

Prova 1 - Exercícios práticos

Redes Complexas para Ciência da Computação

Luben Miguel, Rodrigo Lassance

Índice

Exercício 1	1
Item (a)	2
Item (b)	3
Item (c)	3
Item (d)	3
Exercício 2	4
Item (a)	4
Item (b)	6
Item (c)	11
Exercício 3	12
Item (a)	12
Item (b)	13
Exercício 4	13
Item (a)	13
Item (b)	14

Exercício 1

Para esse exercício utilizamos as redes de conexão entre as fibras do cérebro de ratos [bn-mouse_visual-cortex_2](#), rede de co-autoria de artigos da área de redes complexas [ca-netscience](#) e a rede de emails mandados entre funcionários da empresa de energia Enron que deixou de existir em 2007 [email-enron-only](#).

Selecione essas redes e utilizamos como métodos de detecção de comunidades o método de Louven, Girvan-Newman e propagação de rótulos.

Item (a)

Selecionamos a rede de coautoria para visualização de suas respectivas comunidades estimadas em cada método como pode ser visualizado pela Figura abaixo:

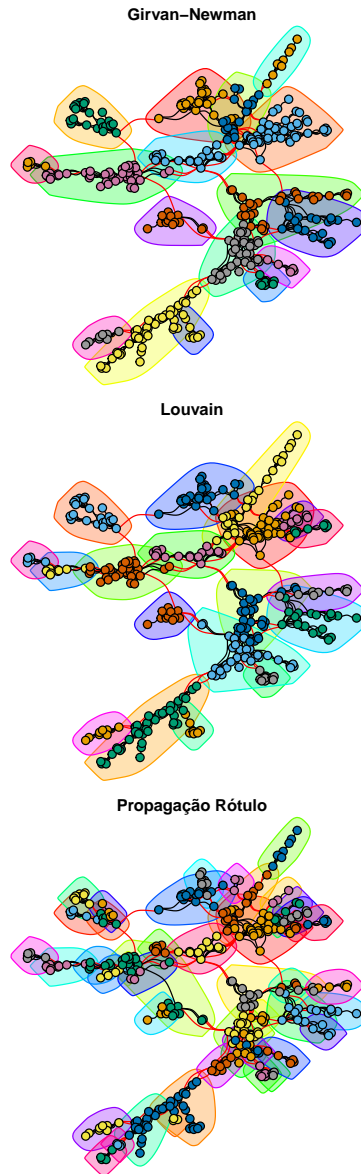


Figura 1: Comunidades detectadas para a rede de co-autoria para cada método

Item (b)

A tabela a seguir sumariza o número de comunidades encontrados para cada rede (nas linhas) por cada método (nas colunas):

Tabela 1: Número de comunidades para cada rede usando cada método

Redes	Girvan-Newman	Louvain	Propagação de Rótulo
Rede do cortex dos ratos	10	9	10
Rede de emails	9	7	4
Rede de co-autoria	17	19	46

Item (c)

A seguir, assim como no item (b), construímos uma tabela com os valores calculados de modularidade de cada método (nas colunas) para cada rede (nas linhas).

Tabela 2: Modularidade para cada rede usando cada método

Redes	Girvan-Newman	Louvain	Propagação de Rótulo
Rede do cortex dos ratos	0.7499236	0.7523037	0.7385689
Rede de emails	0.5345388	0.5663787	0.4123642
Rede de co-autoria	0.8421850	0.8472006	0.7742825

Conforme visto em aula, a modularidade tende a ser maior em redes maiores. Ou seja, não sabemos se a modularidade da rede de co-autoria, com 379 vértices, é maior que a rede de ratos, com 193 vértices, por estar melhor separada ou por simplesmente ser uma rede maior.

Item (d)

Para cada rede utilizada, podemos calcular a similaridade entre as partições encontradas por cada algoritmo através do NMI. Assim, computando o NMI entre os diferentes métodos de detecção de comunidades obtemos os resultados mostrados pela Tabela abaixo:

Nota-se primeiramente que para a rede de emails, o valor do NMI calculado na comparação entre os métodos Girvan-Newman e Louvain versus a propagação de rótulos é razoavelmente baixo. Isso pode se dever ao fato de que no método de propagação de rótulos, o número resultante de comunidades é pequeno comparado aos outros dois métodos. Já, para a rede de co-autoria, esperavamos que o valor do NMI fosse maior nessas mesmas comparações devido à quantidade grande de partições obtidas pela propagação de rótulo, mas não foi isso o observado, vide os valores modestos de NMI sumarizados na tabela.

