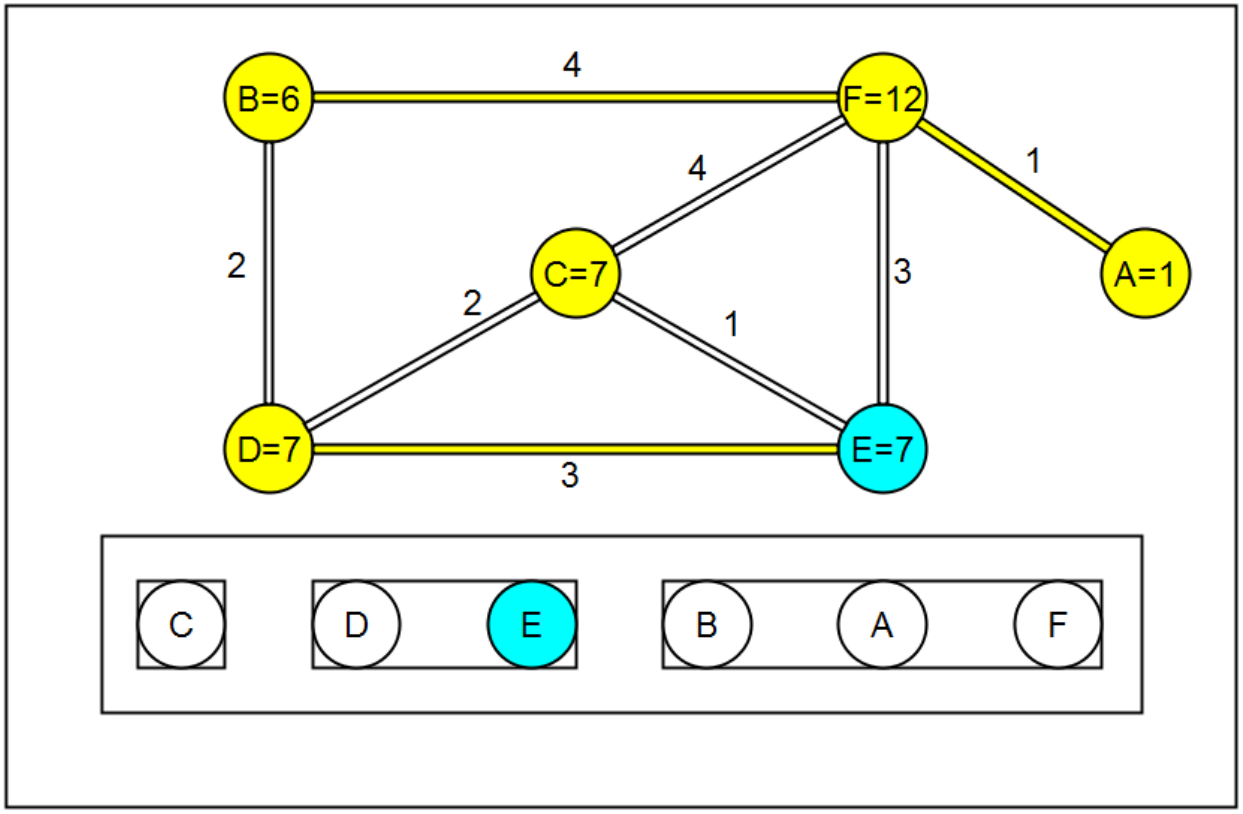

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

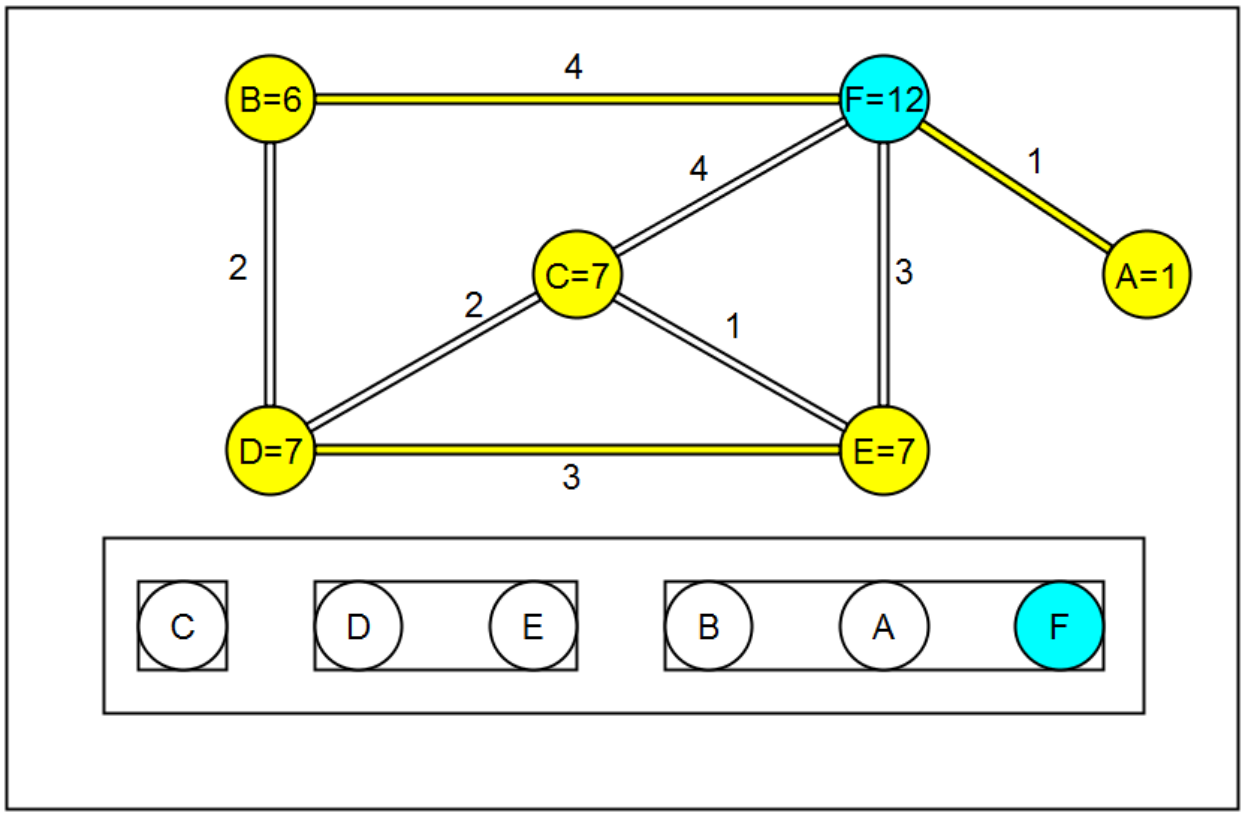

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

