DATA.ML.430 Complex Networks
Project Task

Assignment

The Project Task will combine all the knowledge acquired during the course. The aim of the project is the ability of the student to analyze the network and recognize the areas for network applications. The students will select a network data set of interest to them, preferably from their own study field, build a network, analyze it and present it. The project is mandatory assignment of the course and it will give 4**0% of the final grade** for the course.

Part I

1. Register a **group of 2 (individual projects also possible)** students in Moodle using *Group registration.* One member of the group creates the group and sets a password for joining the group. Then, he or she gives the group number and password for the other student in the group, who then joins the group. **Deadline: 24.01.2021.** You can use the forum in Moodle to find a pair.
2. Think about the problems in your own field that can be solved and visualized as complex networks. Select the data for your network. The students are encouraged to collect the data and build the network themselves, e.g. from the internet sources by web scrapping the data, e.g. relationships between scientists in paper collaborations from Google Scholar. Those who will collect the network data themselves will receive extra points for the project. **Note:** you do not need to collect all of the possible links for your network, but you should rather apply some filtering to solve your problem and justify your choices.
3. Ready-to-use network data sets can be downloaded, e.g., from http://networkrepository.com/. The grades will not be reduced for such projects.
4. Requirements for the network:
    a. at least 500 nodes in the connected component
    b. the nodes have some sort of attributes (age, gender, race, …)
    c. data set is preferably from the own field of one of the students in the group
    d. network is built from a manually (not advised) or digitally (preferred) collected data or from the ready network data sets
    e. Example libraries for collecting HTML data with Python:
    https://www.crummy.com/software/BeautifulSoup/bs4/doc/
    https://gawron.sdsu.edu/python_for_ss/course_core/book_draft/web/web_intro.html
5. Write a brief description (no more than 1 page, pdf) of the data set that you are planning to collect and use and describe what are the nodes and links and why this data set is of interest to you. Submit your project proposals in Moodle before **02.02.2021 23:59.** (Only 1 submission per group, write your group number when submitting.)

Part II

1. Collect the data with interactions of at least 500 nodes and build your network. **Note:** you will have to explain in the final project submission how you collected the data and submit your data set with data collection and network mapping codes.

2. Submit the slides of the presentation with the first visualization of the network (e.g. from gephi) and the size of the network (number of nodes, number of links and average degree) by **05.03.2021 23:59**. (1 submission per group).
3. **Preliminary project presentations: 08.03.2021.** Present your collected dataset, the problem and the direction of your analysis during the lecture. No more than **3-4 minutes** per group, no more than **3 slides**.
4. Analyze the crawled social/complex network sample and compare the obtained network statistics with 2 random graph models having the same number of nodes and edges, i.e. Erdős-Rényi (ER) random network model and the Barabási-Albert (BA) preferential attachment network model.
5. Submit a project file (e.g. pdf) with your network project description, network statistics, analysis, visualizations for your crawled data and ER and BA synthetic graphs, and at least 2 implemented network problems out of 5 (a-e):
    a. **Community Discovery:**
    Analyze and compare the modular structure of the crawled network sample. The results of at least three different algorithms should be presented. Among all the possible choices you can use:
        - K-clique (available in NetworkX)
        - DEMON (code available at https://github.com/GiulioRossetti/demon)
        - Louvain (code available at http://perso.crans.org/aynaud/communities/)
        - Infomap (https://www.mapequation.org/)
    b. **Tie Strength:**
    Define a way to assess tie strength and analyze the impact of strong/weak ties on the connectedness and resilience of the crawled network.
    c. **Spreading:**
    Simulate a spreading process (SIS and/or SIR) both on the crawled data and on random graphs (i.e., ER and BA).
    d. **Threshold model:**
    Implement and evaluate a threshold model both on the crawled data and on random graphs (i.e., ER and BA).
    e. **Curiosity Driven:**
    Formulate a network problem specifically tailored upon the crawled data and solve it using social network analysis tools (network measures, community discovery, tie strength, spreading, link prediction…)
6. Submit your project report (no more than **10 pages** excluding images, pdf), your codes and your slides for the final presentation. **Note:** make sure your codes are properly documented and well-formatted!
7. Final project submission deadline: **15.04.2021 23:59**. All projects will be presented by the students during the last two lectures (**19.04.2021** and **26.04.2021**), **10-15 mins** per presentation.

Contacts & Feedback

Send your queries regarding the project, data set ideas, data collection or comments and feedback about the course to Kestutis Baltakys kestutis.baltakys@tuni.fi and Margarita Baltakiene margarita.baltakiene@tuni.fi.