Tabela 3: NMI associado a cada combinação dois a dois dos métodos para cada rede

Comparação	Rede de ratos	Rede de emails	Rede de coautoria
Givarn-Newman vs Louvain	0.9488052	0.7754347	0.8683370
Girvan-Newman vs Label Prop	0.8762523	0.4213453	0.6952310
Louvain vs Label Prop	0.8782549	0.4315381	0.7260035

Exercício 2

Item (a)

A seguir, podemos visualizar o scatter-plot entre cada par de medida para cada rede a seguir:
- Rede de conexão de fibras do cortex de ratos

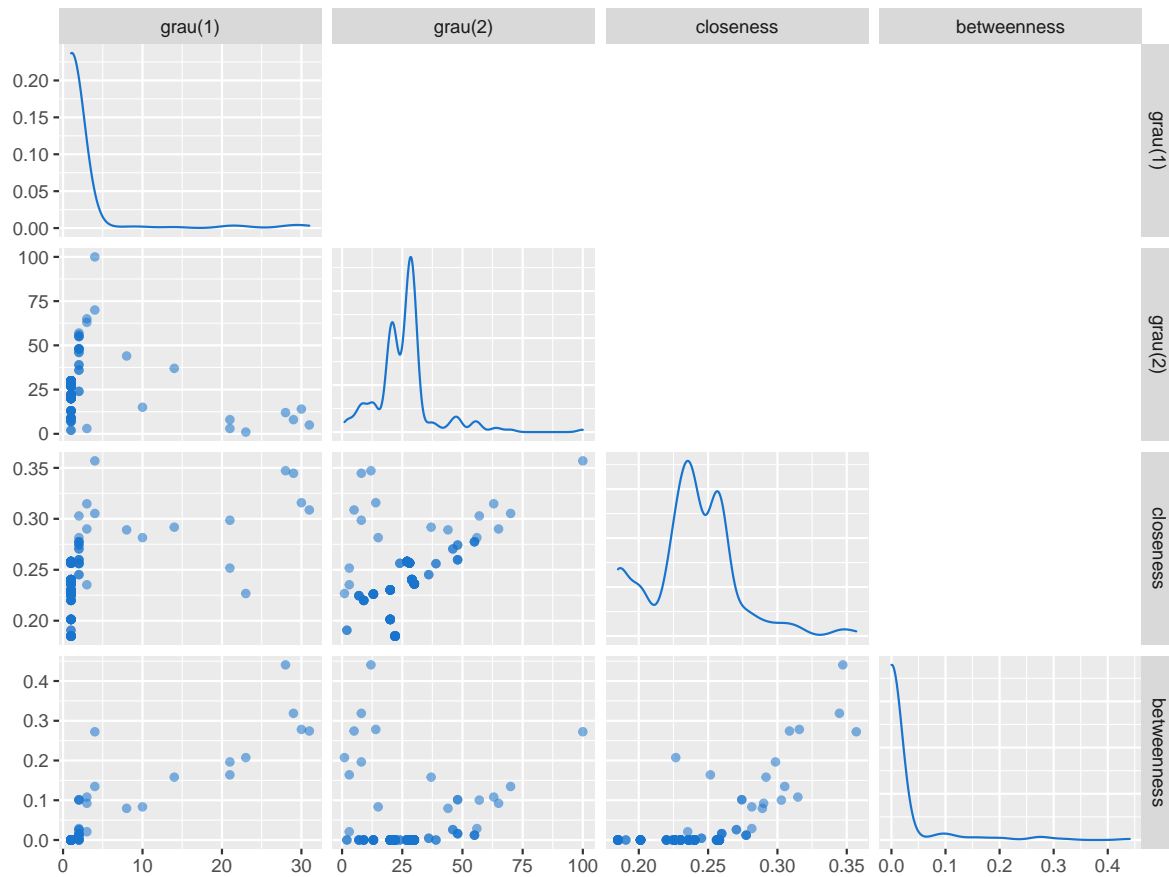


Figura 2: Gráficos de dispersão das medidas de centralidade para a rede de cortex de ratos

