Weighted Stochastic Block Model implementation on A Song Of Ice And Fire data

Contents

1 Introduction

1.1 Context

Famous for its intricate plots of fantasy medieval politics, A Song Of Ice And Fire is a franchise composed of 5 books (currently), written by George R. R. Martin, with an equally well known TV adaptation *Game of Thrones*, wich ended its run in 2019, to disappointing reviews.

This work aims to use ASOIAF character interactions data, obtained from Network of Thrones, to implement a weighted stochastic block model (WSBM) algorithm made by myself, based on the theoretical work in *James Ng, 2021*[2] e *El Haj, 2020*[1], trying to find groups between the characters based purely on the cluster model for their interactions, better described in section 2.

Only data for the book interactions were considered. In this dataset, an interaction is counted if two characters' names or aliases appear within 15 words of one another [5].

1.2 Data

All data comes from the `all_edges.csv` file, containing 3 variables[1]: **Source** and **Target**, identifying the characters in the interaction, and **Weight**, the interaction count for the two characters.

There is a **connection** between two characters if they interact at least once in the dataset, and connections are bilateral, making no distinction between target and source.

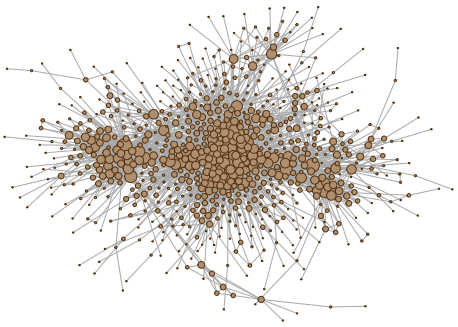

Figure 1: Dataset network

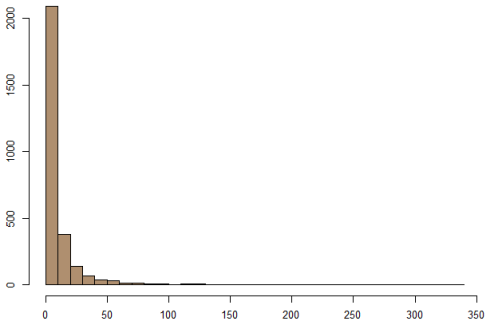

Figure 2: Histogram of connection weight

Based on figures 1 and 2, with a seemingly exponential curve for the **weight** distribution, and a network with many edge nodes who only connect with one other node, a hypothesis can be made of two groups: protagonists and secondary characters, the latter being larger, disconnected, and interacting more often with the former than with itself.

The data contains 2823 observations for interactions between 796 characters.

[1] "Id" was discarted

2 Methodology and tools

2.1 Weighted Stochastic Block Model

A stochastic block model models, through clusters, the probability of a node connecting to another: $\mathbb{P}(X_{ij} = 1 | Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql}, \;\; 1 \leq i, j \leq n, \; 1 \leq q, l \leq Q$, where X is the connection matrix, $X_{ij} = 1$ if character i connects with character j, $X_{ij} = 0$ otherwise. Q is the number of clusters, and $z_{iq} = 1$ if character i belongs to cluster q.

The *"Weighted"* term of the model comes from the addition of a **weight** to every connection. This can have many interpretations for different uses, in this case it is the number of interactions between the two characters. This weights is represented by the weight matrix Y, with the condition that $X_{ij} = 0 \Rightarrow Y_{ij} = 0$

Many distributions can be used to model the Y weight matrix. This project follows the example of *T. L. James Ng*[2], making use of a gamma distribution due to its flexibility, being able to take the form of other distributions with the right parameters, the exponential between them. A model with Q clusters has the proprieties:

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

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

A cluster Q forms a *community* if the probability of connection between two nodes of Q is higher than the connection between a node of Q and any other cluster, $\pi_{qq} \geq \pi_{qq'}, \forall q' \neq q$.

2.2 Algorithm implementation

Due to apparent unavailability of functions that implement WSBM in R, all functions were developped by myself, based on the theory in *James Ng, 2021*[2] and *El Haj, 2020*[1]. The packages [4] were used, and the code is available on GitHub[3].

