LINMA2472-Homework 1

Groupe 21
Cédric Antoine, Romain Graux, Lionel Lamy

Octobre 2020

1 Introduction

In this work we are going to analyse some relations between characters of the book « Harry Potter and the Sorcerer » for the course LINMA2472 (UCLouvain). In order to do this we are going to use the 'networkx'and 'spacy' package of python to compute, plot and analyse these relations.

You can find our code on Github @RomainGrx/LINMA2472-Homeworks

2 Part I

In the graph below, we have represented every character of the book and linked together all character whose names were in the same paragraph. For each occurrence, we have increased the weight of the relation (represented here by the thickness of the line). Moreover, the size of the nodes represents the number of different relationships.

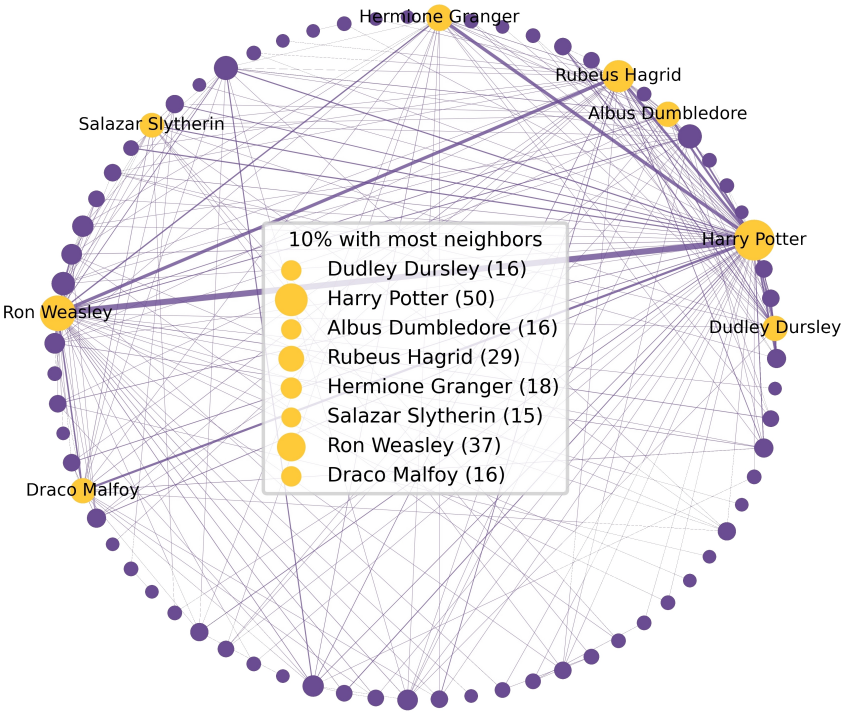

Figure 1: Characters interactions

2.1 Louvain algorithm

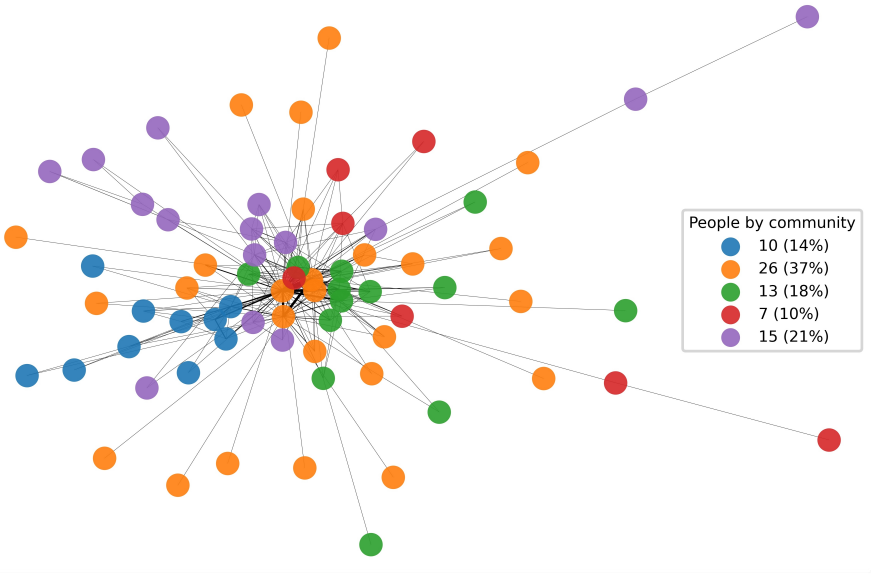

Figure 2: Community detection using the Louvain method

We observe that the Louvain algorithm splits the population of the book into 5 communities. The orange community with the most characters includes Harry, Hagrid, Hermione and Ron who are the main characters of the book (as seen it in Figure 1).