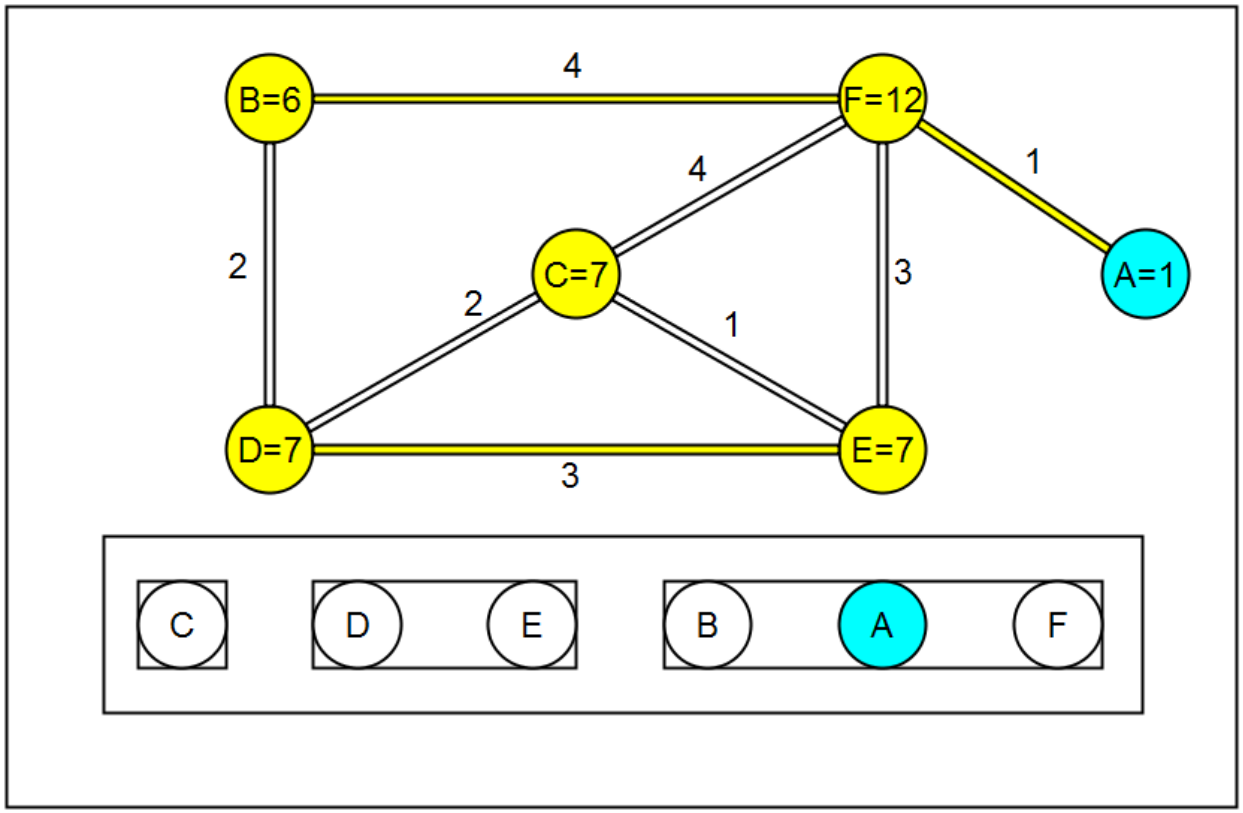

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

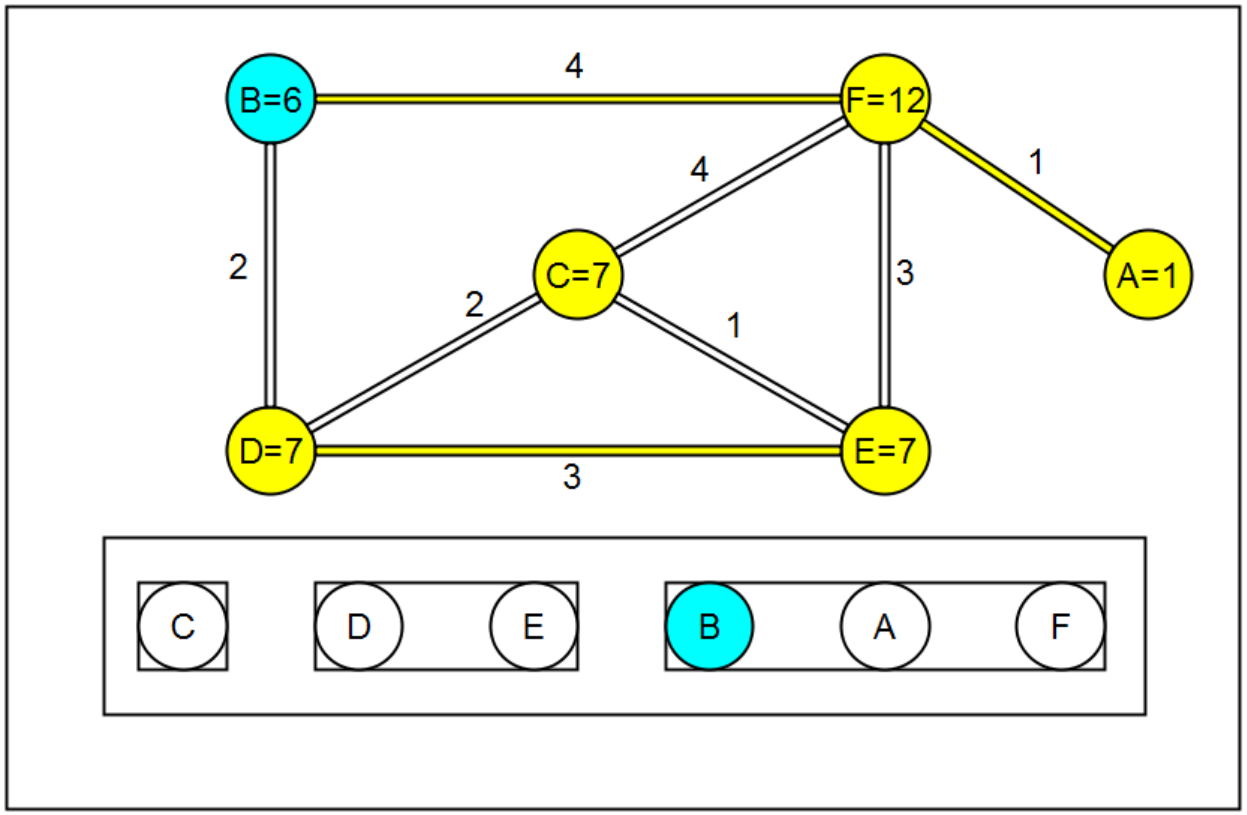

