ME921 - Aprendizado de Máquinas Não-Supervisionado
Universidade Estudual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Bacharelado em Estatítica

Caio Théodore Genovese Huss Oliveira - 247005

Implementação de Weighted Stochastic Block Model em dados de Game of Thrones

1 Introdução

1.1 Contexto

Este trabalho utiliza dados do site Network of Thrones, um projeto que catalogou todas as interações entre personagens nos 5 livros da franquia *A Song of Ice and Fire*, mais conhecida pela sua adaptação à série *Game of Thrones*. O objetivo é encontrar grupos entre personagens por uma implementação própria de modelagem de rede, melhor descrita na secção 2. Uma interação é contada quando os nomes ou apelidos dos personagens aparecem a uma distância de até 15 palavras no texto.

1.2 Dados

Os dados foram obtidos no arquivo `all_edges.csv`, possuindo 3 variáveis[1]: **Source** e **Target**, identificando os personagens interagindo, e **Weight**, o número de vezes que a interação ocorre.

Há uma **conexão** entre dois personagens caso eles interajam pelo menos uma vez ao longo dos livros.

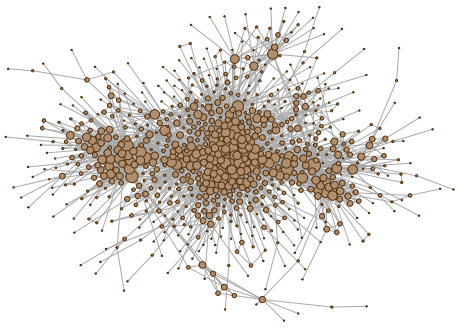

Figura 1: Rede dos dados

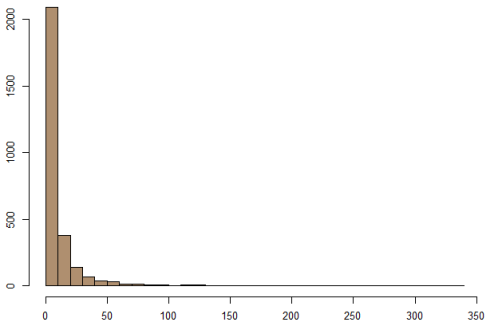

Figura 2: Histograma de weight

Percebe-se que a rede parece ser concentrada em torno de alguns personagens, com muitos nodos possuindo apenas uma conexão, e muitos valores baixos de **weight**, cujo histograma se assemelha a uma distribuição exponencial. Podemos construir a hipótese que há um grupo de protagonistas, que interage frequentemente entre si, e coadjuvantes, que aparecem menos frequentemente.

[1] a coluna "Id"foi descartada

2 Materiais e Métodos

2.1 Weighted Stochastic Block Model

O modelo de blocos estocásticos modela, por clusters, a probabilidade de um nodo conectar-se a outro, apenas com a informação do clusters a que pertencem:
$\mathbb{P}(X_{ij} = 1|Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql}, \quad 1 \leq i, j \leq n, \ 1 \leq q, l \leq Q$, onde X é a matriz de conexão, Q o número de clusters, e $z_{iq} = 1$ se a obs. i pertence ao cluster q. O termo "weighted" do modelo se refere à adição de um peso à conexão (número de conexões, duração, etc). A intensidade é representada pela matriz de pesos Y, com a condição $X_{ij} = 0 \Rightarrow Y_{ij} = 0$. Pode-se modelar o peso das conexões por diversas distribuições.

Neste trabalho, segui o exemplo de *T. L. James Ng, T. B. Murphy*[3], utilizando a distribuição gamma, por abranger outras distribuições em sua flexibilidade, entre elas a exponencial. Um modelo com Q clusters se dá por:

$$Z \sim \mathbb{M}(1, \theta_1, ..., \theta_Q), \ (X_{ij}|Z_{iq} = 1, \ Z_{jl} = 1) \sim Bernoulli(\pi_{ql})$$

$$(Y_{ij}|X_{ij} = 1, \ Z_{iq} = 1, \ Z_{jl} = 1) \sim \Gamma(\alpha_{ql}, \ \beta_{ql}), \quad X_{ij} = 0 \Rightarrow Y_{ij} = 0^{2}$$

onde $Z_{iq} = 1$ se a obs. i pertence ao cluster q. Diz-se que há uma **comunidade** em um cluster caso $\pi_{qq} \geq \pi_{qq'}, \forall q' \neq q$, ou seja, maior probabilidade dos elementos do cluster ligarem-se com elementos do mesmo cluster, comparados a outros.