- Rede de emails mandados entre funcionários da empresa Enron

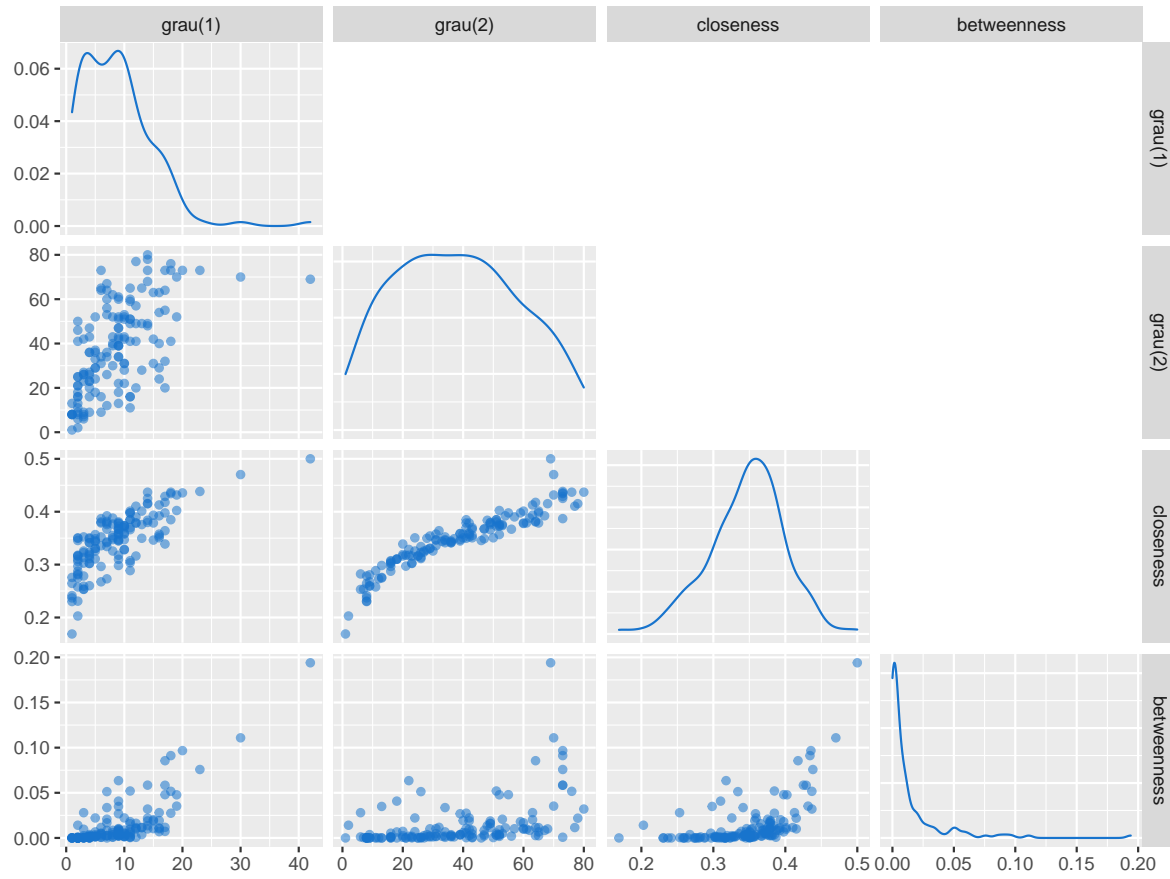


Figura 3: Gráficos de dispersão das medidas de centralidade para a rede de emails

- Rede de co-autoria de artigos

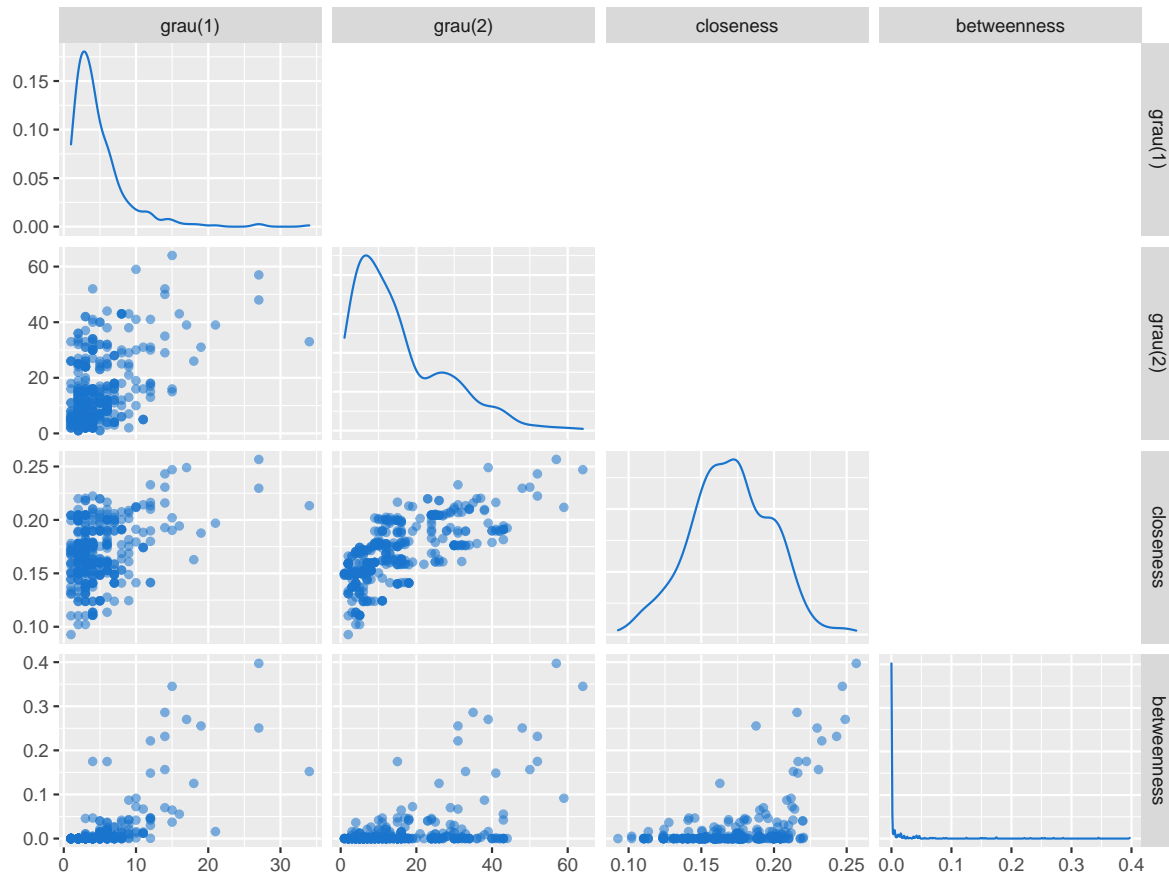


Figura 4: Gráficos de dispersão das medidas de centralidade para a rede de coautoria

