Clusterização da Estrutura de Pré-Requisitos de Disciplinas na Unicamp

Autor: Lucas Perondi Kist (RA: 236202)

1 Introdução

Diversas disciplinas são oferecidas para os cursos de graduação a cada semestre na Universidade Estadual de Campinas (Unicamp). Nesse sentido, algumas delas só podem ser cursadas após certos pré-requisitos terem sido cumpridos, os quais são classificados como plenos (aprovação em outra disciplina), parciais (frequência e média mínimas em outra disciplina) ou especiais (Coeficiente de Progressão mínimo ou autorização da Coordenação).

Assim, como cada matéria é identificada por um código XXYYY, onde XX é a sigla do departamento e YYY são três dígitos, é possível construir uma relação de associações entre departamentos a partir da análise das matérias que devem ser cursadas antes de outra. Dessa forma, o objetivo deste trabalho é não só realizar essa construção, mas também aplicar métodos de *clusterização* para encontrar agrupamentos nessa rede. Os códigos utilizados para a elaboração deste trabalho estão disponíveis em: `https://github.com/lpkist/Trabalho3_ME921/`.

2 Materiais e Métodos

As análises foram realizadas na linguagem de programação R (R Core Team (2018)). Inicialmente, foi realizado *webscraping*, utilizando os pacotes *rvest* e *xml2*, das páginas com as disciplinas ofertadas por cada departamento, acessadas a partir deste link. Na sequência, para cada matéria, foram extraídos os códigos dos pré-requisitos e, destes, obtiveram-se as respectivas siglas do departamento.

Na sequência, foi construída a matriz de adjacências entre os departamentos $\boldsymbol{A}$, com $a_{i_1 i_2} = 1$ se $i_1 \neq i_2$ e o departamento i_1 é pré-requisito para alguma disciplina do departamento i_2 e 0 caso contrário, para todos os pares (i_1, i_2). Ademais, foram removidos da análise aqueles que não tivessem relação de pré-requisitos com outro departamento, isto é, i tal que $\sum_{k=1}^{n} a_{ki} = 0$ e $\sum_{k=1}^{n} a_{ik} = 0$, resultando em 82 departamentos e uma matriz de adjacências direcionada. Para a *clusterização*, foram utilizados *Stochastic Block Model* (pacote *mixer*) e *Regularized Spectral Clustering* (pacote *randnet*)

2.1 Stochastic Block Model (SBM)

Conforme apresentado por Bouveyron et al. (2019), assume-se que há uma rede com n nós, com uma matriz de adjacências $\boldsymbol{A}_{n\times n}$, G blocos na população, que a probabilidade de um nó provir do bloco g é τ_g e que existe uma matriz de interação entre os blocos $\boldsymbol{\Theta}_{G\times G}$ com a probabilidade de que um nó do bloco g se ligue a outro do bloco h no elemento θ_{gh}. Assim, definem-se o vetor de probabilidades de pertencimento aos blocos $\boldsymbol{\tau} = (\tau_1, ..., \tau_G)'$, o vetor que indica o bloco a que o i-ésimo nó pertence $\boldsymbol{z}_i = (z_1, ..., z_G)'$, com $z_{ig} = 1$ se pertence ao bloco j e $z_{ig} = 0$ caso contrário e $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$.

A partir disso, a verossimilhança completa é dada por:

$$L_c(\boldsymbol{\tau}, \boldsymbol{\Theta}) = P(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{\Theta})P(\boldsymbol{Z}|\boldsymbol{\tau}) = \prod_{i\neq j} (\boldsymbol{z}_i' \boldsymbol{\Theta} \boldsymbol{z}_j)^{a_{ij}} (1 - \boldsymbol{z}_i' \boldsymbol{\Theta} \boldsymbol{z}_j)^{1-a_{ij}} \prod_{i=1}^{n} \prod_{g=1}^{G} \tau_g^{z_{ig}},$$

que é computacionalmente custosa. Por isso, foi utilizada uma abordagem bayesiana proposta por Daudin et al. (2008), que atribui as seguintes prioris: $\boldsymbol{\tau} \sim Dirichlet(\delta)$ e $\theta_{gh} \sim beta(\alpha, \beta)$. Dessa forma, é possível obter as probabilidades *a posteriori* de cada observação pertencer a cada grupo, bem como calcular a incerteza dessa atribuição e o *Integrated Completed Likelihood* (ICL).