The degree of assortativity of our graph is -0.294. This means that in our graph, nodes with a high degree have a lower propensity to connect to nodes of similar degree than nodes with a lower degree. This is most notably the case in hierarchical networks.

2.2 K-core decomposition

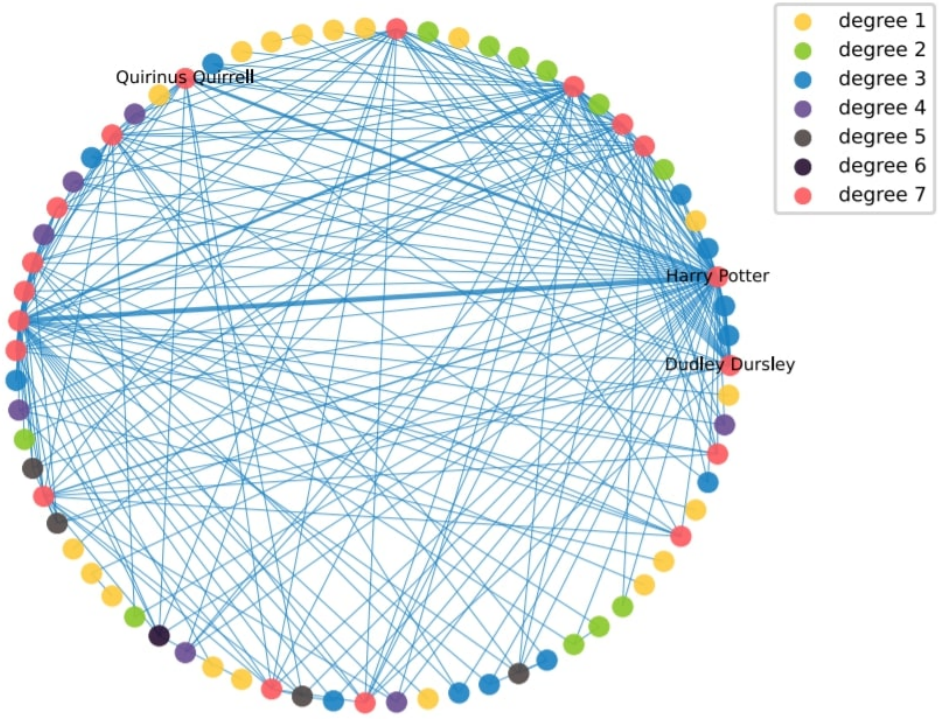

Figure 3: K-core decomposition

Thanks to the k-core decomposition we can observe that the nodes are separated in 7 cores.
Unlike what we could think we notice that « Harry Potter », despite being the main character, is not the only important characters. We see characters like « Dudley Dursey » or « Quirinus Quirell »,who are less known, have the same degree of importance as Harry.
Thanks to the k-core decomposition we can clearly observe which characters have a high or low degree of implication with others in the story.

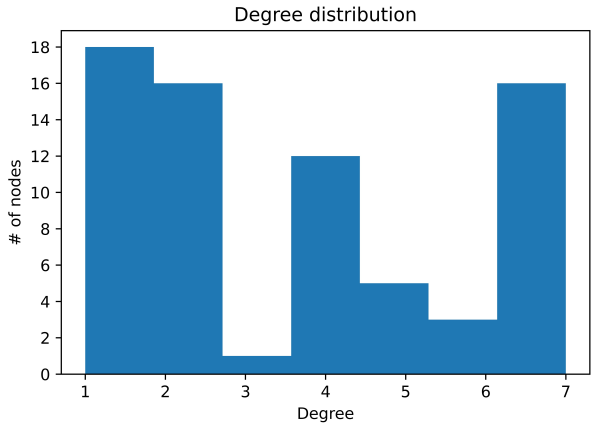

Figure 4: Degree distribution of nodes

In the histogram (FIGURE 4) we see the number of characters in each core. We can conclude that most of the characters have either a very low degree(1 or 2) or a very high one(7). We notice that a minority of nodes have a medial degree(3-6).

