# PROJECT: NETWORK SCIENCE-BASED ANALYSIS OF CCDS FACULTY COLLABORATION

## SD6127 NETWORK SCIENCE

**TOTAL MARKS: 100**

**Due Date: April 18, 2025**

## PROJECT DESCRIPTION

The College of Computing and Data Science (CCDS) 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 CCDS 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 2025, *DBLP* indexes over 7.6 million publications, published by more than 3.7 million authors. Specifically, it contains the temporal history of publications of each author (e.g., institutions, year of publication, co-authorship, publication venue) including CCDS 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, a list of current and former CCDS 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). Some of the members data is incomplete or incorrect (like any real data). Hence, for these members your task will be to complete and clean the data before using it for analysis.

Your project should seek to answer the following questions:

- What are the network properties of the CCDS faculty network? Note that the network should only contain CCDS faculty members as nodes. No other individual should be part of this network. Can these properties occur by chance?

- 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)
- 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 can be found at https://csrankings.org/. 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 CCDS 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.

Additionally, you may implement additional components for ease-of-use of your software (this is not mandatory).

## DEVELOPMENT ENVIRONMENT

You must use **Python** and **Windows** environment for your project. You are free to use any publicly available libraries for your development or any additional visualization packages and software.

## SUBMISSION REQUIREMENTS

Your submission should include the followings:

- **Software:** In order to facilitate grading, you should submit at least **three** main program files: *faculty.py*, *preprocessing.py*, and *project.py*. 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 PyCharm IDE) to execute the software. Make sure your code follows good coding practice: sufficient comments, proper variable/function naming, etc.
- **Presentation and demonstration:** Each group shall present their solution and demonstrate their software on **Week 14**. Details related to presentation will be posted nearer to the date.
- All submissions will be through NTU Learn. The submission site will be opened closer to the deadline.

## GRADING POLICY

The technical content of the report and the software carries **50%** of the marks. The remaining **50%** will be for the presentation. Late submission will be penalized. You will lose 10 marks/day.