2.2 Regularized Spectral Clustering (RSC)

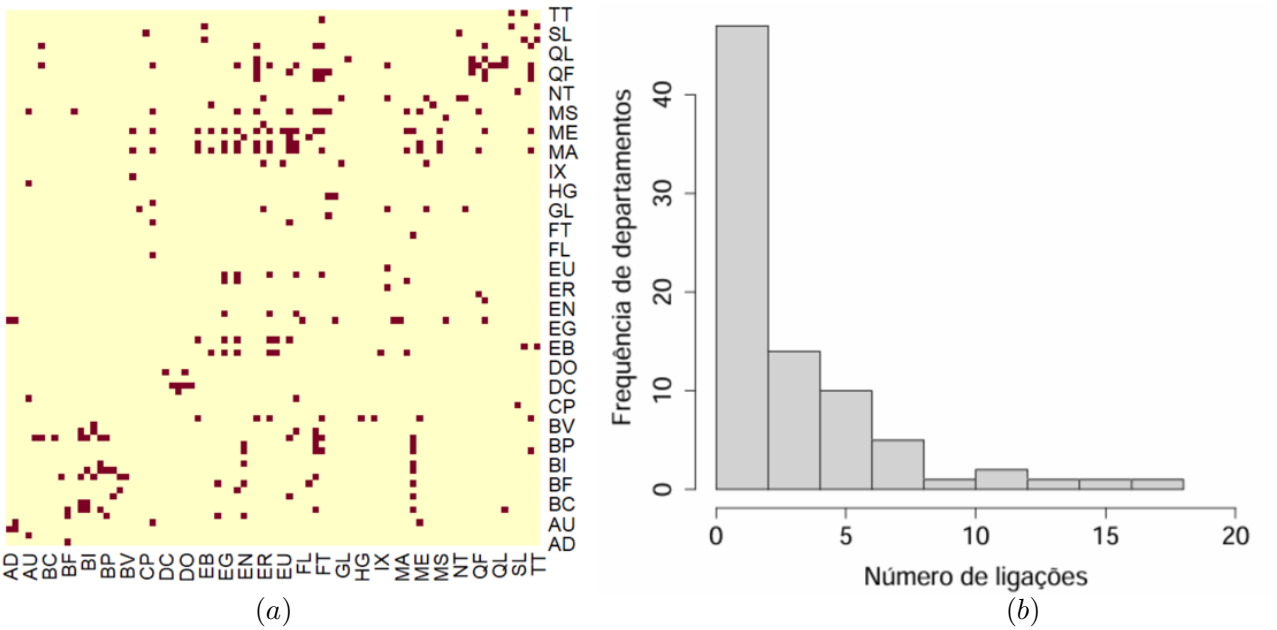
Já o *Regularized Spectral clustering*, ou *clustering* espectral regularizado, segundo apresentado por Qin e Rohe (2013), se baseia no Laplaciano do grafo regularizado, dado por $\boldsymbol{L}_\tau = \boldsymbol{D}_\tau^{-1/2} \boldsymbol{A} \boldsymbol{D}_\tau^{-1/2}$, onde $\boldsymbol{D}_{n\times n}$ é uma matriz diagonal tal que $D_{ii} = \sum_j A_{ij}$, $\boldsymbol{D}_\tau = \boldsymbol{D} + \tau \boldsymbol{I}$ e $\tau \geq 0$ é o parâmetro de regularização. Com base nele, cria-se a matriz $\boldsymbol{X}$, cujas colunas são os K autovetores de $\boldsymbol{L}_\tau$ associados aos K maiores autovalores, normaliza-se suas linhas e, considerando cada linha um ponto, aplica-se o método $K - means$ com K *clusters*. Assim, o vértice i é atribuído ao agrupamento k se a i-ésima linha de $\boldsymbol{X}$ foi atribuída ao agrupamento k.