2.2 Desenvolvimento do algoritmo

Não encontrei códigos ou funções que implementam um algoritmo para o modelo com pesos, portanto todas as funções foram construídas por mim, com base em *James Ng, 2021*[3] e *El Haj, 2020*[1]. Os pacotes [2] foram utilizados, e todos os códigos estão disponíveis no GitHub. O algoritmo segue passos similares ao algortimo EM[4], com um passo maximizando a verossimilhança do cluster atribuído a cada observação, e o segundo ajustando os parâmetros com base nas etiquetas, iterando até a convergência[3]

Obtemos os estimadores variacionais[3] dos parâmetros $\widehat{\theta}_{1\times Q}$, $\widehat{\pi}_{Q\times Q}$, $\widehat{\alpha}_{Q\times Q}$, $\widehat{\beta}_{Q\times Q}$, e a estimação dos clusters $\tau_{1\times n}$. Um modelo "sem pesos" utiliza apenas θ *e* π. Os chutes iniciais de τ são feitos por uma clusterização por *K-médias* em cima da matriz X[4][1].

O algoritmo se mostra estável e convergente quando os dados seguem as suposições teóricas, no entanto em casos extremos os estimadores não convergiram. Para resolver este problema, o algoritmo é reiniciado caso $\exists q \in [1, Q] \,|\, \theta_q \leq 0 \, ou \, \theta_q \geq 1$.

O algoritmo possui uma complexidade computacional $\mathbb{O}(n^2)$[3], portanto o número de iterações é uma escolha importante. Repetindo as simulações em [3] para o exemplo de 3 clusters com 100 nodos e 100 iterações do algoritmo, obtiveram-se as diferenças médias:

$$||\widehat{\theta} - \theta||_2 \approx 0; \ ||\widehat{\pi} - \pi||_F \approx 0.05; \ ||\widehat{\alpha} - \alpha||_F \approx 2.74; \ ||\widehat{\beta} - \beta|| \approx 1.87$$

Com base em resultados observados e no fato que as aproximações melhoram com o aumento no número de nodos[3], optou-se por utilizar 50 iterações para o cálculo dos critérios de seleção, e 100 iterações para o modelo final.

[2] $Y_{ij}|X_{ij} \sim \delta\{0\}$[3], optei por simplificar.
[3] Os cálculos são muito extensos para este relatório, mas se encontram em [3] e [1].
[4] A teoria não especifica, mas sob testes a matriz X se mostrou mais precisa que a matriz Y.

2.3 Critérios de seleção

O *Bayseian Information Criterion*(BIC) e *Integration Classification Likelihood*(ICL) foram aproximados com base na log-verossimilhança[1] $\mathbb{L}(Y, X, Z; \theta, \pi, \alpha, \beta, m_Q)$:

$$\widehat{BIC} = 2\mathbb{L} - (3Q^2 + Q) \times log(n); \quad \widehat{ICL} = \mathbb{L} - \frac{3}{2}Q(Q+1) \times log(n(n+1)) - \frac{Q-1}{2}log(n)$$

O ICL tende a penalizar mais o número de clusters. Análise visual será utilizada caso os dois critérios possuam resultados diferentes, e é importante ressaltar que os critérios são aproximações, portanto deiferenças pequenas não serão consideradas, optando-se pelo menor número de clusters, em favor do menor custo computacional.

3 Resultados

3.1 Número de clusters

Devido ao alto custo computacional com um banco de dados deste tamanho, apenas modelos com 1 a 5 clusters foram considerados. Observa-se que os dois critérios, assim como a log-verossimilhança, seguem um comportamento parecido, com os modelos de 3 e 5 clusters como candidatos. Optou-se pelo modelo que maximizou os dois critérios, o modelo de 5 clusters, apesar de seu alto custo computacional.