Item (b)

- Rede de conexão de fibras do cortex de ratos

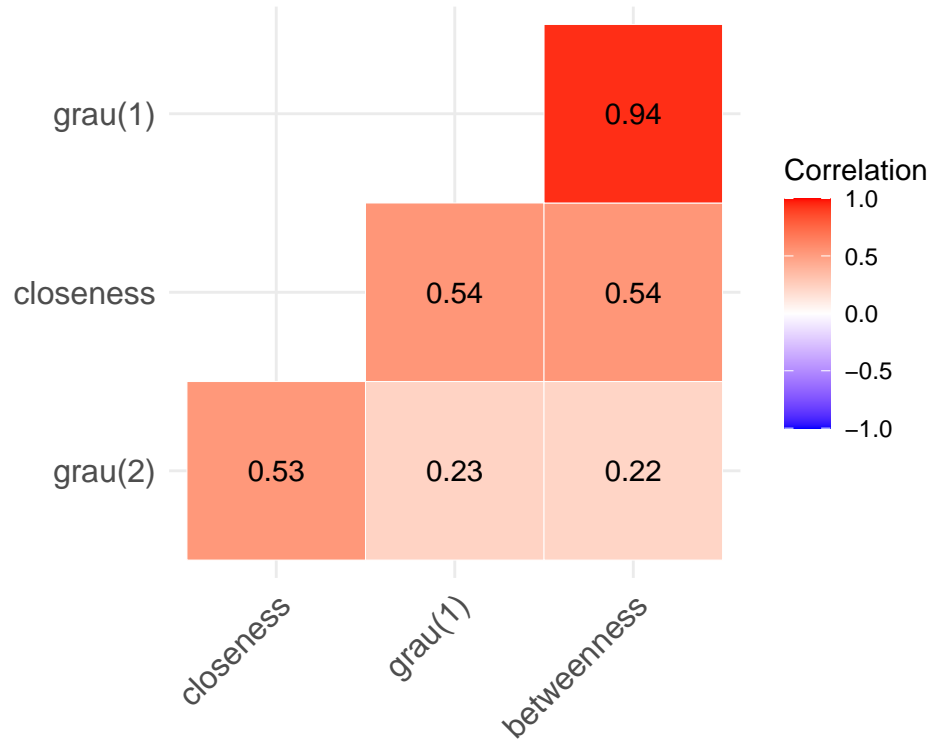


Figura 5: Matriz de correlação das medidas de centralidade para a rede de cortex de ratos

Nota-se pela Figura 5 uma alta correlação entre o betweenness e o grau hierárquico de grau 1, muito próxima de 1. Isso provavelmente se deve ao fato de que, nesse caso, vértices com alto grau tem um número elevado de vizinhos que só se ligam a ele. Desse modo, ao remover esses vértices da rede, torna-se impossível acessar esses outros nós de grau menor. De fato, esses nós menores todos possuem em torno de grau 1 ou 2, como se pode visualizar pelos gráficos de dispersão e densidade associadas ao grau hierárquico de grau 1 observados na Figura 2. Em uma interpretação de comunidades, pode-se dizer que esses nós de graus maiores de cada comunidade, são aqueles também conectam sua comunidade a outras.

Também visualiza-se na Figura que o grau hierárquico de grau 2 possui baixa correlação com o grau hierárquico de grau 1 e o betweenness. Isso ainda se relaciona com a discussão acima, visto que tanto os nós ligados a outros vértices com alto grau e betweenness quanto estes mesmos vértices vão se conectar em segundo grau com nós dotados de vizinhanças grandes. Isso pode também ser observado nos gráficos de dispersão e densidade associada ao grau hierárquico de ordem 2, que em geral, tem grande concentração de valores grandes. Essas constatações da matriz de correlação podem ser ilustradas pelo formato da rede do cortex visualizada por meio da Figura 6