2.3 K-means

Este método de *clusterização* é baseado em encontrar as K médias que minimizam a dissimilaridade em relação aos elementos de cada grupo. Isto é, dada uma matriz de características $\boldsymbol{X}$, o algoritmo parte de K pontos, conecta-os às observações mais próximas e, para cada agrupamento formado, recalcula a média até que as observações em nenhum *cluster* se alterem. Mais detalhes podem ser encontrados em Brian S. Everitt (2011).

3 Resultados

A Figura 1 apresenta o mapa de calor da matriz de adjacências, bem como a frequência de departamentos de acordo com número de ligações por pré-requisitos. Já a Figura 2 ilustra a disposição dos vértices no grafo, com as siglas associadas a cada um deles.



(a) (b)

Figura 1: (a) Mapa de calor dos pré-requisitos e (b) número de siglas por quantidade de ligações

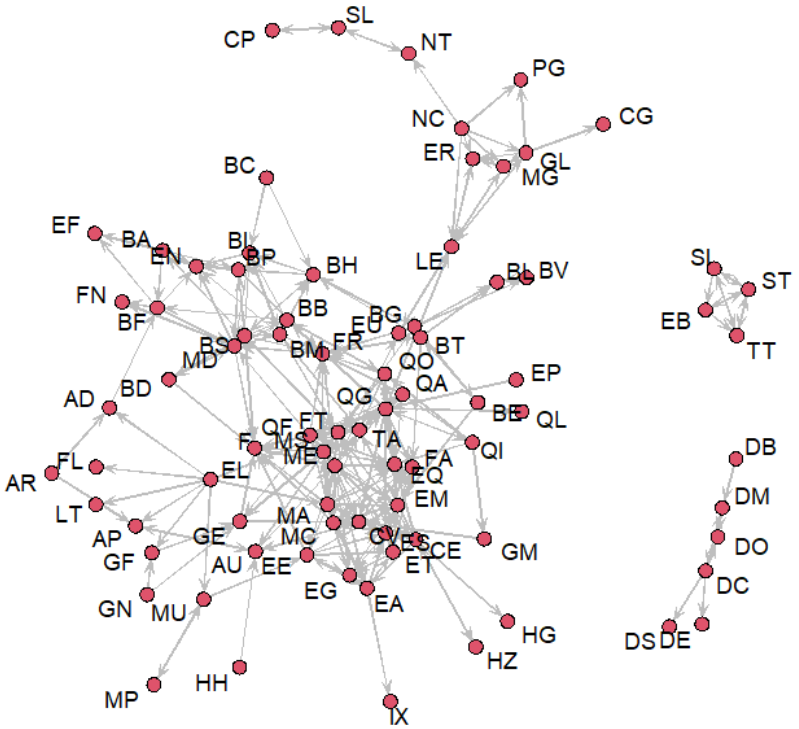


Figura 2: Rede de pré-requisitos entre as siglas das disciplinas

Na Figura 3 é possível comparar o ICL obtido ao aplicar *Stochastic Block Model* para alguns números de agrupamentos. Os três melhores modelos, e respectivos ICLs, foram: 5 (-1779), 4 (-1786) e 7 (-1795). As comparações das incertezas *a posteriori* deles estão na Figura 4. Já o grafo colorido de acordo com os agrupamentos produzidos pelo modelo com 7 *clusters* está representado na Figura 5. O número de elementos em cada grupo, bem como a matriz $\boldsymbol{\Theta}$ estão exibidos, respectivamente, na Tabela 1 e na Tabela 2.