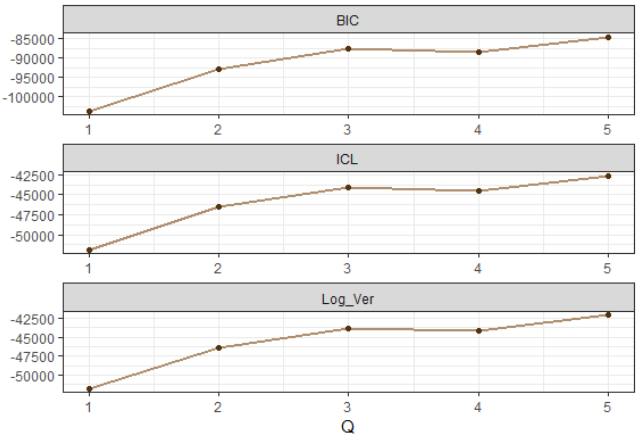

Figura 3: Critérios de seleção

3.2 Análise de perfil

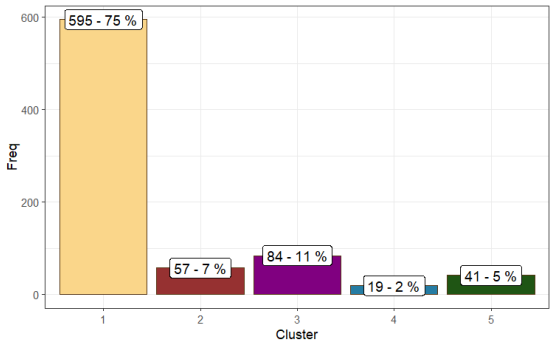

Figura 4: Histograma dos clusters

O cluster 1 se destaca por seu alto número de nodos, apoiando a hipótese do grupo de coadjuvantes. Isto é esperado, já que a série de livros contém um alto número de personagens com poucas aparições, entrando e saindo da história para acrescentar ao desenvolvimento ou construir o cenário[5]. A distribuição dos clusters restantes é visível na imagem 5, e suas conexões na matriz da imagem 6. A tabela 1 nos mostra os 3 personagens com mais interações em cada cluster, para termos uma ideia dos elementos representados pelos clusters. Observa-se que o cluster 1 não está aglomerado em torno de si mesmo, fortalecendo a hipótese de ser um cluster de coadjuvantes.

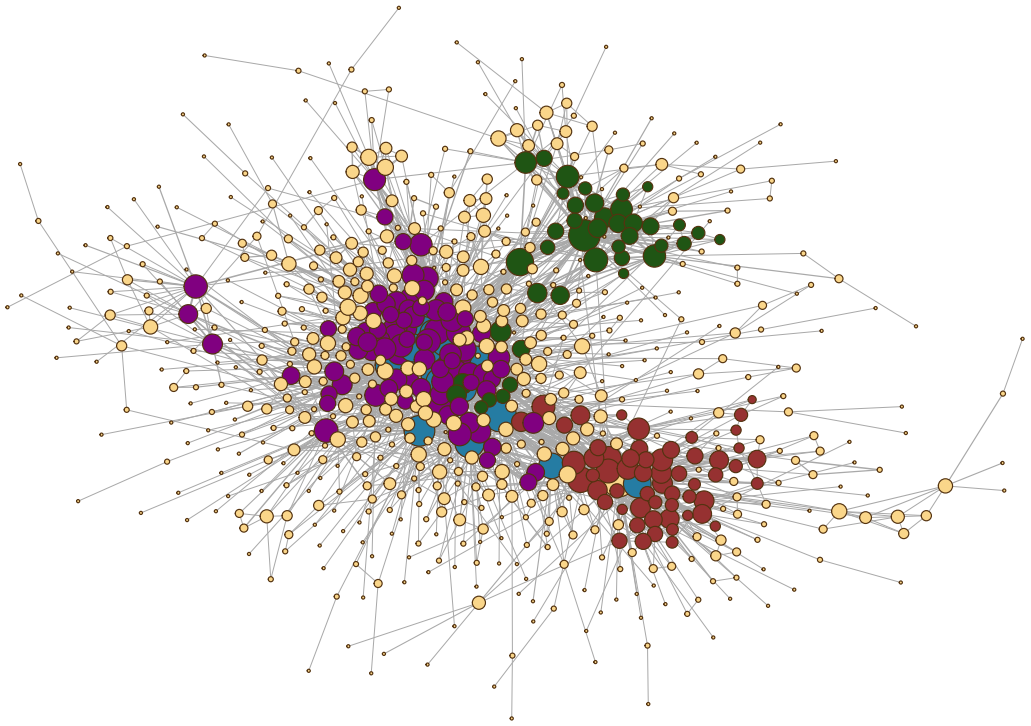

Figura 5: Rede após clusterização com WSBM

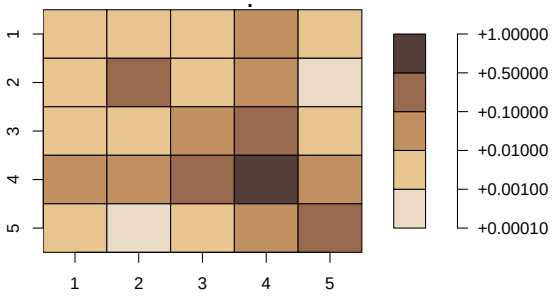

Figura 6: Probabilidade de conexão entre clusters

Tyrion Lannister e *Jon Snow* parecem outliers entre os elementos do cluster 1, sendo dois dos protagonistas da franquia. Isto pode ser explicado pelo fato dos dois personagens viajarem ao redor do mundo nos livros, interagindo com muitos coadjuvantes do cluster 1[5], resultando em sua inclusão no mesmo. Com estas duas exceções levadas em conta, o cluster 4 aparenta conter os personagens principais da série em *Westeros*, o continente principal, contendo os personagens com o maior número de interações. Já o cluster 5 aparenta conter os personagens de *Essos*, continente onde se passa a maior parte da história de **Daenerys Targerian**, uma das protagonistas, que é frequentemente mencionada em conselhos de guerra pelos personagens do cluster 4[5], explicando a maior probabilidade de conexão.

