PROJECT 1: REPUTATION AND SUCCESS IN DATA SCIENCE

CZ4071 NETWORK SCIENCE

TOTAL MARKS: 100

Due Date: March 13, 2020 – April 17, 2020

PROJECT DESCRIPTION

In this open-ended project, the goal is to explore the following question: ***Can network science help us to quantify reputation and success in data science?*** This question is similar to the question posed in [1] in arts (discussed in lectures). Hence, you should read this article prior to exploring solution to the above question.

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 2020, *DBLP* indexes over 4.9 million publications (more than 5,800 conferences and 1,650 journals), published by more than 2.4 million authors. Specifically, it contains the temporal history of publications of each author (e.g., institutions, year of publication, co-authorship, publication venue) including data scientists. In this project, your goal is to analyze this data source (you can download it from https://dblp.uni-trier.de/xml/) to answer following intriguing questions:

- What network science measure reflects the prestige of venues/authors?
- How prestige of institutes and authors who publish in premium venues are correlated?
- What is the impact of network effect on the reputation and success of data scientists? We assume that a data scientist is **successful** if he/she can publish in premium venues (Tier 1) consistently. A successful data scientist has high reputation.
- How the location of institution a scientist belongs to plays a role in success?
- How likely an author can move from non-premium venues to premium venues in his/her career? How about the reverse scenario?
- How will the career of data scientists be? Does initial reputation of publication venues predict success?
- …..

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. In order to work with manageable size of data, you should **only focus** on the following publication venues in the data set (this list should be configurable through a configuration file) where Tier 1 represents premium venues (most prestigious):

- **International Conference on Management of Data (SIGMOD) – Tier 1** *(Not SIGMOD Record)*
- **Very Large Data Bases Conference (VLDB)/ Proceedings of the VLDB Endowment (PVLDB)** (PVLDB is formerly called VLDB) **– Tier 1**
- **Knowledge Discovery and Data Mining (KDD) – Tier 1**
- International Conference on Extending Database Technology (EDBT) – Tier 2
- IEEE International Conference on Data Engineering (ICDE) – Tier 2
- IEEE International Conference on Data Mining (ICDM) – Tier 2
- SIAM International Conference on Data Mining (SDM) – Tier 2
- International Conference on Information and Knowledge Management (CIKM) – Tier 2
- International Conference on Database Systems for Advanced Applications (DASFAA) – Tier 3
- Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) – Tier 3
- European Conference on Principles of Data Mining and Knowledge Discovery (PKDD) – Tier 3
- International Conference on Database and Expert Systems Applications (DEXA) – Tier 3

Note that the list of venues that you have considered should be clearly articulated in the report. Also, carefully study the format used in DBLP to represent these venues. Note that all these venues have affiliated workshops. You should **ignore all** workshop papers.

You can use any major university/institution ranking systems to find premium institutions. This should also be specified in the configuration file.

Finally, to facilitate visualization of your insights and analysis create a graphical user interface (GUI) that takes the DBLP data source as input and enables you to visualize various questions you have posed on it.

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*, *science.py*, *preprocessing.py*, and *project.py*. The file *interface.py* contains the code for the GUI. The *science.py* contains code for analyzing the DBLP network and gaining insights on reputation and success in data science. The *preprocessing.py* file contains code that takes DBLP dump in XML format as input and constructs the **relevant 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.
- Submit the **configuration file** you have used for conference venues and institution ranking (this file should be input to your software).
- **Softcopy report** containing details of the features supported by your software, analysis and insights of various questions related to the reputation and success in data science. Lastly, you should also discuss limitations of the software (if any).
- More details related to the submission will be provided closer to the date.

Note: Unlike previous years, we give you a rolling deadline. You can submit your assignment earliest by March 13, 2020 and latest by April 17, 2020. However, you are only allowed to submit one version of your project. Late submission after April 17 will be penalized. Groups may be asked to demonstrate their projects after the submission deadline.

References:

[1] Fraiberger et al., **Quantifying reputation and success in art.** *Science* 10, 2018