COMP30850 Assignment 1

Deadline: Thursday 16th February 2023

Overview:

The objective of this assignment is to construct and analyse network representations of <u>two</u> movie-related datasets from IMDb. These networks should model the *co-starring* relations between actors in the two datasets - i.e. the collaboration network of actors who appear together in the same movies.

The assignment should be implemented as a single Jupyter Notebook (not a script). Your notebooks should be clearly documented, using comments and Markdown cells to explain the code and interpret the results of your analysis.

Note: The final task in the assignment requires using the Gephi visualisation tool. The rest of the analysis should be performed within your notebook.

Tasks:

Complete the following five tasks:

1. **Network construction**

   - Download your two personal datasets, which are in a single ZIP file at:

     http://mlg.ucd.ie/modules/COMP30850/movies/<STUDENT_NUMBER>.zip

     For example, if your student number is 195023491, your ZIP file is at:

     http://mlg.ucd.ie/modules/COMP30850/movies/195023491.zip

   - Each dataset is represented as a JSON file, containing information about the actors and the movies in which they starred. For each dataset:

     a) Parse the JSON data and create an appropriate *co-starring* network using NetworkX, where nodes represent individual actors. You should justify your choice of network representation.

     b) Identify and remove any *isolated* nodes from the network.

2. **Network characterisation**

   - For each of the networks created in Task 1:

     a) Apply a range of different methods to characterise the structure and connectivity of the network.

     b) Apply different centrality measures to identify *important* nodes in the network. Clearly explain and interpret the results which are produced.

   - Based on the results from the above, discuss how the two networks are different (or similar) to one another in terms of their structure.

3. Network thresholding

- For each of the networks created in Task 1:

    - Examine and discuss the effects of applying different threshold levels to the network, to convert it from a *weighted* network to a new *unweighted* network.

4. Ego-centric analysis

- For each of the networks created in Task 1:

    - Select one of the *important* nodes in the network as identified in Task 2(b). Generate an *ego network* for this node. Characterise and discuss the structure of the new ego network.

5. Network visualisation

- For each of the networks created in Task 1:

    a) Export the network as a GEXF file.

    b) Load the GEXF file in Gephi. Use the layout functionality to produce one or more useful visualisations of the network. Export these visualisations as PNG files.

Guidelines:

- The assignment should be completed <u>individually</u>. All submissions will be subject to plagiarism checking. Any evidence of plagiarism will result in a 0 grade.

- The grade awarded will depend on the complexity of the analysis and level of detail, i.e. network construction, analysis, interpretation etc.

- Submit your assignment via the COMP30850 Brightspace page. Your submission should be in the form of a <u>single ZIP file</u> containing:

    1. Your Jupyter notebook (i.e. IPYNB file). In the notebook please clearly state your student number.

    2. The exported GEXF files for your two networks.

    3. The final PNG images for your two networks, produced with Gephi.

- Hard deadline: Submit by the end of 16th February 2023. Penalties will apply for late submissions:

    - 1-5 days late: 1 grade point deduction, e.g. B to B-

    - 6-10 days late: 2 grade point deduction, e.g. B to C+

    - Assignments will not be accepted later than 10 days without Extenuating Circumstances formally approved by UCD.