The algorithm follows similar steps to an EM algorithm, maximizing the likelihood of the cluster attributed to each node, and the adjusting the parameters based on the new clusters, iterating these two steps until conversion. Formulas for the likelihood and parameter estimation are available in [2] and [1]. The variational estimators[2] $\widehat{\theta}_{1\times Q}$, $\widehat{\pi}_{Q\times Q}$, $\widehat{\alpha}_{Q\times Q}$, $\widehat{\beta}_{Q\times Q}$ are obtained, with the cluster estimation $\tau_{1\times n}$. A model whitout weights (SBM) uses only $\theta \; e \; \pi$. Initial cluster guesses for the algorithm are made implementing a K-means algorithm upon X[4][1].

The algorithm shows convergence when theoretical assumptions are satisfied, but fails on extreme cases, to protects against such cases, the algorithm is restarted if $\exists q \in [1, Q] \,|\, \theta_q \leq 0 \, ou \, \theta_q \geq 1$.

The algorithm's complexity is $\mathbb{O}(n^2)$[2], therefore the number of iterations is an important choice. Using the simulations made in [2], in simulated data with 100 nodes and running 100 iterations, we obtain the average differences:

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

[2] $Y_{ij} | X_{ij} \sim \delta\{0\}$[2], simplified for comprehension.
[3] May be incomplete due to file loss, but all functions are present
[4] Theory doesn't specify wich matrix to use, but upon simulations X wields better results

Based on these results, and that the approximations get better with more nodes[2], 50 iterations will be used to choose the number of clusters, and 100 for the final model.

2.3 Selection criteria

The *Bayesian Information Criterion*(BIC) and *Integration Classification Likelihood*(ICL) were approximated with the log-likelihood $\mathbb{L}(Y, X, Z; \theta, \pi, \alpha, \beta, m_Q)$[1]:

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

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

The ICL penalises more heavily the number of clusters. Visual analysis will be used if criteria diverge in the choice. Given that these are approximations, very small differences won't be taken into account, prefering the smaller number of clusters in favor of less computational cost.

3 Results

3.1 Number of clusters

Due to the high computational cost with a dataset of this size, only models with 1 to 5 clusters were considered, where a 1 cluster model would just mean there are no clusters in the data. The criteria and the log-likelihood follow a similar curve, observed in figure 3 and both a 3 cluster and 5 cluster model seem like good candidates. The five cluster model was picked due to higher criteria and uninteresting results from the 3 cluster model in some trials.

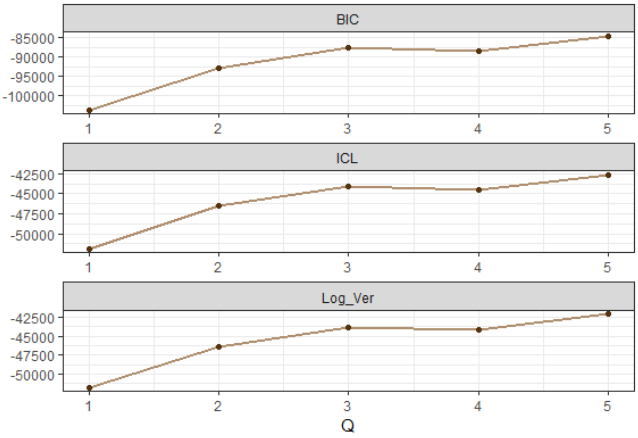

Figure 3: Selection criteria

3.2 Cluster profiles

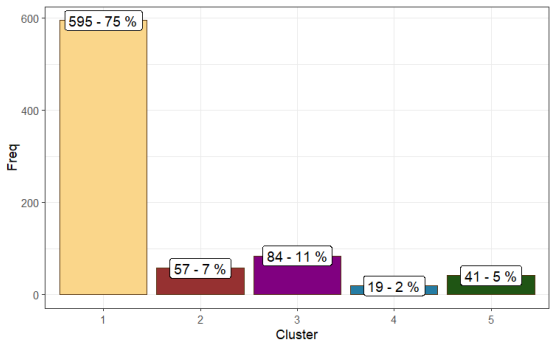

Figure 4: Frequency of each cluster

Cluster 1 stands out with 3 quarters of the data, supporting the secondary characters hypothesis. This is expected, as the franchise contains many characters with few apparitions, coming in and out of the story to further the plot[3]. The distribution of the remaining clusters is observed in figure 5, and its connection probabilities in figure 6. Tables 1 and 2 shows the 3 characters with most interactions in each cluster, with the number of interactions, and the first 20 characters for each cluster can be found in the appendix[5].

