Homework #1

This homework is complete and will not be changed. The homework does not require a lot of writing, but may require a lot of thinking. It does not require a lot of processing power, but may require efficient programming. It accounts for 12.5% of the course grade. All questions and comments regarding the homework should be directed to Piazza.

Submission details

This homework is due on **March 21st** at 3:00pm, while late days expire on **March 24th** at 3:00pm. The homework must be submitted as a hard-copy in the submission box in front of R 2.49 and also as an electronic version to eUcilnica. It can be prepared in either English or Slovene and either written by hand or typed on a computer. The hard-copy should include (1) this cover sheet with filled out time of the submission and signed honor code, (2) short answers to the questions, which can also demand proofs, tables, plots, diagrams and other, and (3) a printout of all the code required to complete the exercises. The electronic submission should include only (1) answers to the questions in a single file and (2) all the code in a format of the specific programming language. Note that hard-copies will be graded, while electronic submissions will be used for plagiarism detection. The homework is considered submitted only when both versions have been submitted. Failing to include this honor code in the submission will result in **10% deduction**. Failing to submit all the developed code to eUcilnica will result in **50% deduction**.

Honor code

The students are strongly encouraged to discuss the homework with other classmates and form study groups. Yet, each student must then solve the homework by herself or himself without the help of others and should be able to redo the homework at a later time. In other words, the 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 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. Finally, failing to name the correct study group members, or filling out the wrong date or time of the submission, is also considered an honor code violation. Honor code violation will not be tolerated. Any student violating the honor code will be reported to **faculty disciplinary committee** and vice dean for education.

Name & SID: _____

Study group: _____

Date & time: _____

I acknowledge and accept the honor code.

Signature: _____

Grand graph challenge (extras)

Consider an Erdős-Rényi random graph [ER59] with n nodes and $\frac{n \ln n}{8}$ links. What can you say about the expected fraction of nodes in the largest connected component S? Implement an efficient algorithm that constructs such a random graph and computes the fraction S. The algorithm can use any network representation and any method for computing connected components. It should merely output S for a given n. What is the size of the largest random graph in terms of n you are able to analyse on your computer within about one minute (Table 1)?

n	m	S	Graph	Search
100 000	143 912	?	0.06 sec	0.06 sec
1 000 000	1 726 939	?	2.42 sec	0.44 sec
10 000 000	20 147 620	?	26.2 sec	8.20 sec
12 500 000	25 533 186	?	49.1 sec	10.1 sec
15 000 000	30 981 676	?	1.38 min	18.3 sec
25 000 000	53 232 457	?	2.28 min	32.5 sec
50 000 000	110 797 085	?	3.78 min	2.23 min

Table 1: **Results for 2.3 GHz Intel Core i7 with 16 GB memory**

What to submit?

Submit your answers and results through the grand graph challenge submission option at eUcilnica. The grand graph challenge ends on **March 28th at 11:00am**.

1 Networkology (5 points)

1.1 Node degrees (0.25 points)

Figure 1 shows wiring diagrams of two networks with 256 nodes and the same average degree $\langle k \rangle = 4$. By observing the networks, what can you say about their degree sequences $\{k_i\}$ and degree distributions p_k? *(Either reason that the networks are the same or highlight the differences between the networks.)*

