Homework #1

The homework does not require a lot of writing but may require some thinking. It does not require a lot of processing power but may require efficient programming. It accounts for 12.5% of the course grade. Any questions or comments should be posted on Piazza.

Submission details

This homework is due on **March 29th** at 11:59pm. It must be submitted through (1) Gradescope (entry code **JKX87N**) and (2) eUcilnica. (1) Submission to Gradescope should include answers to all questions, each on a separate page, which may also demand pseudocode, proofs, tables, plots, diagrams and/or other. (2) Submission to eUcilnica should include this cover sheet with signed honor code and all the programming code used to complete the exercises (preferably in `.py` format). The homework is considered submitted only when *both* (1) and (2) have been submitted. Failing to include the honor code in the submission will result in **10% deduction**. Failing to submit all the developed code will result in **10% deduction**.

Honor code

Students are strongly encouraged to discuss the homework with other classmates and form study groups. But each student must then solve the homework by herself/himself without the help of others and should be able to redo the homework at a later time. In other words, students are encouraged to collaborate, but should not copy from one another. Referring to any solutions obtained from classmates, course books, previous years, found online, AI tools or other is considered an honor code violation. Also, stating any part of the solutions in class or on Piazza is considered an honor code violation.

Any violation of the honor code will be reported to the **faculty disciplinary committee** and vice dean for education.

SID: ____________________________

Full name: ____

Study group: ____

I understand and accept the honor code.

Signature: ____________________________

1 Connected components (15%)

Assume a simple undirected graph with n nodes, m edges, and c connected components. Show that the following two inequalities hold. *(Use mathematical induction for the first inequality and simple reasoning for the second.)*

$$n - c \leq m \leq \binom{n-c+1}{2}$$

Using these two inequalities, provide a criterion for m that *ensures* a connected graph. Is the criterion practically useful for real networks?

What to submit?

Prove both inequalities ($2 \times 5\%$). Provide a criterion for m (2.5%). Give a brief answer to the question (2.5%).

2 Node linking probability (12.5%)

Consider a graph model where the edges are placed independently between each pair of nodes i and j with the probability p_{ij} proportional to

$$p_{ij} \sim v_i v_j,$$

where $v_i \geq 0$ is some number associated with node i.

First show that the expected node degree $\langle k_i \rangle$ is proportional to v_i. *(Show $\langle k_i \rangle = C v_i$ for some constant C.)* Next, derive an exact expression of p_{ij} in terms of *only* the degree sequence $\{k\} = k_1, k_2, \ldots, k_n$ and recognize the result.

What to submit?

Show $\langle k_i \rangle$ is proportional to v_i (5%). Derive an expression of p_{ij} in terms of $\{k\}$ (5%). Comment on the result (2.5%).

3 Random node selection (7.5%)

Erdős-Rényi graph model [ER59] requires efficient implementation of a random selection of nodes, which can be easily achieved with most network representations. More realistic models [BA99] require more sophisticated random selection procedures and associated network representations.

Design an algorithm that does not select the nodes uniformly at random, but proportional to their degree. More precisely, the node i should be selected with the probability $\frac{k_i}{2m}$, where k_i is its degree and m is the number of edges. The algorithm must run in constant time $\mathcal{O}(1)$, while you can assume *any* standard network representation. *(Think more about the network representation than the algorithm.)*

What to submit?

State the network representation (2.5%). Describe the algorithm or provide a pseudocode (5%).

4 Weak & strong connectivity (20%)

Depth-first search is the most efficient algorithm for finding connected components of undirected networks. What would the same algorithm find in a directed network if one would follow the links in any direction? What would the algorithm find if one would follow the links only in the right direction? What would the algorithm find if one would follow the links only in the opposite direction?

Based on your answers design an algorithm for finding strongly connected components in directed networks. Implement the algorithm and find strongly connected components of Enron e-mail communication network [KY04]. *(You should not use the Kosaraju algorithm.)*

Compute the number of strongly connected components and the size of the largest one. Are the results expected?