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

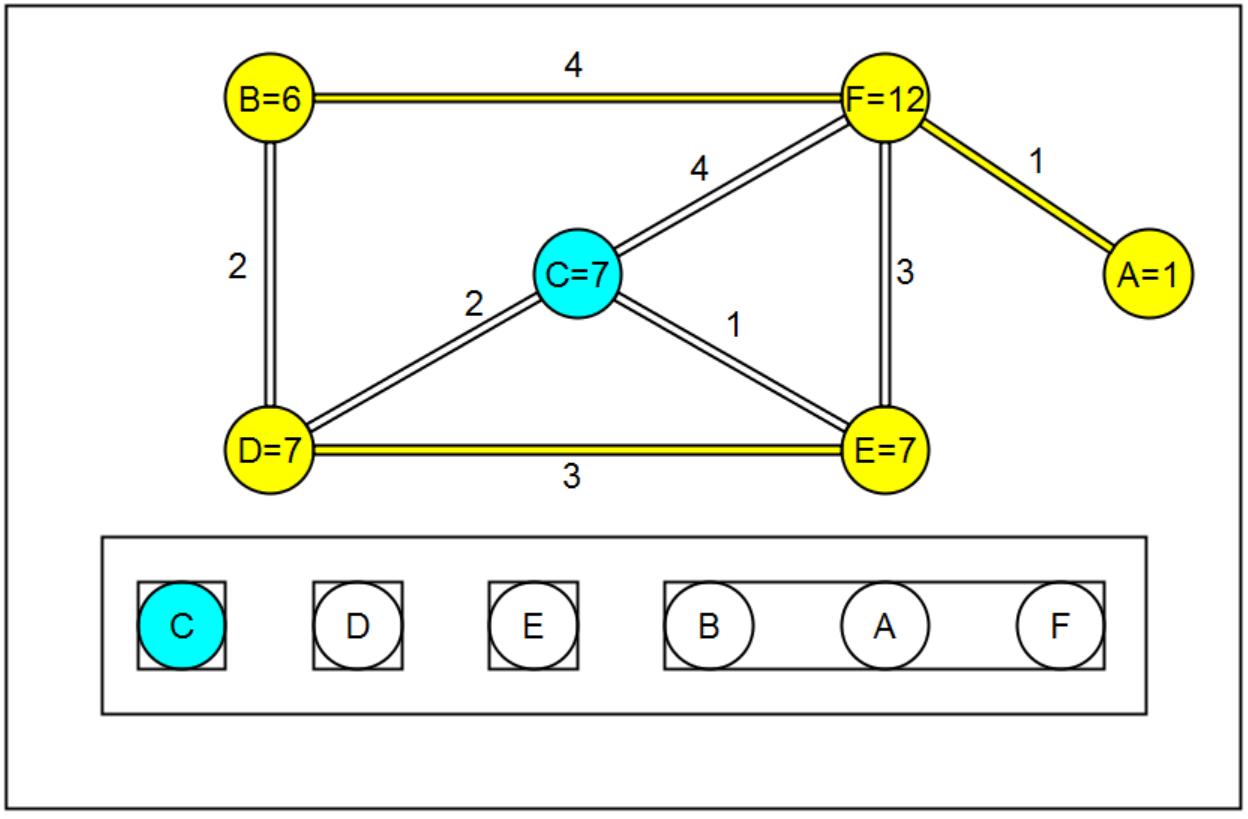

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

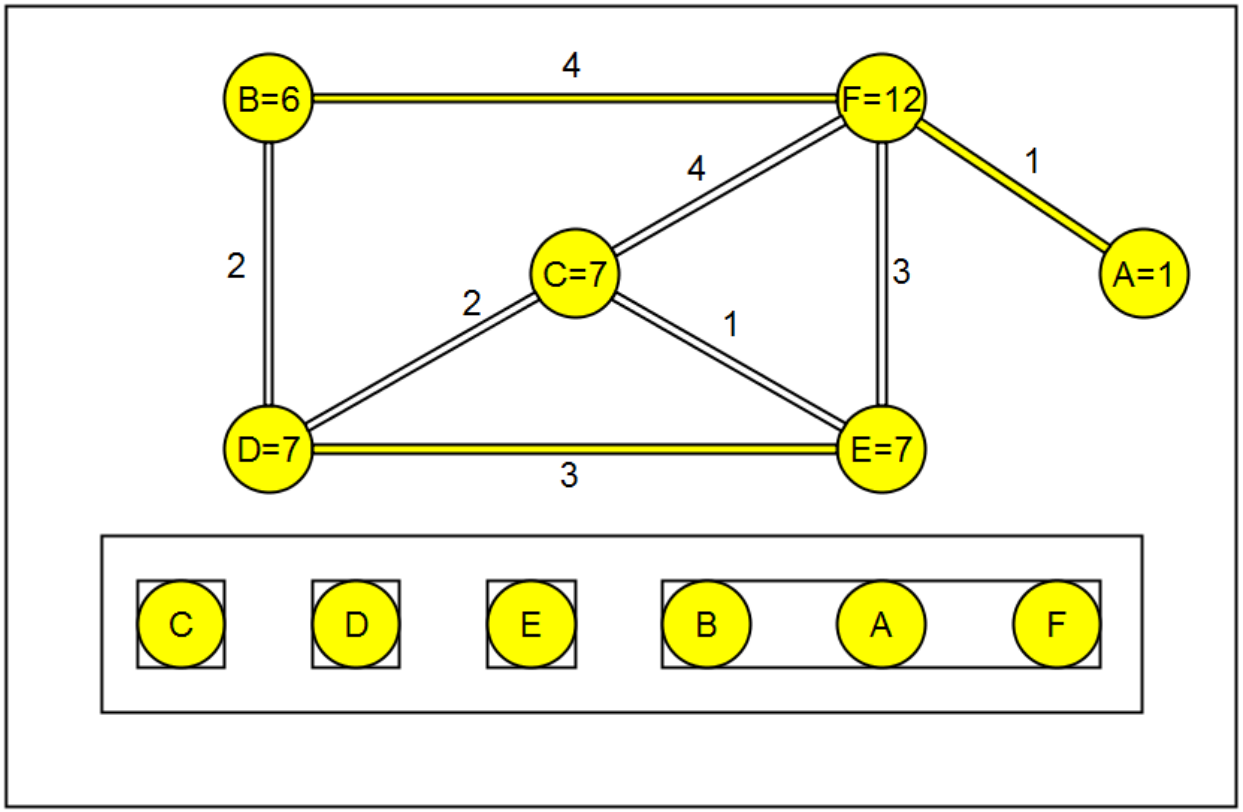

B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F


B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F


B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F


B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F


B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F


B=6
F=12
C=7
A=1
D=7
E=7
4
4
1
2
3
2
1
3
C
D
E
B
A
F

Resultado final da partição :