DATA ANALYSIS II – FINAL TASKS

Select a data collection (either network data or vector data). The data collection must be real, either referential or custom, and must contain at least a thousand instances or thousand network vertices.

1. DATA ANALYSIS (10-20 points) (expected time required to carry out the task 4-6 hours)
The goal is the application of selected methods in the analysis of the real network with thousands to tens of thousands of vertices. The network can also be constructed from real vector data.

- The analysis involves the use of systems such as R, Gephi, Pajek, or existing methods implementation (you can also use your own).
- The methods used will be related to the areas discussed, particularly community detection, network development and network robustness, sampling.
- The output of the analysis will be a PDF document containing a description of the network being processed, including the source, where and how the data was retrieved, and a list and a brief description of the methods that were used.
- The outputs will be represented by tables with statistics and suitably chosen visualizations of outputs (networks, distribution, etc.). The outputs of the methods will be analysed and interpreted in order to find characteristics that cannot be obtained without applying the methods of analysis.

2. IMPLEMENTATION TASK (12-24, points) (expected time required to carry out the task - 4 - 6 hours)
The goal is to implement a program that works with a selected algorithm or algorithms discussed at lectures (or similar from the lecture areas).

- The program will include a simple user interface that allows you to enter the parameters of the algorithm, respectively specify the network that will be processed by the algorithm.
- The output of the program will be a text file or (report) containing algorithm settings and other information. E.g. for generating networks, sampling and community detection, this information will be the statistical properties of the networks (frequency, distribution, etc.).
- For the calculation of the statistical properties it is possible to incorporate the implementation from the first semester, resp. calls to R system libraries (or other libraries of your choice). Implementations will be designed in general and with regard to principles providing separate interfaces.
- It is assumed to work with networks with thousands to tens of thousands of vertices.

Data Sources

1. Network data

- https://snap.stanford.edu/data/
- http://vlado.fmf.uni-lj.si/pub/networks/data/
- http://www-personal.umich.edu/~mejn/netdata/
- http://dblp.uni-trier.de/xml/
- http://www.weizmann.ac.il/mcb/UriAlon/download/downloadable-data
- https://www3.nd.edu/~cone/software_data.html
- http://www.imsc.res.in/~sitabhra/research/neural/celegans/index.html
- http://www.cs.toronto.edu/~tsap/experiments/download/download.html
- http://toreopsahl.com/datasets/
- http://math.nist.gov/~RPozo/complex_datasets.html
- Etc.

2. Vector Data

- https://www.kaggle.com/datasets
- In the Weka format http://archive.ics.uci.edu/ml/datasets.html, http://www.hakank.org/weka/
- https://www.kdnuggets.com/datasets/index.html
- https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- etc.