It's noticeable how cluster 1 doesn't seem to agglomerate, doesn't have a higher probability of connecting with itself, and has, beside two possible outliers, the characters with the fewest interactions.

Tyrion Lannister and *Jon Snow* seem to be outliers among cluster 1, being the two characters with the most interactions in the whole franchise. This could be explained by how much these characters travel in the story[3], showing a different connection pattern than others, and connecting with characters from multiple clusters, especially 1.

[5] the full list for all clusters can be found in the "clusters_5q.csv" file

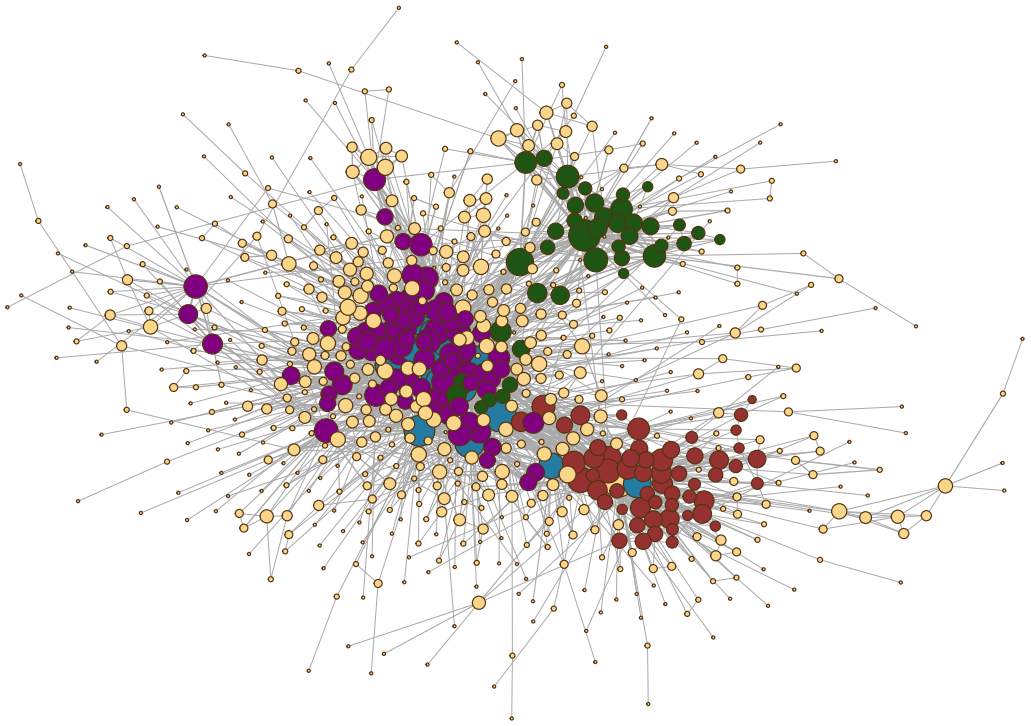

Figure 5: Clusterized netwrok

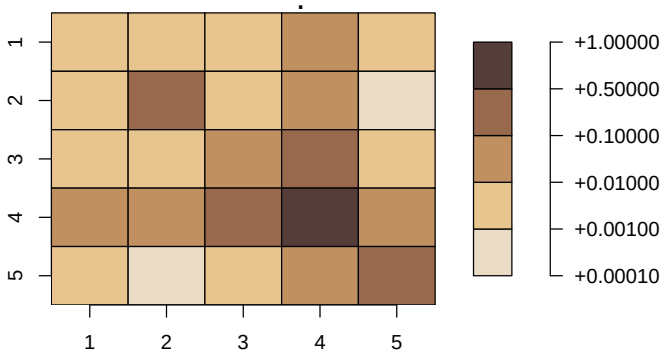

Figure 6: Connection probability between clusters

With the two exceptions in cluster 1 partially explained, cluster 4 contains all major westerosi protagonists, having all of the major *Starks* and *Lannisters*, as well as the royal family and some important court presences, *Varys* and *Peter Baelish*. It has relatively high connection probabilities with all clusters, and the highest of all with itself at 72.5%, expected since the story mainly revolves around these characters.