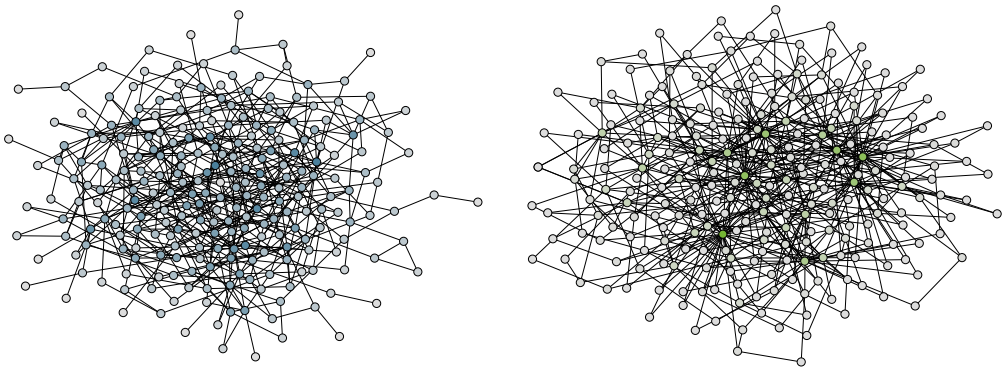

Figure 1: **Networks with** 256 **nodes and** $\langle k \rangle = 4$

1.2 Connected components (1 point)

Assume a simple undirected network with n nodes, m links and c connected components. Show that the following two inequalities hold. *(Use induction for the first inequality and simple reasoning for the second.)* Using these inequalities give a criterion for m that ensures a connected network. Is the criterion practically useful? Why?

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

1.3 Weak & strong connectivity (1.5 points)

In labs you saw an efficient algorithm for finding connected components of undirected networks. What would the same algorithm find in a directed network if one could follow the links in any direction? What would the algorithm find in a directed network if one could follow the links only in the proper direction? What would the algorithm find in a directed network if one could follow the links only in the opposite direction? Based on your answers, design an efficient algorithm for finding strongly connected components in directed networks. Implement the algorithm and find strongly connected components in Enron e-mail communication network [KY04]. Compute the number of strongly connected components and the size of the largest one. Are the results surprising? Why?

1.4 Node & network clustering (0.75 points)

In lectures you saw two measures of local density in a network, namely the average clustering coefficient $\langle C \rangle$ [WS98] and the network clustering coefficient C [NSW01]. Although the measures are similar, they are not equivalent. Course book [Bar16] describes a double star network for which $\langle C \rangle \to 1$ and $C \to 0$ when $n \to \infty$, where n is the number of nodes in the network (Figure 2). Think of another example network for which $\langle C \rangle \to$ const. and $C \to 0$ when $n \to \infty$. *(Study what gave this property in a double star network.)*

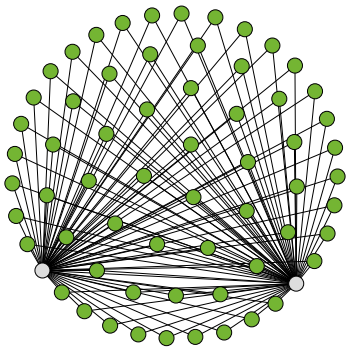

Figure 2: **Double star with node colors corresponding to clustering coefficient**

1.5 Effective diameter evolution (1.5 points)

In labs you saw an efficient algorithm for computing the diameter D of a connected undirected network. Note that D is a very sensitive measure as a single chain of nodes extending out of the main part of the network already gives large D. A smoothed version of D is called 90-percentile

effective diameter E_{90} [LKF07] that measures the maximum number of hops needed to reach 90% of the nodes in a network. Design an efficient algorithm for computing E_{90} of a connected undirected network. Implement the algorithm and compute E_{90} for citation networks of physics papers published by American Physical Society in the periods 2010-2011, 2010-2012 and 2010-2013. *(Although the networks are directed, treat them as undirected graphs. Note that these computations may easily take an hour.)* Are the results surprising? Why? Compute also the number of nodes n and the average degree $\langle k \rangle$ of all three networks, and discuss the results.

What to submit?

1.1 Briefly reason why both networks are the same or highlight the differences in $\{k_i\}$ and p_k (0.25 points).

1.2 Give proofs of both inequalities (2×0.25 points). Derive a criterion for m (0.25 points) and provide brief answers to both question (0.25 points).

1.3 Provide brief answers to all three questions (0.25 points). Give a pseudocode of the designed algorithm (0.5 points) and a printout of the implementation (0.25 points). State the number of strongly connected components in Enron network and the size of the largest one, and briefly comment on the results (2×0.25 points).

1.4 Provide brief description or a wiring diagram of the example network (0.25 points). Give proofs for the values of $\langle C \rangle$ and C (2×0.25 points).

1.5 Give a pseudocode of the designed algorithm (0.5 points) and a printout of the implementation (0.25 points). State n, $\langle k \rangle$ and E_{90} for all three citation networks and briefly comment on the results (3×0.25 points).

