

# TEXTBOOK PROJECTS AT THE NETWORKS COURSE

## General description of the projects

The term projects are aimed at providing a test ground for the skills and knowledge you picked up during the course. The textbook projects in most cases correspond to numerical studies of one of the problems discussed during the course. To fulfill a textbook project you have to go through basically the following steps:

- understand the project posed, (i.e., dig up the slides from the course related to the project, look up in Wikipedia, etc.),
- write/download the program(s) needed to carry out the project,
- calibrate and run your programs: in some cases parameters have to be tuned, in other cases the same routine has to be re-run several times with different parameters, etc.
- prepare figures showing the results,
- prepare your report (4-8 pages, including plots / visualizations). In your report you should
  - explain the project to reader,
  - outline how you solved it,
  - show the results.

## The projects

1. Write a program that can randomize a network by preserving the degree sequence and enhancing or decreasing the frequency of a chosen motif consisting of 3 nodes. (The user should be able to change the motif and to decide whether its frequency should be increased or decreased).

Demonstrate how the program works on the network examples used in the practical sessions: show a few figures of the networks before and after the randomization, and show a figures of the changing of some network parameters during the randomization.

(Project advisor: Dániel Ábel or Gergely Palla)

2. Compare random graphs generated with the Holme-Kim model and with the configuration model. Set the number of nodes and links to be the same, and also the degree distribution to be the same. Is there any difference between the behavior of the clustering coefficient, the behavior of the average nearest neighbors degree, and the closeness?

In the presentation show pictures of the networks generated by the two different algorithm, and also present figures showing the examined statistics (e.g., the clustering coefficient as a function of degree, the  $k_{nn}$  as a function of degree, etc.).

(Project advisor: Dániel Ábel or Gergely Palla)

3. Write a program that can extract the triad motif significance profile of an input network. Compare the motif profile of the networks used in the practical session.

In the presentation (beside the motif significance profiles) show how the frequency of a given motif is changing when randomizing the network, and also snapshots of the networks before and after the randomization.

(Project advisor: Dániel Ábel or Gergely Palla)

4. Study the percolation transition of the E-R graph numerically. How does the size of the largest connected component scale with the system size at the critical point? How does the size of the largest connected component grow with  $p$  at the critical point?

In the presentation show pictures of the generated graphs both below and above the critical point, and also figures revealing the scaling of

the largest connected component at  $p_c$ .

(Project advisor: Dániel Ábel or Gergely Palla)

5. Study the average shortest path length in the W-S model numerically. If  $p$  is kept fixed and  $N$  is varied, how does the critical  $N_c$  (where the small world effect takes place) scale with  $p$ ? Reproduce the data collapse for  $\langle l \rangle$  discussed in the course.

Include pictures of the generated graphs both below and above the critical point in the presentation, and show figures depicting the scaling and the data collapse.

(Project advisor: Dániel Ábel or Gergely Palla)

6. Study epidemics numerically in the SIS model on computer generated directed networks. What is the difference between spreading on Erdős-Rényi, and on scale-free networks? In case of scale-free networks, examine whether it makes a difference if only the in- (or out-) degree distribution is scale-free? Reproduce the main figures from the following paper: Romualdo Pastor-Satorras, Alessandro Vespignani, “Epidemic dynamics and endemic states in complex networks”, *Phys. Rev. E* **63**, 066117 (2001). (free download from <https://arxiv.org/pdf/cond-mat/0102028>)

(Project advisor: Dániel Ábel or Gergely Palla)

7. Compare the networks used in the practical sessions according to the resilience against random breakdown and attack! Analyze the problem of resilience on the corresponding random graphs as well. Use an ensemble of ER graphs and also randomized networks with preserved degree distribution.

(Project advisor: Dániel Ábel or Gergely Palla)

8. Implement the Girvan-Newman algorithm and test it on the “classical” computer generated test bed and on the networks used in the practical sessions.

Details:

- The computer generated test bed is defined as follows: consider 4 groups of nodes with 32 members each, these correspond to the communities the algorithms should find. The links are drawn between the nodes at random, however the average number of links going from one node to other nodes in its own community

is given by  $z_{\text{in}}$ , whereas the average number of links going to the rest of the network is given by  $z_{\text{out}}$ , and  $z_{\text{in}}$  and  $z_{\text{out}}$  always add up to  $z_{\text{in}} + z_{\text{out}} = \langle k \rangle = 16$ . When  $z_{\text{out}}$  is small and  $z_{\text{in}}$  is close to 16, the algorithms have an easy job, the communities can be spotted even by eye, since they the links are very dense inside and occur very rarely in between the communities. However, as  $z_{\text{out}}$  is increased the communities become less and less well defined, and around  $z_{\text{out}} = 8$  the graph becomes homogeneous.