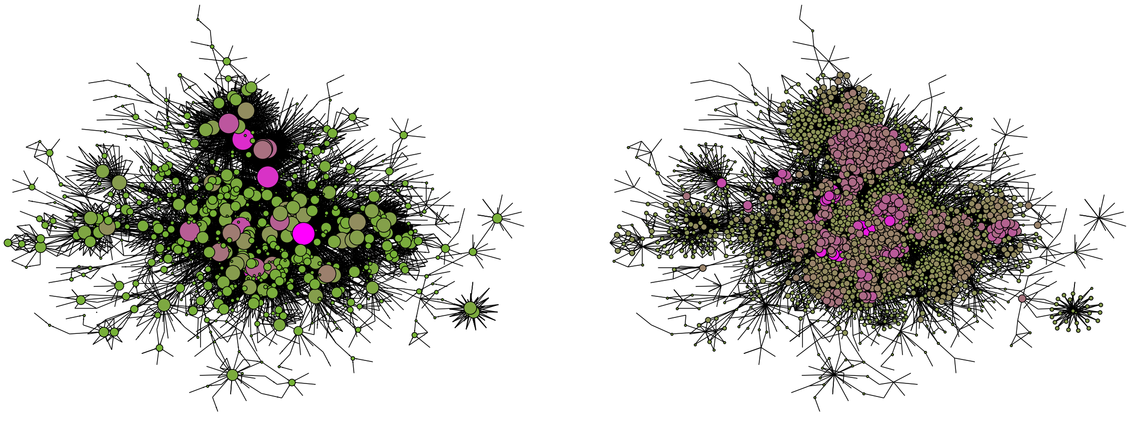
What to submit?

Give brief answers to the first three questions ($3 \times 2\%$). Provide a pseudocode of the algorithm or a printout of the implementation (4%). State the number of strongly connected components and the size of the largest one ($2 \times 4\%$). Comment on the results (2%).

5 Is Java software scale-free? (20%)

Consider a software class dependency network [ŠB11] representing Lucene search engine (Figure 1). This is a directed network where node i links to node j if the software class represented by i depends on or *uses* the class represented by j.

Compute the degree distribution p_k and plot it on a log-log plot by representing each distinct (k, p_k) with a single dot, $k > 0$. *(Transformation to logarithmic axes should be done by your plotting software.)* Next, compute also the in-degree distribution $p_{k^{in}}$ and the out-degree distribution $p_{k^{out}}$, and superimpose them on the same plot using dots of different colors.



(a) Colors show in-degrees k^{in}

(b) Colors show out-degrees k^{out}

Figure 1: **Lucene class dependency network**

Compare all three degree distributions $p_{k^{\cdot}}$ and highlight the differences between them. Do the distributions appear to be scale-free following a power-law $k^{-\gamma_{\cdot}}$? For the distributions not to appear like a power-law, reason why. For the distributions that do seem to follow a power-law, reason why and compute their power-law exponent $\gamma_\cdot$ using the maximum likelihood formula [Bar16]

$$\gamma_\cdot = 1 + n_\cdot \left[\sum_{i=1}^{n} \ln \frac{k_i^\cdot}{k_{min}^\cdot - \frac{1}{2}} \delta(k_i^\cdot \geq k_{min}^\cdot) \right]^{-1},$$

where $k_i^\cdot$ is the $\cdot$-degree of node i, $k_{min}^\cdot \geq 1$ is some reasonable choice for the minimum degree cutoff and $n_\cdot \leq n$ is the number of nodes thus considered.

What to submit?

Draw a plot with three distributions ($3 \times 2\%$). Reason whether the distributions are seemingly scale-free and why ($3 \times 2\%$). Compute $\gamma_\cdot$ of scale-free distributions for $k_{min}^\cdot = 5$ (4%). Provide a printout of all the code used to solve the exercise (4%).

6 Five networks problem (20%)

You are given five networks in edge list format.

- Flickr social affiliation network
- IMDb actors collaboration network
- Small part of a university Web graph
- Sample of computer science citation network
- Realization of the Erdös-Rényi random graph

All networks were reduced to their largest (weakly) connected component and represented with simple graphs. *(Notice that two of the networks are directed.)*

Your task is to figure out which network is which by examining its structure. Describe every step of your reasoning and/or provide necessary computations.

What to submit?

State which network is which by referring to their filenames and support your reasoning ($5 \times 4\%$). Provide a printout of all the code used to solve the exercise.

7 Who to vaccinate? (5%)

The probability that an individual would spread some viral disease through her/his social network is proportional to k^2 [New10], where k is the degree of the corresponding node.

Consider two immunization schemes for preventing the spread of diseases. In the first scheme, you randomly select some number of individuals and vaccinate them. In the second scheme, you first select the same individuals, but then rather vaccinate a random acquaintance of theirs.

Which of the two schemes will provide better immunization? Why?

What to submit?

Give brief answers to both questions ($2 \times 2.5\%$).

References

[BA99] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[Bar16] A.-L. Barabási. *Network Science*. Cambridge University Press, Cambridge, 2016.

[ER59] P. Erdős and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[KY04] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 217–226, Pisa, Italy, 2004.

[New10] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 2010.

[ŠB11] Lovro Šubelj and Marko Bajec. Community structure of complex software systems: Analysis and applications. *Physica A*, 390(16):2968–2975, 2011.