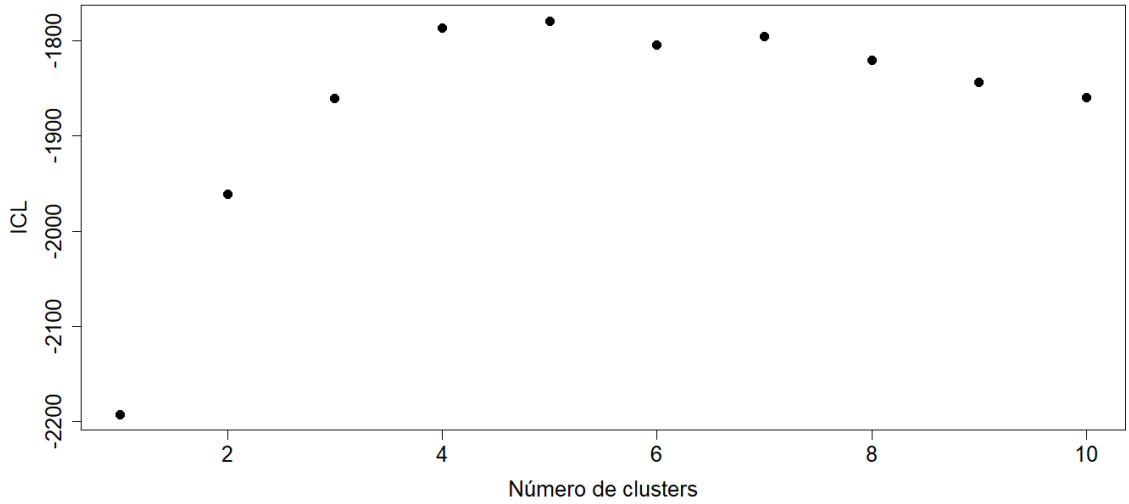


Figura 3: ICL do SBM para alguns números de *clusters*

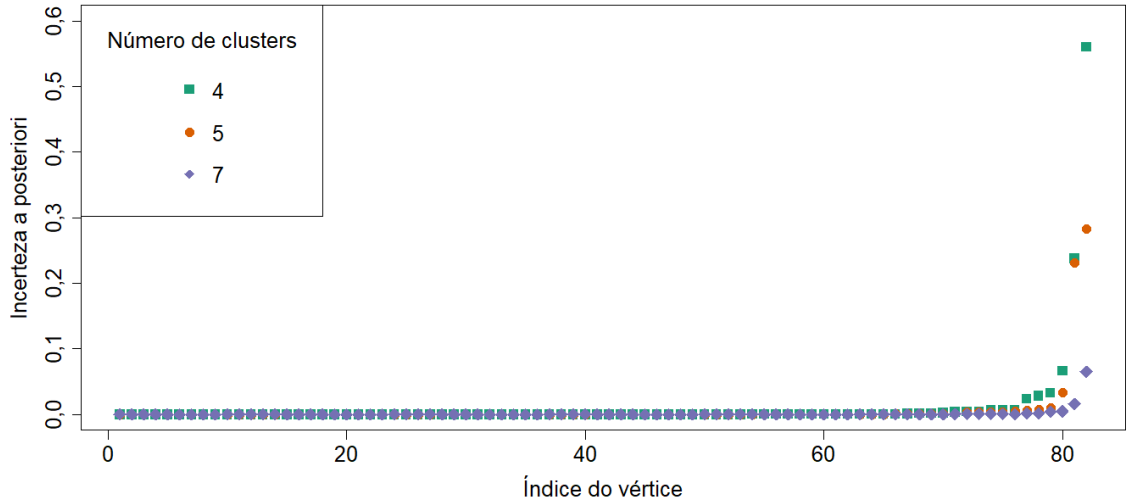


Figura 4: Incertezas *a posteriori* do SBM com os três melhores números de *clusters*