2.3 Preferential attachment network

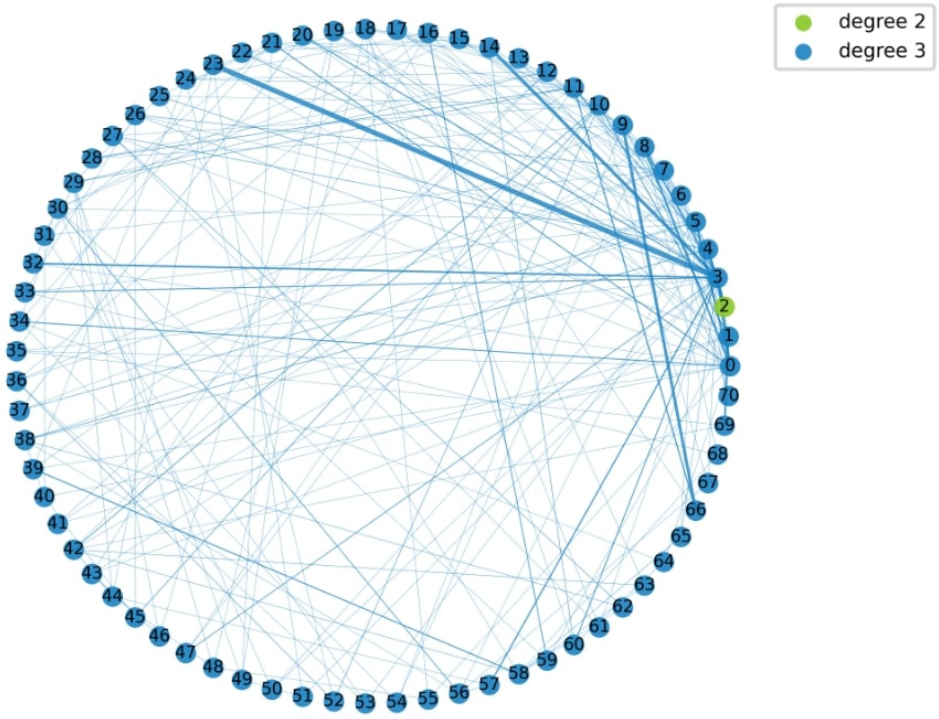

Figure 5: Preferential attachment applied to k-core decomposition

In the preferential attachment network, we can spot a major difference to our network. We see that nearly all nodes have degree 3, contrary to our graph where nodes had a degree from 1-7. This is probably due to the fact that the albert-barabazi model has power-law degree distributions and that therefor most of the nodes tend to have the same degree. From this wee can interpret that in the preferential attachment network nearly every node has the same 'importance' whereas in our network some nodes are clearly more or less important.

3 Part II

3.1 Influence maximisation problem

When we use the maximisation influence problem to determine the k=5% best starting nodes, we come across the 3 nodes Harry Potter, Rubeus Hagrid and Ron Weasley as shown in the graph 6.