Cluster 5 contains our characters from *Essos* and the only remaining protagonist, *Daenerys Targerian*, with the lowest probability of connection with other clusters due to the physical isolation of most of these characters. It still shows a decent connection probability with cluster 4, probably due to how often *Daenerys* is mentionned in war councils, along with her advisors, and vice-versa, and also to the presence of *Barristan Selmy* in cluster 5.

Table 1: Characters with the most interactions in clusters 1, 2 and 3

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

Table 2: Characters with the most interactions in clusters 4 and 5

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

Clusters 2 and 3 are composed of secondary, but frequent, characters from *The Wall* and *King's Landing*, respectively.

Figure 7: Estimated gamma distributions for connection weights, with estimated mean and variance

Figure 7 shows us the different connection distribution between characters of the clusters, when there is one. Connections between characters of cluster 4 stand out with the highest mean, at over 40 interactions, and having the flattest curve due to the frequent interactions of said characters. Connections between characters of clusters 2 and 3 stand out as the lowest interaction averages, being highly concentrated around a mean of 4.

Surprisingly, despite having the lowest connection probability at 0.04%, connections between characters of clusters 5 and 2 show the second highest variance. This might be due to an outlier or, more likely, to the low sample size of connections between characters of these two clusters.

4 Discussion

Throughout this report, a WSBM model was proposed, explained, implemented and interpreted, with the goal of finding groups in the interactions of characters in *A Song Of Ice And Fire*. Five clusters were found, with distinct characteristics:

1. **Minor characters and travelers:** Contains three quarters of the characters, who do not appear frequently and don't connect with many others, scattered around the world and the timeline, except for *Jon Snow* and *Tyrion Lannister*, potential outliers, who might have been alocated to this cluster due to frequent interactions with many of the other clusters.
   Doesn't form a community.

2. **The Wall:** Secondary characters from the wall, where most of *Jon Snow*'s story happens. Has a high probability of connection with itself, but low connection weight means, suggesting many characters who appear together, but not so frequently.
   Forms a community.

3. **King's Landing:** Secondary characters in King's landing, also containing the *Tyrells* and *Tommen*, due to them only becoming relevant later in the story. Has a higher probability of connection with cluster 4 than with itself, suggesting these characters' plots revolve around the actions or stories of the characters in cluster 4. Has higher connection weight means with itself and cluster 4.
   Does not form a community.

4. **Protagonists:** The main westerosi protagonists, containing the *Starks*, *Lannisters*, the *Baratheons*, as well as characters very important to the story of them, such as *Theon Greyjoy, Varys and Peter Baelish (littlefinger)*. Has the highest probability of connection, the highest average connection weight, highest connection variance, and is the smallest cluster. Has connection probabilities above 1% with all other clusters.
   Forms a community.

5. **Essos:** Essos characters, following the story of *Daenerys Targerian*. Phisically isolated, has low connection probabilities with most other clusters, except for 4. Connection probability with characters of the same cluster is the second highest at 23%, and connection weights are concentrated on low values with other clusters.
   Forms a community.

The gamma distribution adequates well to the data, allowing for curves highly concentrated around low values, like the connections between clusters 2 and 3, but also much flatter curves, like the connections of cluster 4 with itself. A negative of this algorithm is the high computational cost, with a runtime of over an hour for the final model.

The initial assumption seems to be partially correct, as a cluster of minor characters that revolve around the others was found, however there are many different patterns between the protagonists and secondary characters.

References

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

[2] T. L. James Ng and T. B. Murphy. "Weighted Stochastic Block Model". In: *Statistical Methods & Applications* (2021).

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

[4] R packages: Tidyverse, Matrix, MASS, igraph.

[5] https://networkofthrones.wordpress.com/.