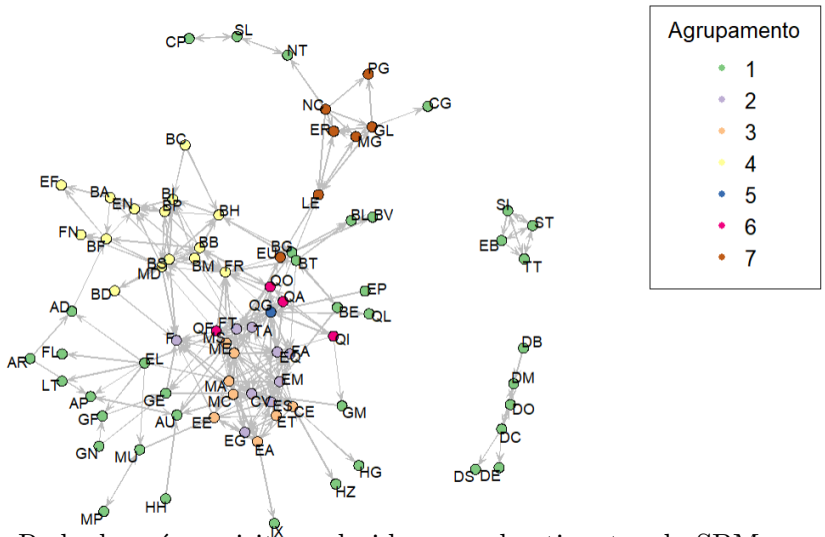


Figura 5: Rede de pré-requisitos colorida segundo etiquetas do SBM com 7 grupos

Tabela 1: Número de elementos em cada agrupamento produzido pelo SBM com 7 grupos

1	2	3	4	5	6	7
45	5	5	14	1	4	8

Tabela 2: Matriz de ligação ($\boldsymbol{\Theta}$) do SBM com 7 *clusters*

Agrupamento	1	2	3	4	5	6	7
1	0,03	0,00	0,03	0,00	0,04	0,01	0,01
2	0,01	0,05	0,72	0,14	0,40	0,70	0,05
3	0,01	0,00	0,60	0,00	0,00	0,00	0,05
4	0,00	0,01	0,01	0,22	0,07	0,02	0,00
5	0,04	0,00	0,20	0,00	0,12	1,00	0,00
6	0,00	0,05	0,10	0,02	1,00	0,25	0,00
7	0,02	0,00	0,57	0,00	0,48	0,00	0,36

Já a soma de quadrados entre os grupos resultantes da aplicação do *clustering* espectral regularizado com 2 a 20 grupos e $\tau = 1$ está apresentado na Figura 6, sendo importante ressaltar que a figura foi gerada várias vezes, e em todos os resultados foram semelhantes. A rede colorida de acordo com as etiquetas produzidas por esse método com 7 *clusters* está apresentada na Figura 7. Já o número de siglas em cada grupo, bem como as proporções observadas de ligação entre eles (análogo à matriz $\boldsymbol{\Theta}$, mas para os dados observados) podem ser encontrados, respectivamente, nas Tabelas 3 e 4.