- The evaluation of the community finding results is based on a matching between the predefined 4 groups and the found communities. Choose the matching which gives the highest number of correctly classified nodes. (E.g., index the original groups by  $A, B, C, D$ , then index the found communities also with  $A, B, C, D$ , and search for the permutation of these indices maximizing the number of nodes on which the two community index is the same).
- In the presentation plot the ratio of correctly classified nodes over the total number of nodes as a function of  $z_{\text{out}}$ . Also show figures about the found communities in the test graph and in the real networks.

(Project advisor: Dániel Ábel or Gergely Palla)

9. Implement a greedy optimization of the modularity  $Q$  and test it on the “classical” computer generated test bed and on the networks used in the practical sessions.

The outline of the greedy optimization is the following:

- In the initial state each node is considered to be in a separate community.
- At each step merge a pair of communities into a single community. Select the pair for which the increase in  $Q$  resulting from the merging is maximal.
- Carry on until the network contains only a single community. Choose the partition with highest  $Q$  value in the process.

The further details are the same as in the previous project.

(Project advisor: Dániel Ábel or Gergely Palla)

10. Study random walks on Watts-Strogatz networks. Generate large W-S graphs with fixed  $N$  and varying rewiring probability  $\beta$ , and record

the number of visited nodes  $s$  for a random walker as a function of the time-steps. On a linear chain we would expect  $s(t) \sim \sqrt{t}$ , whereas in an E-R random-graph  $s(t) \sim t$ , (at least as long as  $s \ll N$ ). Assume that in case of a W-S graph we have  $s(t) \sim \sqrt{t} f(t\beta^\alpha)$ , where  $f(x)$  is some universal function, and  $\alpha$  is a yet unknown exponent. Plot  $s(t)/\sqrt{t}$  as a function of  $t\beta^\alpha$ , and find the  $\alpha$  exponent for which the datapoints collapse onto the universal  $f(x)$  function.

(Project advisor: Dániel Ábel or Gergely Palla)

11. Study the extent of hierarchy in random graphs. Hierarchical organization is a ubiquitous feature of a wide range of systems from cells through living organizations to animal flocks and the human society. There were several different hierarchy measures introduced in the scientific literature for quantifying the extent of hierarchy in the structure of directed networks. The task is to examine the behavior of a couple of these measures around the critical point of percolation in the directed Erdős–Rényi graph.

The directed E-R graph is very similar to the undirected one, i.e., directed links are introduced independently with probability  $p$  between  $N$  nodes, (where we consider both directions for every pair). If  $p$  is very low, the obtained directed random graph is dispersed and is consisting of many disconnected components, whereas if  $p$  is large enough, a giant (weakly connected) component is emerging, which will eventually “swallow” all smaller component if  $p$  is increased further.

The task is to monitor the value of three hierarchy measures as functions of the average degree around the critical point of the percolation using numerical simulations. Since the structure of the giant component is tree-like in the undirected E-R graph at the critical point, and trees are highly hierarchical objects, it is an interesting question whether we observe any peaks in the hierarchy measures for the directed E-R graph at the percolation transition? The hierarchy measures to be studied are the following:

- the fraction of links not participating in any directed cycles,
- the Global Reaching Centrality introduced in E. Mones, L. Vicsek and T. Vicsek, Hierarchy measure for complex networks. *PLoS ONE* **7**, e33799 (2012).
- the Random Walk Hierarchy introduced in D. Czégel and G. Palla, Random walk hierarchy measure: What is more hierar-

chical, a chain, a tree or a star? *Scientific Reports* **5**, 17994 (2015).

(Project advisor: Gergely Palla)

12. There are several indications that hidden metric spaces might exist behind the often observed general properties of real networks. A remarkable network model exploiting this idea is the Popularity-Similarity-Optimization (PSO) model, where nodes are placed one by one on the Poincaré disk representation of the 2D hyperbolic plane with a logarithmically increasing radial coordinate and a random angular coordinate, and links are drawn with probabilities determined by the hyperbolic distance between the node pairs. In vague terms, the degree of nodes is determined by their radial coordinate (lower distance from the origin corresponds to larger degree), and the angular proximity of the nodes can be interpreted as a sort of similarity, where more similar nodes have a higher probability to be connected. Random graphs generated this way usually exhibit a scale-free degree distribution, high clustering coefficient and the small world property at the same time.

The goal of the project is to implement the PSO model on the computer based on the original paper by F. Papadopoulos et al, <https://doi.org/10.1038/nature11459> (A free version is available on the arXiv: <https://arxiv.org/abs/1106.0286>)

Generate random graphs according to this approach (up to a size with at least 1000 nodes) and study their properties, and prepare plots proving that they are indeed small world, highly clustered and scale-free. Provide layouts of the obtained graphs both on the Poincaré disk and using Cytoscape.

(Project advisor: Gergely Palla)