5 Appendix

5.1 Cluster tables

Table 3: top 20 characters by number of interactions, cluster 1

	char	interactions	cluster
1	Tyrion-Lannister	2873	1
2	Jon-Snow	2757	1
3	Dontos-Hollard	103	1
4	Lommy-Greenhands	99	1
5	Axell-Florent	98	1
6	Nan	91	1
7	Cleos-Frey	86	1
8	Pylos	77	1
9	High-Sparrow	76	1
10	Salladhor-Saan	74	1
11	Shagga	70	1
12	Obara-Sand	68	1
13	Mycah	65	1
14	Jaqen-Hghar	64	1
15	Garin-(orphan)	63	1
16	Hyle-Hunt	59	1
17	Nestor-Royce	58	1
18	Tyene-Sand	57	1
19	Elia-Martell	56	1
20	Brandon-Stark	55	1

Table 4: top 20 characters by number of interactions, cluster 2

	char	interactions	cluster
1	Jeor-Mormont	484	2
2	Aemon-Targaryen-(Maester-Aemon)	395	2
3	Mance-Rayder	389	2
4	Melisandre	327	2
5	Grenn	265	2
6	Pypar	232	2
7	Gilly	216	2
8	Selyse-Florent	207	2
9	Janos-Slynt	196	2
10	Craster	193	2
11	Bowen-Marsh	187	2
12	Tormund	180	2
13	Qhorin-Halfhand	166	2
14	Ygritte	165	2
15	Alliser-Thorne	159	2
16	Eddison-Tollett	154	2
17	Rattleshirt	129	2
18	Benjen-Stark	125	2
19	Val	125	2
20	Satin	101	2

Table 5: top 20 characters by number of interactions, cluster 3

	char	interactions	cluster
1	Tommen-Baratheon	539	3
2	Sandor-Clegane	528	3
3	Margaery-Tyrell	442	3
4	Gregor-Clegane	410	3
5	Hodor	383	3
6	Luwin	376	3
7	Pycelle	357	3
8	Rickon-Stark	345	3
9	Lysa-Arryn	341	3
10	Loras-Tyrell	329	3
11	Bronn	313	3
12	Meryn-Trant	298	3
13	Edmure-Tully	291	3
14	Meera-Reed	291	3
15	Ilyn-Payne	268	3
16	Myrcella-Baratheon	264	3
17	Jojen-Reed	261	3
18	Rodrik-Cassel	253	3
19	Ramsay-Snow	241	3
20	Gendry	238	3

Table 6: top 20 characters by number of interactions, cluster 4

	char	interactions	cluster
1	Cersei-Lannister	2232	4
2	Joffrey-Baratheon	1762	4
3	Eddard-Stark	1649	4
4	Jaime-Lannister	1569	4
5	Sansa-Stark	1547	4
6	Bran-Stark	1508	4
7	Robert-Baratheon	1488	4
8	Arya-Stark	1460	4
9	Robb-Stark	1424	4
10	Stannis-Baratheon	1375	4
11	Catelyn-Stark	1230	4
12	Samwell-Tarly	934	4
13	Theon-Greyjoy	782	4
14	Petyr-Baelish	733	4
15	Tywin-Lannister	690	4
16	Renly-Baratheon	639	4
17	Varys	609	4
18	Brienne-of-Tarth	521	4
19	Davos-Seaworth	483	4

Table 7: top 20 characters by number of interactions, cluster 5

	char	interactions	cluster
1	Daenerys-Targaryen	1608	5
2	Barristan-Selmy	470	5
3	Jorah-Mormont	401	5
4	Drogo	307	5
5	Hizdahr-zo-Loraq	292	5
6	Quentyn-Martell	246	5
7	Irri	220	5
8	Rhaegar-Targaryen	202	5
9	Daario-Naharis	196	5
10	Jhiqui	183	5
11	Jon-Connington	169	5
12	Viserys-Targaryen	169	5
13	Belwas	153	5
14	Illyrio-Mopatis	138	5
15	Aegon-Targaryen-(son-of-Rhaegar)	136	5
16	Haldon	135	5
17	Skahaz-mo-Kandaq	135	5
18	Jhogo	131	5
19	Missandei	121	5
20	Aggo	112	5