PROJECT 1: NETWORK SCIENCE-BASED ANALYSIS OF SCSE FACULTY COLLABORATION

CX4071 NETWORK SCIENCE

TOTAL MARKS: 100

Due Date: March 14, 2021 – April 16, 2021

PROJECT DESCRIPTION

Over the years, the School of Computer Science & Engg. (SCSE) of NTU has grown in reputation for the scientific contributions made by faculty members. This is postively reflected in various global ranking metrics for CS departments. In this open-ended project, the goal is to explore the following question: ***Can network science help us to understanding research collaboration among SCSE faculty members over time and, if possible, explain the reputation growth?*** Here we measure research collaboration as co-authorship among faculty members in scientific papers/articles.

In order to understand research collaboration of faculty members, we need to have access to the list of publications of each faculty. To this end, you should use the *DBLP computer science bibliography* (https://dblp.uni-trier.de/) to seek answer to the grand question. It is an on-line reference for bibliographic information on major computer science publications. It has evolved from an early small experimental web server to a popular open-data service for the whole computer science community. As of January 2021, *DBLP* indexes over 5.4 million publications, published by more than 2.6 million authors. Specifically, it contains the temporal history of publications of each author (e.g., institutions, year of publication, co-authorship, publication venue) including SCSE faculty members. In this project, your goal is to analyze this data source (you can download individual faculty member's data in XML format from DBLP), to answer following intriguing questions. For ease of reference, the list of SCSE faculty members and their rank, gender, management position held (Y/N), DBLP address, and key research area are provided to you as input (*Faculty.xlsx* file).

- What are the network properties of the SCSE faculty network? Note that the network should only contain SCSE faculty members as nodes. No other individual should be part of this network.
- How has the network and its properties evolved since year 2000? That is, the program should be able to analyze the network properties over time (at yearly granularity).

- Analyze the collaboration between faculty of different ranks (e.g., Professor vs Assistant Professor).
- Analyze the collaboration between faculty holding or held management position and non-management faculty.
- Analyze the collaboration between faculty of different areas in computer science (data management vs AI/ML)
- Given a set of faculty member as input, track their **collaborative properties** over time. Think how to define various collaborative properties.
- Are the **central** nodes of the network as measured using network properties (e.g., degree centrality, betweenness centrality) identical to **excellence** nodes? We define that a faculty is an **excellence** node if he/she has published in the top venue *frequently* (in the last 10 years or since his/her first publication if the first publication appears less than 10 years ago) in his/her respective area. Analyze and compare these two types of nodes. What insights can you draw from them? The list of top venues for different areas is given in *Top.xlsx* file. Also, carefully study the format used in DBLP to represent these venues (they may not be identical to the input file). Note that all these venues have affiliated workshops. You should **ignore all** workshop papers.
- Assume now SCSE would like to hire *at least* 1000 faculty members to handle growing demand of its CS/CE program (assume NTU has unlimited financial resources 😊). Select at least 1000 co-authors (it is up to you to determine how to select them) of the faculty members as potential hires and add them to the network. Analyze how the network properties of the modified network differ from the original faculty network.
- …..

Note that the question set is **not exhaustive** in order to facilitate unleashing of your creativity. You are free to pose additional questions that you think are relevant to this project.

Finally, to facilitate visualization of your insights and analysis create an interactive graphical user interface (GUI) that takes the data made available (faculty.xlsx, top.xlsx files) to you as input and enables you to visualize various questions you have posed on it. For these questions, you are free to take additional input from end users. Your software needs to be event-driven though the GUI. Note that the input files are configurable and hence you should not be hard coding anything.

Optional challenge: For those who wish to push further, you may take up the following challenge. You may note that the DBLP data of some faculty (e.g., Liu Zhiwei, Cheng Long) are not accurate as DBLP has failed to disambiguate their names with different

authors with the same name (i.e., namesake problem). Given such a faculty and its DBLP address as input, extract the correct publications and co-authors of this faculty.

DEVELOPMENT ENVIRONMENT

You must use **Python 3.0** in **Windows** environment for your project. You are free to use any publicly available libraries for your development.

SUBMISSION REQUIREMENTS

Your submission should include the followings:

- In order to facilitate grading, you should submit **four** program files: *interface.py*, *faculty.py*, *preprocessing.py*, and *project.py*. The file *interface.py* contains the code for the GUI. The *faculty.py* contains code for analyzing the faculty network and gaining insights on aforementioned issues. The *preprocessing.py* file contains code that takes DBLP information of faculty in XML format as input and constructs the **faculty network** for your analysis (Note from your lectures, choosing the correct network representation is a key task in network science). Lastly, the *project.py* is the main file that invokes all the necessary procedures from these three files. **Note that we shall be running the project.py file** (either from command prompt or using the Pychamp IDE) to execute the software. Make sure your code follows good coding practice: sufficient comments, proper variable/function naming, etc.
- **Softcopy report** containing details of the features supported by your software, analysis and insights of various questions related to research collaboration among SCSE faculty members. Lastly, you should also discuss limitations of the software (if any).
- **Peer assessment report** from each member of the team. Each individual member of a team needs to assess contributions of the group members. Details of peer assessment form will be provided closer to the submission date.
- More details related to the submission will be provided closer to the date.

Note: We give you a rolling deadline due to pandemic. You can submit your assignment earliest by March 14, 2021 and latest by April 16, 2021. However, you are only allowed to submit one version of your project. Late submission after April 16 will be penalized. Groups may be asked to demonstrate their projects after the submission deadline.