Rede do cortex visual de rato

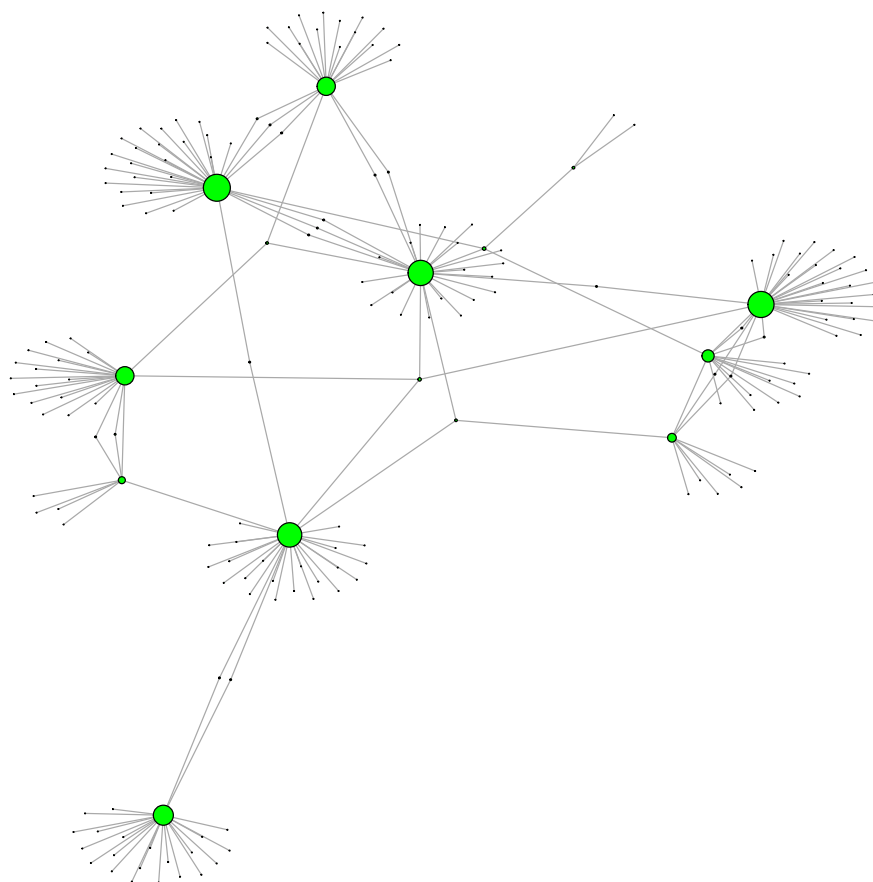


Figura 6: Rede do cortex visual de rato

- Rede de emails mandados entre funcionários da empresa Enron

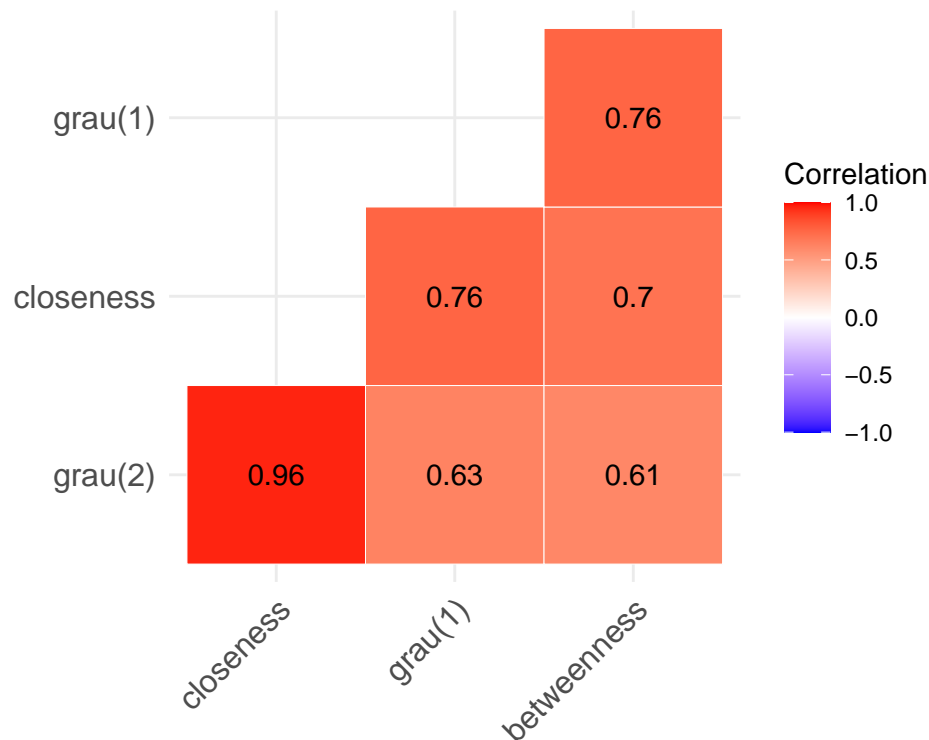


Figura 7: Matriz de correlação das medidas de centralidade para a rede de emails

Primeiramente, constata-se pela Figura 7 acima que todas as medidas de centralidade possuem correlação alta, acima de 0.6, o que pode estar relacionado a rede ter um número elevado de arestas e cada vértice estar bem conectado nesta. Isso é ilustrado pelos gráficos de dispersão e densidades do grau de ordem 1 presentes na Figura 3, em que se observa uma boa dispersão dos graus em maiores valores, com maiores densidades no intervalo de 0 a 10. Destaca-se também a alta correlação entre o grau hierárquico de ordem 2 e o closeness que podem estar relacionados ao fato de que, quando se tem um grau hierárquico de ordem 2 alto, também teremos um tamanho médio de caminhos menores, ou seja um alto closeness.

Isso implica que muitos nós nessa rede podem ser alcançados em até dois passos, o que faz sentido por ser uma rede de emails institucional, havendo uma noção de hierarquia associada. Já, para nós mais periféricos, em que não temos muitas conexões de grau 2, é natural que o closeness seja menor, visto que serão necessários mais passos para chegar nesse vértice a partir de todos os outros. No contexto de emails, pode-se dizer que tais usuários fazem parte de grupos mais separados na empresa.

