LINMA2472 – Algorithms in Data Science

HW 1 – module "Networks"

This is the first part of the first assignment (the next part will follow after this week's lecture). This assignment is to be completed in groups of 2 or 3, please form your groups on moodle (using the activity called "Group choice for assignment 1"). If you need help to look for teammates, use the "Teammate finder" forum. If you have any practical questions, please email us on remi.delogne@uclouvain.be .

The deadline for this assignment (and the next part to come) is on Sunday the 1st of November at 23h59.

Assignment 1: co-occurrence network of characters

Please choose one of the following options:

- Find an appealing *book* (for example, use the Project Gutenberg https://www.gutenberg.org/ to find the text), parse the textual information in order to reconstruct the co-occurrence network of characters. For example, two characters can be linked if they appear in the same paragraph. Examples for inspiration: Lord of the Rings, War and Peace, Les Miserables, etc.
- Find a *screenplay* from your favorite movie (there are many resources can be found by Googling, for example, https://thescriptsavant.com/free-movie-screenplays-am/). Convert the .pdf to text using any online tool and parse the textual information to reconstruct the co-occurrence network of characters, where two characters can be linked if they appear in the same scene. Scenes are usually distinguished in bold notation. Examples for inspiration: Harry Potter, Lord of the Rings, Zootopia, etc.

The only requirement here would be to choose a book or a movie with many characters (ideally more than 50). Tools for text processing were discussed on the first lecture.

- Find degree assortativity of the network and perform community detection using Louvain algorithm. Visualise the results. What can you tell from the them?
- Write the code for k-core decomposition (do not use the preprogrammed instance in *networkx*) and apply it to the network. What can you infer from it?
- Generate the preferential attachment (Barabasi-Albert) network with similar average degree and size. Perform same operations on this network. Describe any differences or similarities you can spot.

Report guidelines:

- Write in a concise and structured manner. No long sentences, only relevant information.
- You may present your data and the preprocessing steps, but remember that this isn't the main goal of the report
- Any numerical result that can be presented in a table should be presented so.
- Round numbers up to 3rd digit, unless it's really necessary. Don't copy-paste 10 digits floats.
- Plots must be easy-to-read. Must include labels on axes, legend if more than one curve is shown, title or a caption, explaining what the plot is about.
- Network properties (k-core shell, community index, etc) can be visualized in color. When doing so, it's a good practice to add a colourbar (k-core shell) or a summary of each or most representative communities.