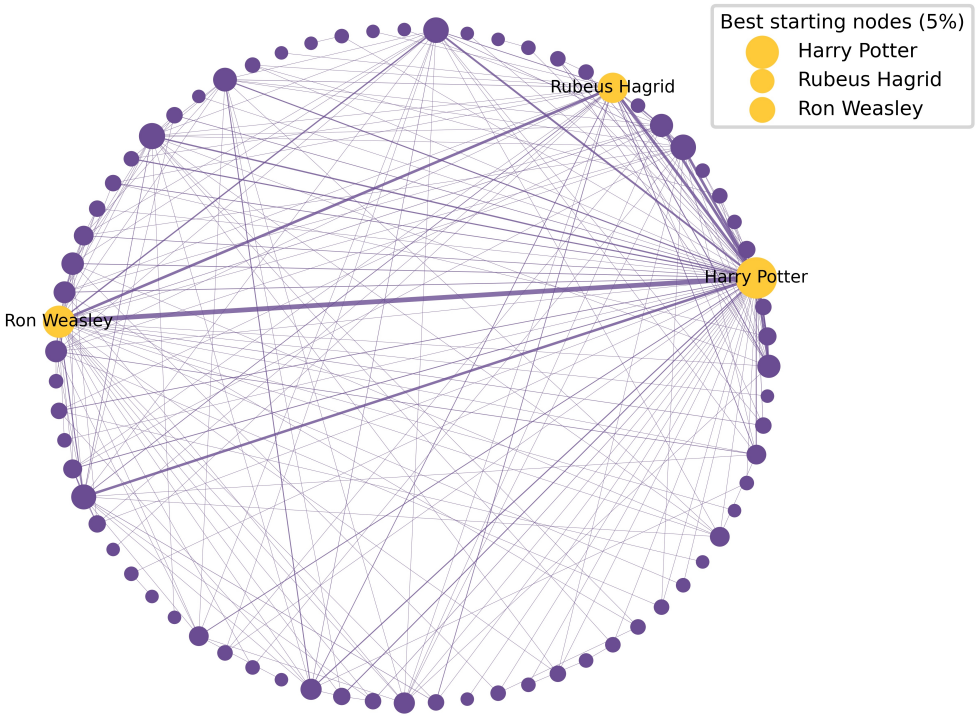

Figure 6: Influence maximisation with k=5%

3.2 Comparaison between influence maximisiation, highest degree and random selection

To compare the effect of IMP with other choices of nodes (random and highest degree), we studied 2 cases :

1. the effect of the size k with a fixed p
2. the effect of p with a fixed k

On Figure 7, we can see that the random leads to a smaller final spread which is obviously normal. Moreover, given that they very often have neighbours in our graph we can see that we have to choose our starting nodes skillfully, at the risk of being stuck in "dead ends" if we choose nodes without too much influence, like in the random case. IMP and highest degree have spread on all the graph.

Finally, on Figure 8, we can observe the non-linearity (decreasing derivative) of the 3 cases. This non-linearly decreasing derivative could be explained by the fact that when the probability is small, changes can lead to a large change in the spread, but the closer the probability approaches 1, the less the spread will be impacted. This can be explained by the fact that many nodes are of a very large degree and therefore if one node is affected, many others are directly affected all at once. On the other hand, at the end there are only a few knots left that have only one relationship and are difficult to reach.

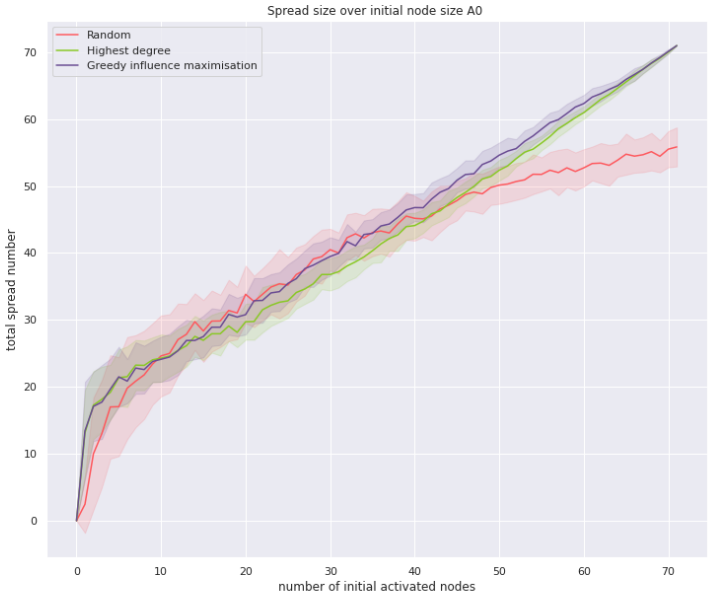

Figure 7: Total spread number over initial size A0 with p=.1 and 24 simulations per run

