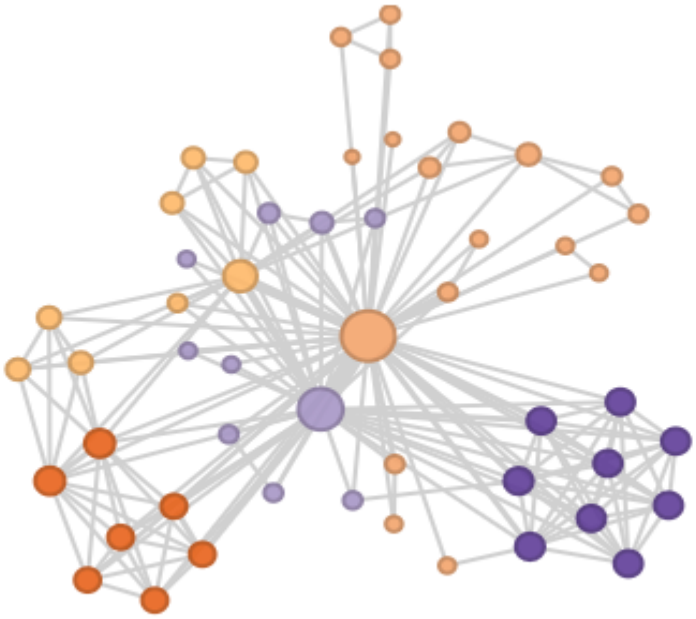
ORF 387: Networks

Project component

Instructor: Elizaveta Rebrova

This is about a part of the class when you work on **a project of your choice**, in the general topic of **networks**. This document has 3 parts: (1) logistics, (2) about the projects, and (3) additional links to help you with data and ideas.



Thanks to Micklos Racz for the initial version of this document.

1. Logistics

Scope: the project accounts for 25% of the course grade, so it has the same weight as all six HW sets. Accordingly, the amount of work that you put in it should be roughly what you would for six HW sets. Of course, there is no upper limit, but efforts of this order will be considered as fully satisfying the project component.

Milestones: there will be three milestones throughout the semester:

1. Project proposal, due Wednesday, October 5. This should be one full page: half page background / setup, and half page plans and goals.
2. Progress report, due Wednesday, November 9.
3. Final report, due Friday, December 16 (Dean's Date).

Collaboration: You are encouraged to work in groups of 2-3 people. More than 3 people in the group are not allowed. You should form groups before the initial project proposal, and submit the project proposal jointly: each of you submits a copy of the same document on Gradescope. Consider using Ed discussions and precepts for collaborator search!

Submission process: The main document is a pdf submitted via the Gradescope, experimental results are included in it as figures and tables. You can also include links to the code if desired, but the main document should be a pdf.

Coding and collaborating: For analyzing data sets, you may use the environment of your choice. Python and, in particular, Jupyter notebooks would be a great choice if you are unsure. Networkx is a nice Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. You can also use Jupyter notebooks through Google Collab for synced codes with your collaborators (warning: it might be slower than running things on your computer). You are encouraged to use Overleaf to latex joint project reports (you will likely need to learn this one anyway, at least for your senior thesis, if you are at ORFE). It is also a good practice to use git for the code version control if several people write the code in your group (which is expected), but it might result in a certain learning curve if you have never used it before. Another thing that might take too much time, and this class does not have dedicated time to address it, is a very natural computational challenge coming from the huge size of many real-world networks.
*So, every group should make a conscious decision about the **size of data** and **coding practices** (language, packages, version control system etc) that result in best and most stable performance of the model, but learning and trouble-shouting them would not take too much time from the **core of your project**, namely, doing **conceptually interesting network analysis**.*

2. About the projects

The projects can have various forms, including, but not limited to:

1. **Data analysis.** A natural type of project is to look at some existing data set(s) and analyze them in order to answer some interesting questions. There are many network data sets online, or you can start by checking out some of the links we compiled below.
2. **Simulation studies.** Another natural type of project is to study processes on networks or dynamics of networks via simulation. This makes sense in situations where data is hard or impossible to obtain, and one can start with some initial data and model the evolution of it. Typical examples of such projects include modeling the spread of a disease or a news topic.
3. **Reproducing/replicating existing research papers.** A course project is a perfect opportunity to attempt to reproduce experiments from an existing research paper. One learns a tremendous amount from trying to do so: it is a way to dive deep into one particular approach, and to learn-by-doing all its caveats (that are so much harder to notice just by reading a paper).
   One should probably be slightly more careful with the abundance of papers available online for the balance between hardness to read/access the data and the importance and relevance of the proposed method. We encourage you to briefly discuss the paper you consider with your AI to see if it is a good fit. We will also add a list of some popular networks papers below.
4. **Your ideas!** This course covers a broad range of topics and of course we do not have time to cover everything. The project is a perfect opportunity for you to pursue your personal interests.

3. Additional links

Network data repositories:

- Stanford Network Analysis Project **http://snap.stanford.edu/**
- The Colorado Index of Complex Networks **https://icon.colorado.edu/#!/**
- Network Dynamics project **https://dynamics.cs.washington.edu/data.html**
- Network datasets from Barabasi's online textbook **http://networksciencebook.com/translations/en/resources/data.html**
- Network data compiled by Mark Newman **http://www-personal.umich.edu/~mejn/netdata/**
- Awesome Public Datasets grouped by topic/application **https://github.com/awesomedata/awesome-public-datasets**

Single project/theme websites:

- Connectome project data **https://neurodata.io/project/connectomes/**
- Air transportation network OpenFlights **http://konect.cc/networks/openflights/**
- Lexical database of English WordNet **https://wordnet.princeton.edu**
- Facebook friendships, citation and scientific collaboration networks **http://wwwlovre.appspot.com/support.jsp**
- One company e-mail communication network **https://www.ii.pwr.edu.pl/~michalski/index.php?content=datasets#manufacturing**
- Python packages dependency analysis **http://kgullikson88.github.io/blog/pypi-analysis.html**
- HIV transmission network **https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/22140**

Papers:

This list is also open, you can take a paper you are interested in! Show it to us first to discuss the fit. If original data is hard to get, another interesting task is to test proposed findings on other conceptually similar dataset.

1. *"Collective dynamics of ' small-world' networks',* Watts and Strogatz; Nature 1998 https://www.nature.com/articles/30918

2. *"The structure of scientific collaboration networks"*, Newman; PNAS 2001 https://www.pnas.org/content/pnas/98/2/404.full.pdf

3. *"Network enhancement as a general method to denoise weighted biological networks",* Wang et al.; Nature Communications, 2018 https://cs.stanford.edu/people/jure/pubs/ne-natcom18.pdf

4. *"Hierarchical structure and the prediction of missing links in networks"*, Clauset, Moore, Newman; Nature 2008 https://www.nature.com/articles/nature06830 (and with appendix https://arxiv.org/pdf/0811.0484.pdf)

5. The story of scale-free networks (https://en.wikipedia.org/wiki/Scale-free_network):
    (a) *"Emergence of Scaling in Random Networks"*, Barabasi and Albert; Science 1999 https://science.sciencemag.org/content/sci/286/5439/509.full.pdf (direct link https://arxiv.org/pdf/cond-mat/9910332.pdf or https://barabasi.com/f/67.pdf)
    (b) *"Scale-free networks are rare"* Broido, Clauset; Nature Communications 2019 https://www.nature.com/articles/s41467-019-08746-5
    (c) Thread with a link (PDF) to a rebuttal by Albert-László Barabási https://twitter.com/barabasi/status/971068797342879745