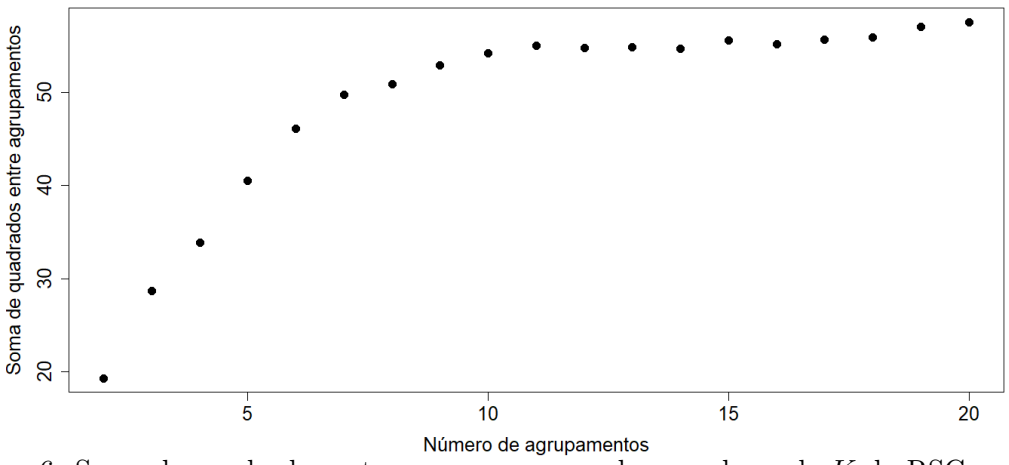


Figura 6: Soma de quadrados entre os grupos para alguns valores de K do RSC com $\tau = 1$

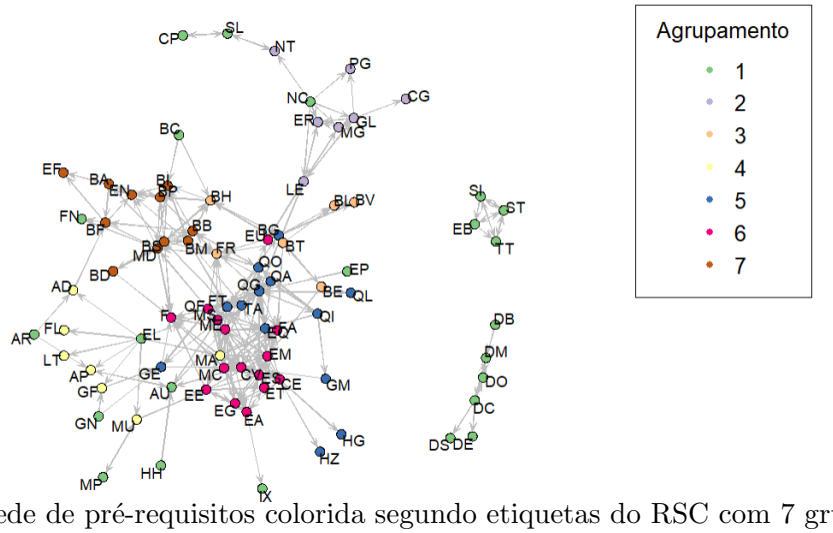


Figura 7: Rede de pré-requisitos colorida segundo etiquetas do RSC com 7 grupos e $\tau = 1$

Tabela 3: Número de siglas em cada grupo produzido pelo RSC com 7 agrupamentos

1	2	3	4	5	6	7
23	7	6	7	13	15	11

Tabela 4: Proporção de ligações dos elementos do grupo da linha ao grupo da coluna

Agrupamento	1	2	3	4	5	6	7
1	0,04	0,04	0,01	0,07	0,01	0,01	0,00
2	0,01	0,20	0,00	0,00	0,00	0,01	0,00
3	0,00	0,00	0,14	0,00	0,00	0,02	0,09
4	0,01	0,00	0,00	0,00	0,04	0,13	0,01
5	0,00	0,01	0,10	0,00	0,12	0,05	0,01
6	0,01	0,01	0,03	0,00	0,11	0,24	0,01
7	0,01	0,00	0,09	0,00	0,03	0,02	0,23

4 Discussão

As Figuras 1 e 2 apresentam como são realizadas as conexões dos departamentos em relação a pré-requisitos. Com base nelas, nota-se que grande parte deles possuem poucas ligações com os demais, apesar de existirem claramente siglas que são pré-requisitos para muitas outras. Para melhor compreensão dessas relações, foi ajustado o *Stochastic Block Model* para 1 a 10 agrupamentos, tendo sido apresentados os respectivos ICLs na Figura 3, da qual se conclui que os três melhores modelos tiveram 5, 4 e 7 *clusters*, todos com valores semelhantes desse critério.