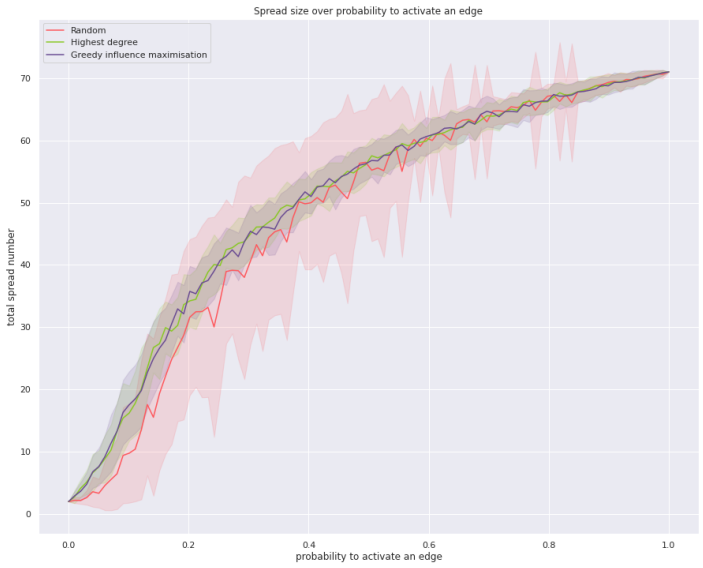

Figure 8: Total spread number over probability to activate an edge with k=.03 and 24 simulations per run

4 Appendix

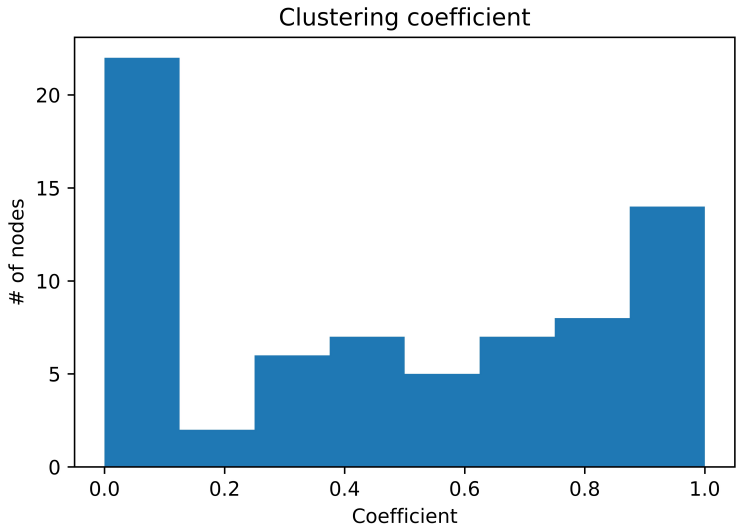

Figure 9: Repartition of the clustering coefficient

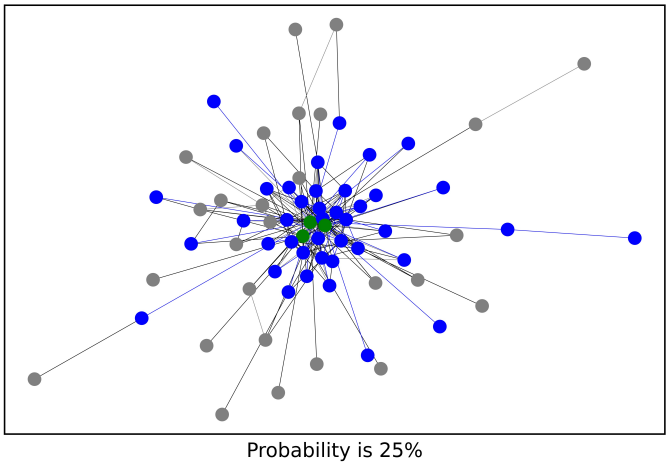

Figure 10: Animation of the Spread's evolution of the Independent Cascade Model by increasing probability (with starting nodes A0 given by the greedy algorithm). LINK HERE