- Rede de co-autoria de artigos

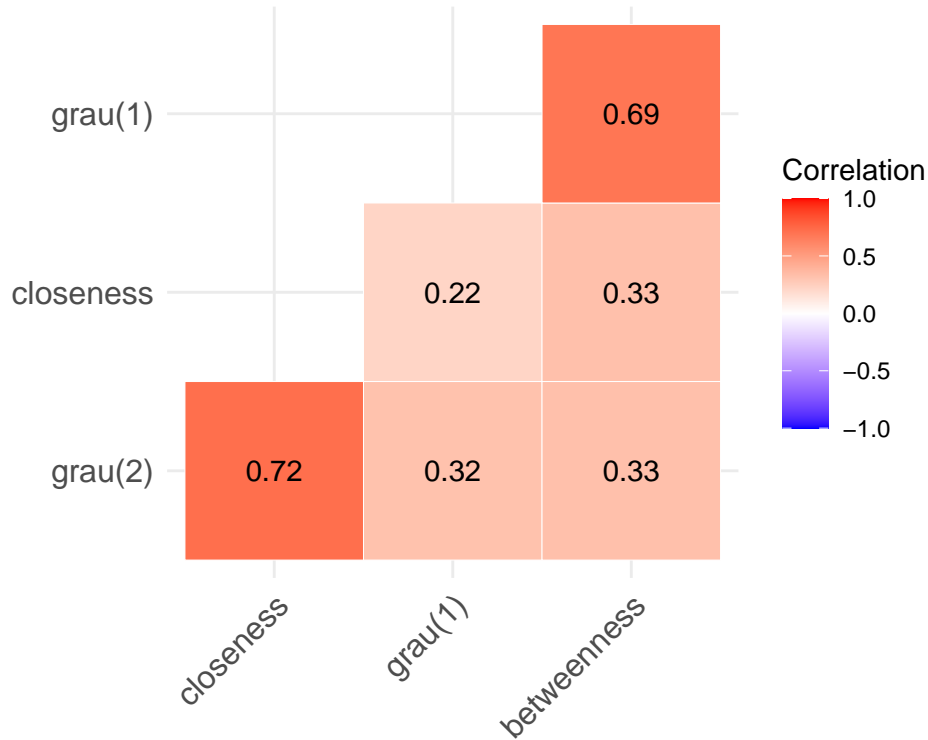


Figura 8: Matriz de correlação das medidas de centralidade para a rede de coautoria

Pela Figura 8, nota-se que tanto o grau hierárquico de ordem 1 tem uma alta correlação com o betweenness quanto também o grau hierárquico de ordem 2 tem alta correlação com o closeness, enquanto o restante dos pares de medidas possuem correlações razoavelmente baixas. Nota-se pela Figura 4 que para essa rede, a distribuição do grau hierárquico de ordem 1 tende a ter maiores concentrações em menores valores, com um pico em sua densidade para valores muito menores que 10. Ou seja, ao contrário da rede de emails, cada vértice tende a ter um número mais seleto de vizinhos. Nesse caso da rede de coautoria, a relação positiva entre grau de ordem 1 e betweenness pode estar relacionada ao fato de que, vértices com uma vizinhança grande representam autores que tendem a trabalhar junto com diversos outros autores, sejam eles autores de outros enfoques, autores com mesmo enfoque e possivelmente orientandos que por sua vez co-autoram com poucos ou nenhum outro pesquisador. Assim, se removermos os “orientadores”, muito provavelmente os “orientandos” e outros pesquisadores perdem qualquer relação com vários outros pesquisadores presentes na rede. Nesse sentido, esses casos remetem ao que se visualizou na rede de Córtex visual dos ratos, porém em um caso menos extremo.

Por sua vez, a alta correlação entre o closeness e o grau de ordem 2 se deve ao fato de que, pesquisadores influentes que co-autoram com outros pesquisadores também influentes tendem a não se distanciar muito de todos os demais pesquisadores, ou seja, são facilmente acessados por qualquer outro vértice da rede. Já, pesquisadores com uma vizinhança de grau 2

menor coautoram com outros pesquisadores que possuem vizinhanças pequenas, ou seja, seriam pesquisadores de enfoques mais específicos e isolados, e portanto, estariam mais distante dos demais. Isso reflete o comportamento hierarquico observado na rede emails, porém de forma menos rígida e sem refletir uma noção de superioridade, e sim uma organização interna da área de pesquisa.

Item (c)

Tomando a rede de emails e as comunidades obtidas pelo método de Louvain, podemos realizar um ranqueamento dos vértices para as maiores comunidades. Primeiramente, a tabela abaixo nos dá o tamanho de cada comunidade:

Tabela 4: Tamanho de cada comunidade

Comunidade	Tamanho
1	29
2	18
3	21
4	26
5	22
6	15
7	12

Visualiza-se que as comunidades 1,3,4 e 5 tem mais que 20 vértices membros. Assim, tomando a comunidade 3, que possui um tamanho grande, podemos tomar o subgrafo correspondente a essa comunidade e computar as medidas de centralidade locais como dada pela Tabela 5:

Tabela 5: Top 10 vértices com maior centralidade local para o subgrafo da comunidade 3

vertices	grau(1)	grau(2)	closeness	betweenness
75	14	6	0.769	0.277
47	11	6	0.625	0.054
9	10	10	0.667	0.035
13	10	10	0.667	0.033
32	10	10	0.667	0.067
49	10	10	0.667	0.028
56	10	10	0.667	0.040
66	10	7	0.606	0.032
74	10	10	0.667	0.079
137	9	11	0.645	0.072

Pela Tabela 5, destaca-se o vértice 75 como o mais central na comunidade 3, tendo maiores valores de grau, closeness e betweenness. Já, para a centralidade global, podemos observar os top 10 vértices mais centrais de forma geral pela Tabela 6:

Tabela 6: Top 10 vértices com maior centralidade global

vertices	grau(1)	grau(2)	closeness	betweenness
105	42	69	0.500	0.194
17	30	70	0.470	0.111
95	23	73	0.438	0.076
48	20	73	0.436	0.097
43	19	70	0.432	0.035
132	19	52	0.402	0.048
31	18	41	0.385	0.027
74	18	76	0.437	0.052
91	18	73	0.434	0.091
20	17	20	0.339	0.007

Vemos que o vértice 75 nem sequer consta entre os 10 vértices mais globalmente centrais. De fato, outros vértices mais relevantes para a comunidade 3 como o 47, 9, 13 e 32 também não aparecem entre os top 10 vértices mais centrais. Ou seja, pode-se afirmar que existem vértices altamente relevantes dentro da comunidade de interesse sem um grande destaque global.

Exercício 3

Item (a)

A seguir, resumimos em uma tabela a taxa de correspondência para cada rede tomando $B = 100$ como o número de repetições do processo descrito no enunciado:

Tabela 7: Valor médio da taxa de correspondência de cada Rede selecionada

Redes	Correspondência
Rede de co-autores	0.0113536
Rede de cortex do rato	0.0056366
Rede de emails	0.3187234

Item (b)

Antes de pensar em como adaptar a ideia de previsão de links para o contexto de redes complexas com comunidades, convém identificar qual seria a sua real serventia nesse tipo de contexto. Já havendo uma noção bem estabelecida de quais vértices fariam parte de quais comunidades, uma pergunta natural seria se essas comunidades são entidades separadas, isto é, se podem ser avaliadas individualmente. Em um contexto assim, a previsão de link poderia ser utilizada para avaliar a real pertinência das arestas que se situam entre comunidades. Assim, caso haja indícios de que essas arestas podem ser removidas, significa que estaríamos lidando com entidades de naturezas distintas.

Com essa ideia em mente, uma proposta de adaptação do método que seria mais eficiente e que levaria em consideração a existência das comunidades seria realizar a previsão de links apenas para arestas que figuram entre comunidades. Desse modo, reduziríamos consideravelmente o escopo do problema, sem com isso serem necessárias grandes alterações na metodologia.

Exercício 4

Item (a)

Pode-se dizer que a afirmação erra ao equivaler a abordagem de particionamento de redes com a de detecção de comunidades. Apesar de o método de particionamento da rede poder ser usado de forma a auxiliar ou aprimorar o processo de identificação de comunidades, ele não fornece critérios adequados para estabelecer se as partições estão fornecendo potenciais comunidades.

Como exemplo dessa falha, apresentamos a Figura abaixo. Os pontos em preto representam vértices, as linhas azuis são arestas e as linhas tracejadas são duas possíveis partições ótimas. Podemos observar que a partição em verde geraria dois grupos (de cima e de baixo) que apresentam potencial de serem comunidades. Porém, a partição gerada em vermelho - que oferece o mesmo *cut size* ótimo da partição em verde - fornece um grupo (à esquerda) que passa a impressão de possuir duas comunidades em seu interior. Afinal, há apenas uma aresta no meio ligando duas regiões bem conectadas.