Dessa forma, eles foram comparados em relação às incertezas *a posteriori* dos vértices, as quais estão na Figura 4. Dela, se conclui que o modelo com 7 agrupamentos se ajustou melhor, pelo fato de possuir incerteza máxima muito menor do que os demais. Por isso, foram utilizados esses grupos para colorir a rede, presente na Figura 5, e analisar os perfis produzidos a partir das Tabelas 1 e 2. Delas, nota-se que existe um grande grupo (1) e outro com apenas uma sigla (6), a matriz $\boldsymbol{\Theta}$ não é diagonalmente dominante, o que significa que não existem comunidades, mas ainda assim há 7 perfis:

1. Este grupo está na parte externa da rede e tem pouca probabilidade de se conectar com outros vértices;

2. Este *cluster* está na parte central da rede e tende a ser pré-requisito dos grupos 3 e 6 e não ter pré-requisitos;

3. Este agrupamento tem pré-requisitos dos grupos 2 e 7 e costuma ser pré-requisito para si;

4. Este grupo possui poucos pré-requisitos, mas, quando possui, são de matérias do *cluster* 2 ou dele mesmo e não é pré-requisito para as demais;

5. Esta sigla é pré-requisito para o grupo 6 e, em menor escala, para o 3 e para si. Além disso, tem pré-requisitos dos *clusters* 2, 6 e 7;

6. Este *cluster* é pré-requisito para a sigla do grupo 5 e para si, mas tem muitos pré-requisitos do agrupamento 2;

7. Este grupo tem pré-requisito apenas de si, mas, além disso, é pré-requisito dos grupos 3 e 5.

Para fins de comparação, foi ajustado um *clustering* espectral regularizado com $\tau = 1$. Para determinar o número adequado de grupos, foi utilizado o método do cotovelo para a soma de quadrados entre os agrupamentos apresentada na Figura 6, onde se observa uma inflexão com 7 *clusters*. Por isso, esse foi o número escolhido para a aplicação, que gerou grupos com os tamanhos apresentados na Tabela 3, que são mais próximos entre si do que os produzidos pelo SBM.

A partir da Figura 7, observa-se que as classes estão mais espalhadas do que as da Figura 5, o que dificulta uma melhor interpretação e sugere que a *clusterização* obtida não é tão boa quanto a anterior. Em relação às ligações entre grupos, apresentadas na Tabela 4 (que também não é diagonalmente dominante), percebe-se que as siglas dos grupos 2, 3, 5, 6 e 7 são pré-requisitos, principalmente, para as matérias do próprio agrupamento, enquanto as do 1 são pré-requisitos para as do 4 e, estes, só possuem pré-requisitos do grupo 1. Assim, apesar de também possuírem uma interpretação, ela é mais complicada e parece ser menos útil na prática.

5 Conclusão

Com base no exposto, nota-se que é possível construir uma rede direcionada de pré-requisitos entre as disciplinas ofertadas na Unicamp. Além disso, existem muitas siglas que são pré-requisitos para poucas matérias, enquanto algumas o são para muitas. Adicionalmente, mostrou-se que é possível agrupá-las em sete grupos interpretáveis a partir de duas técnicas distintas: *Stochastic Block Model* (SBM) e *Regularized Spectral Clustering*.

Ademais, foi possível observar que os agrupamentos produzidos são bastante distintos e que aquele produzido pelo SBM tem melhor interpretabilidade e aplicação prática. Assim, foi possível identificar as relações de pré-requisitos existentes entre esses *clusters*, atingindo satisfatoriamente o objetivo inicial.

Referências

Bouveyron, C., Celeux, G., Murphy, T. B., e Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Brian S. Everitt, Dr Sabine Landau, D. M. L. D. D. S. (2011). *Cluster Analysis, Fifth Edition (Wiley Series in Probability and Statistics)*. Wiley Series in Probability and Statistics. Wiley, 5th edition.

Daudin, J.-J., Picard, F., e Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

Qin, T. e Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.