Cluster_1	Cluster_2	Cluster_3
Tyrion Lannister_2873	Jeor Mormont_484	Tommen Baratheon_539
Jon Snow_2757	Aemon Targaryen_395	Sandor Clegane_528
Dontos Hollard_103	Mance Rayder_389	Margaery Tyrell_442

Cluster_4	Cluster_5
Cersei Lannister_2232	Daenerys Targaryen_1608
Joffrey Baratheon_1762	Barristan Selmy_470
Eddard Stark_1649	Jorah Mormont_401

Tabela 1: Elementos com mais interações de cada cluster

Os clusters 2 e 3 aparentam ser compostos por personagens secundários, mas não irrelevantes, da *Muralha*, lugar onde o protagonista *Jon Snow* passa a maior parte de sua história, e *King's Landing*, cidade real central à franquia, respectivamente.

Figura 7: Distribuições gamma estimadas, com médias e variâncias

A figura 7 nos mostra a distribuição gamma estimada pelo modelo, para os pesos das conexões entre clusters. Nota-se a média extremamente alta do cluster 4 interagindo consigo mesmo. A interação com o cluster 4 é, no geral, também a que possui maior média entre os outros clusters, reforçando a ideia de que este é o cluster protagonista.

4 Discussão

Ao longo deste relatório, um modelo WSBM foi proposto, explicado, implementado e interpretado, com o objetivo de encontrar grupos entre a rede de interações dos 5 livros da franquia *A Song Of Ice And Fire*. Cinco clusters foram encontrados:

- 1: Coadjuvantes ou viajantes, com os outliers em potencial *Jon Snow* e *Tyrion Lannister*, com poucas interações. É de longe o mais numeroso, contendo 75% dos nodos. Não forma uma comunidade.
- 2: Personagens secundários da *Muralha*, linha da história de *Jon Snow*. Possuem muitas interações entre si, mas poucas com outros clusters. Forma uma comunidade.
- 3: Personagens secundários de *King's Landing*. Possuem maior probabilidade maior de conexão com o cluster 4 do que entre si, e as interações também possuem peso médio maior com o cluster 4, condizente com o comportamento de personagens secundários no cenário principal. Não forma uma comunidade.

- 4: O cluster da maioria dos protagonistas de *Westeros*, inclui os principais *Stark* (com a exceção de Rickon, secundário), os principais *Lannister* (com exceção de Tyrion, outlier), e a família real, com exceção de *Tommen*. É o cluster com as maiores probabilidades de conexão com os outros clusters e também entre o próprio cluster, as maiores média e variância de peso de conexão, o menor número de personagens e, apesar de não aparentar na figura 5, forma uma comunidade.

- 5: Personagens localizados em *Essos*, seguindo a história de *Daenerys Targerian*, com probabilidades de conexão com outros clusters mais baixas (mas não-nulas!), devido à história se passar em outro continente. Possui um aparente outlier em *Daenerys Targerian* com um número discrepante de conexões, devido a ela ser a única protagonista do cluster. Forma uma comunidade.

A distribuição Gamma se mostra útil neste modelo, em função da sua alta flexibilidade, adaptando-se a distribuições muito centralizadas, como nos pesos das conexões entre os clusters 2 e 3, e também a distribuições com alta variância, como os pesos das conexões entre elementos do cluster 4. Um ponto negativo deste modelo é o alto custo computacional[3], com tempo superior a uma hora para gerar o resultado.

Pode-se dizer que a hipótese inicial estava parcialmente correta: Há um grupo de coadjuvantes, porém os protagonistas se organizam em vários grupos diferentes, e há espaço para grupos de persongagens secundários, de importância média.

Referências

[1] A. El Haj et al. "Estimation in a binomial stochastic blockmodel for a weighted graph by a variational expectation maximization algorithm". Em: *Communications in Statistics - Simulation and Computation* (2020).

[2] Pacotes: Tidyverse, Matrix, MASS, igraph.

[3] T. L. James Ng e T. B. Murphy. "Weighted Stochastic Block Model". Em: *Statistical Methods & Applications* (2021).

[4] Slides do professor Guilherme Ludwig.

[5] George R. R. Martin. *A Song Of Ice And Fire (all books)*. HarperVoyager.