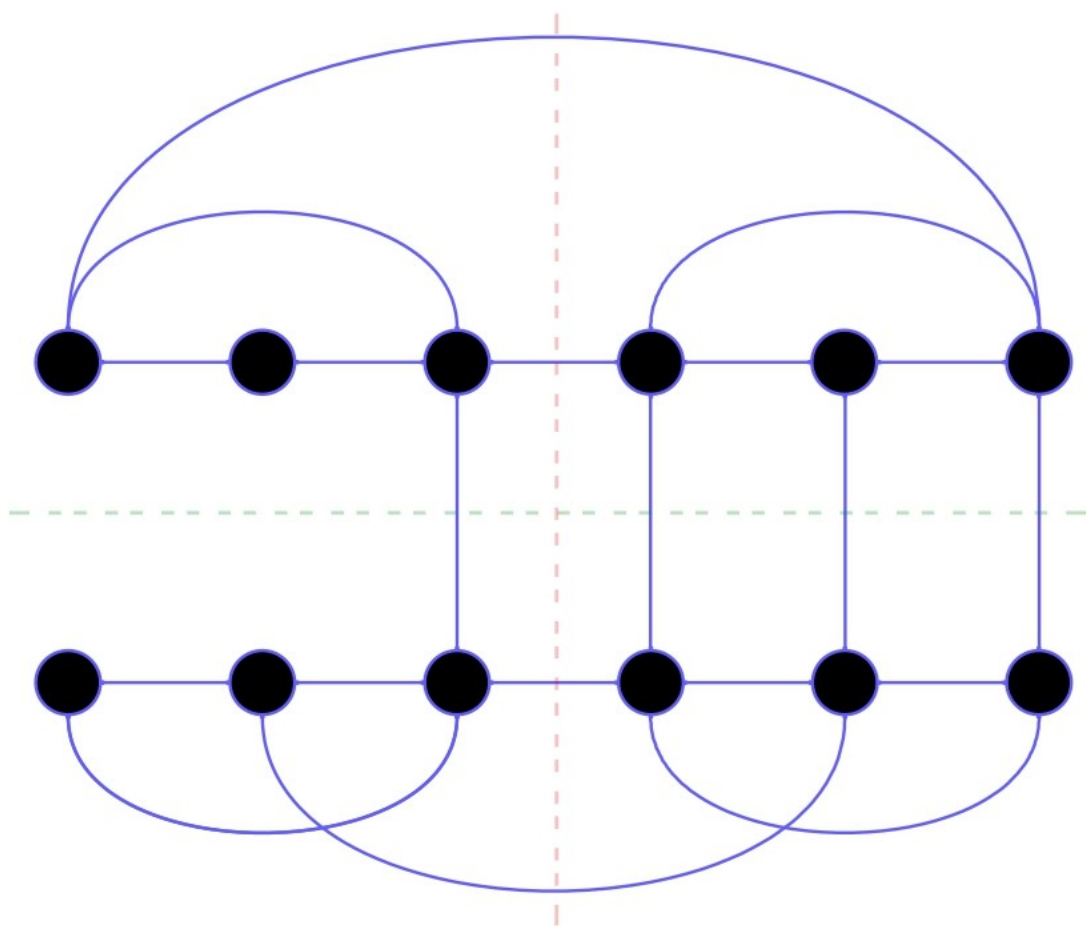


Figura 9: Exemplo de rede.

Item (b)

Ilustra-se a melhor bi-seção por kernighan-lin da rede Zachary's karate club pela Figura 10.

Partição por Kernighan–Lin da rede Zachary's karate club

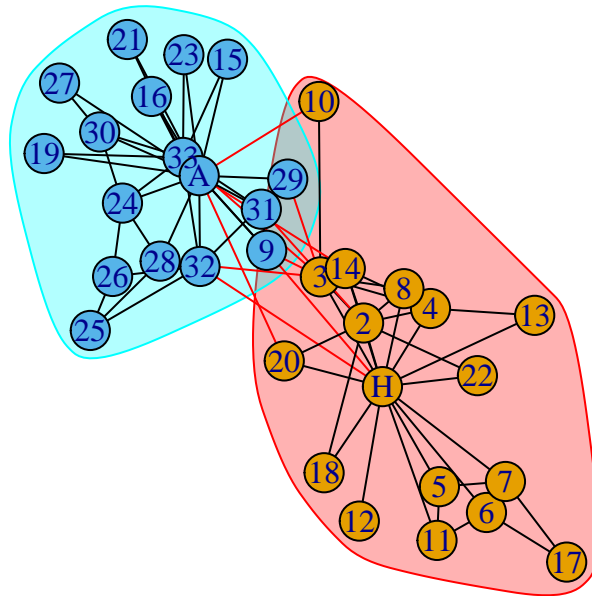


Figura 10: Bi-seccção obtida por Kernighan-Lin

Para comparar com a rede original, visualiza-se a seguir a Figura 11.

Rede original Zachary's karate club

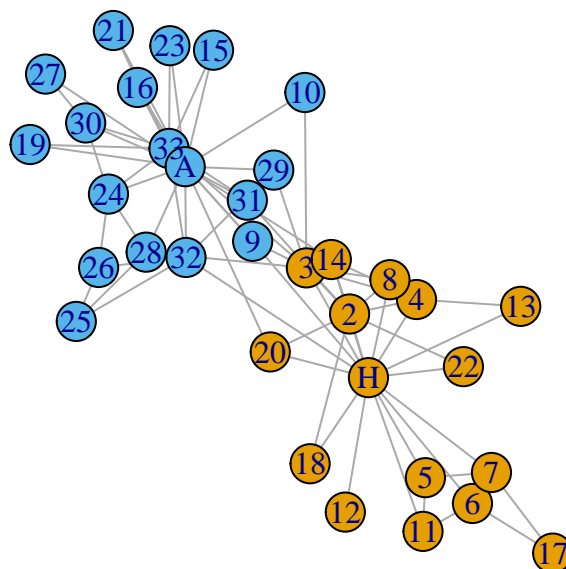


Figura 11: Rede original do Zachary's karate club

Nota-se em geral, pelas duas figuras, que o método de Kernighan-Lin nos dá uma bisecção praticamente idêntica aos dois grupos originais, tendo como única diferença o vértice 10, que está no grupo amarelo ao invés do azul na bisecção. Ou seja, neste caso, podemos dizer que o método da bisecção teve um bom funcionamento, praticamente nos retornando os grupos originais