2 Graph models (3 points)

2.1 Random node selection (0.75 points)

Erdős-Rényi random graph model [ER59] requires an efficient implementation of a random selection of nodes, which can be easily achieved for most network representations. On the other hand, more realistic models [BA99] require more sophisticated random selection procedures. Design an algorithm that does not select nodes uniformly at random, but proportional to their degrees. Thus, node i is selected with probability $\frac{k_i}{2m}$, where k_i is its degree and m is the number of links. The algorithm should run in constant time $\mathcal{O}(1)$, whereas you can assume any standard network representation. *(Think more about the network representation than the algorithm.)*

2.2 Node linking probability (0.75 points)

Consider a random graph model in which links are placed independently between each pair of nodes i and j with probability p_{ij} proportional to $v_i v_j$, where v_i is some non-negative number associated with node i. First show that the expected node degree k_i is proportional to v_i. Next, derive an exact expression for p_{ij} in terms of the degree sequence $\{k_i\}$ and discuss the result.

2.3 Node degree distributions (1.5 points)

Represent a small part of the Facebook social network [VMCG09] as an undirected graph and compute its degree distribution p_k. Plot p_k on a doubly logarithmic or log-log plot by representing each distinct (k, p_k) with a single dot. *(Transformation to logarithmic axes should be done by your plotting software.)* Next, let n and m be the number of nodes and links, and $\langle k \rangle$ the average degree in Facebook network. Construct an Erdős-Rényi random graph [ER59] with parameters n and m, and again compute p_k. Superimpose p_k on the same plot using dots of different color as before. Also, compute the theoretical degree distribution of the Erdős-Rényi random graph $p_k \simeq \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}$ and plot it with a solid line. Finally, construct a random graph according to the preferential attachment model [BA99]. Start with a complete graph on $\lceil \langle k \rangle \rceil + 1$ nodes and add the remaining $n - \lceil \langle k \rangle \rceil - 1$ nodes one at a time. Each newly arrived node selects $\lceil \langle k \rangle / 2 \rceil$ of the existing nodes with probability proportional to their degrees and links to them. *(If you have not solved exercise 2.1, use linear roulette wheel algorithm for random selection by degree.)* Compute p_k of the preferential attachment graph and again plot it using dots of different color. Compare all four degree distributions p_k and highlight the differences among them.

What to submit?

2.1 State the network representation and give a pseudocode of the designed algorithm (0.5 points). Reason or prove why the algorithm gives the correct result (0.25 points).

2.2 Show that k_i is proportional to v_i (0.25 points). Derive an expression for p_{ij} and briefly comment on the result (0.5 points).

2.3 Draw a plot with four distributions and briefly discuss each result (3×0.25 points). Give a printout of the implementation of the preferential attachment model (0.5 points) and a printout of all the code used to compute p_k (0.25 points).

3 Node position (1.5 points)

You are given Slovenian highway network from 2010 with traffic loads at each location (Figure 3). For reasons of simplicity, the network is represented as a simple undirected graph. Your task is to find out which measure of node position could be utilized to best predict the traffic loads. You should consider at least three node measures, namely node degree k_i, node clustering coefficient $C_i = \frac{2t_i}{k_i(k_i-1)}$ [WS98] and node harmonic mean distance $\ell_i^{-1} = \frac{1}{n-1} \sum_j \frac{1}{d_{ij}}$ [New10]. Possibly the simplest way to achieve this goal is to compute Pearson or Spearman correlation coefficient between the values returned by some node measure and the actual traffic loads. Compute the correlation coefficient for each of the three measures. Are the results expected? Why? List also top ten locations according to the best node measure along with the computed values and the actual traffic loads.

What to submit?

State the values of the correlation coefficient and briefly discuss each result (3×0.25 points). List top ten locations according to the best measure (0.25 points). Print out any code you might have used or describe how you solved the exercise (0.5 points).

Figure 3: **Slovenian highways network with node colors corresponding to traffic loads**

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, 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.

[LKF07] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–41, 2007.

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

[NSW01] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.

[VMCG09] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. In *Proceedings of the ACM International Workshop on Online Social Networks*, pages 37–42, Barcelona, Spain, 2009.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.