Homework #1

This homework is complete and will not be changed. 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 13.3% of the course grade. Any questions and comments regarding the homework should be directed to Piazza.

Submission details

This homework is due on **March 19th** at 9:00pm, while late days expire on **March 22nd** at 12:00pm. The homework must be submitted through (1) Gradescope (entry code **NM8ZRM**) 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 other. (2) Submission to eUcilnica should include at least this cover sheet with signed honor code and all the programming code used to complete the exercises. The homework is considered submitted only when *both* parts have been submitted. Failing to include this honor code in the submission will result in **10% deduction**. Failing to submit all the developed code will result in **50% deduction**.

Honor code

Students are strongly encouraged to discuss the homework with other classmates and form study groups. Yet, each student must then solve the homework by her/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 or other material, 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: _____

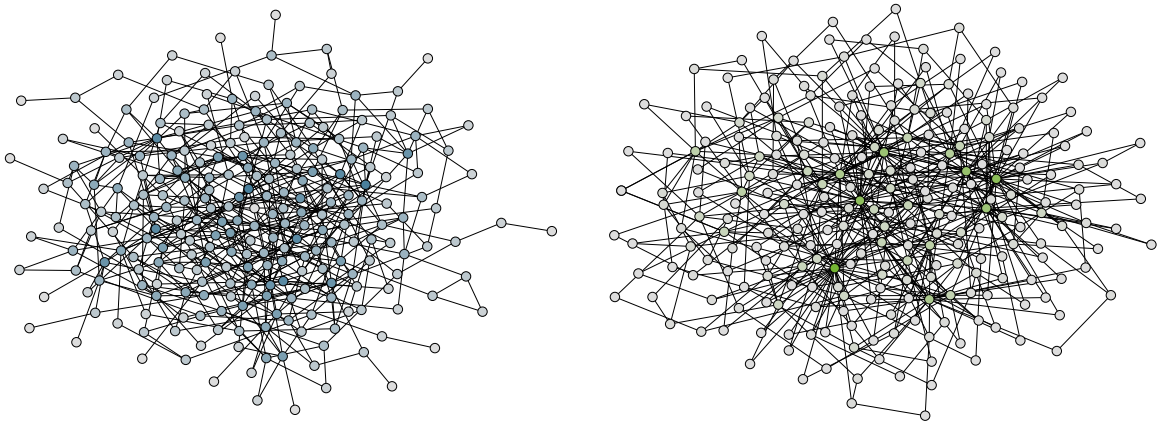

Figure 1: **Networks with** 256 **nodes and** $\langle k \rangle = 4$

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 $\{k\}$ and degree distributions p_k? *(First focus on high degree nodes or hubs.)*

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.)*

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

Using these two inequalities, give a criterion for m that ensures a connected network. Is this criterion practically useful for real-world networks?

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 would follow the links in any direction? What would the algorithm find in a directed network if one would follow the links only in the proper direction? What would the algorithm find in a directed network if one would follow the links only in the opposite direction?

Based on your answers design a 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 expected or surprising?

1.4 Node & network clustering (0.75 points)

In lectures you saw two measures of local density in a network, namely the average node clustering coefficient $\langle C \rangle$ [WS98] and 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 $\langle C \rangle \to c > 0$ and $C \to 0$ when $n \to \infty$. *(Study what gave this property in a double star network.)*

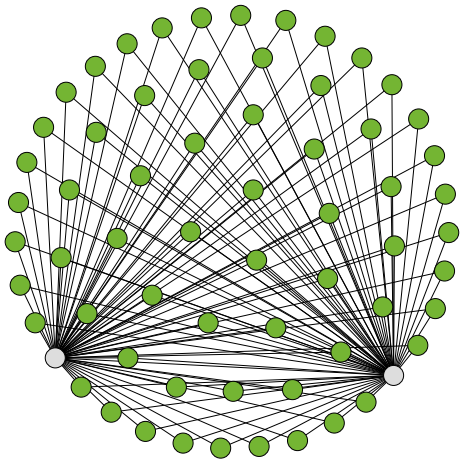

Figure 2: **Double star with node colors corresponding to clustering coefficient**

1.5 Effective diameter evolution (1.5 points)

In labs you saw an efficient algorithm for computing diameter d_{max} of a connected undirected network. Notice that d_{max} is a very sensitive measure as a single chain of nodes extending out of a large main part of the network already gives large d_{max}. A smoothed version of d_{max} is called 90-percentile effective diameter d_{90} that measures the number of hops at which 90% of the nodes in a network are reachable [LKF07].

Design an algorithm for computing d_{90} of a connected undirected network. Implement the algorithm and compute d_{90} of citation networks of physics papers published by American Physical Society in periods 2010-2011, 2010-2012 and 2010-2013. Are the results expected or surprising? *(Although the networks are directed, treat them as undirected graphs. Note that these computations might take some time.)*

What to submit?

1.1 Reason why both networks are the same or highlight the differences in $\{k\}$ 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 answer to the question (0.25 points).

1.3 Provide brief answers to all three questions (0.25 points). Give a pseudocode of the 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 wiring diagram of the example network (0.25 points). Give proofs for $\langle C \rangle \to c > 0$ and $C \to 0$ (2×0.25 points).

1.5 Give a pseudocode of the algorithm (0.5 points) and a printout of the implementation (0.25 points). State d_{90} of 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 random selection of nodes, which can be easily achieved with 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. More precisely, node i should be selected with probability $\frac{k_i}{2m}$, where k_i is its degree and m is the number of edges. The algorithm should run in constant time $\mathcal{O}(1)$, whereas you can assume any standard network representation. *(Think more about 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

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

where v_i is some non-negative 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 for p_{ij} in terms of *only* degree sequence $\{k\}$ and recognize the result.

2.3 Node degree distribution (1.5 points)

Represent a small part of 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, $k > 0$. *(Transformation to logarithmic axes should be done by your plotting software.)*

Next, let n and m be the number of nodes and edges, and $\langle k \rangle$ the average node degree of Facebook network. Construct an Erdős-Rényi random graph [ER59] with parameters n and m, and again compute its p_k. Superimpose p_k on the same plot using dots of different color as before. Compute also theoretical degree distribution of 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 degree 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 superimpose 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 network representation and give a pseudocode of the algorithm (0.5 points). Reason or prove that the algorithm returns the correct result (0.25 points).

2.2 Show that $\langle k_i\rangle$ is proportional to v_i (0.25 points). Derive an expression for p_{ij} in terms of $\{k\}$ and comment on the result (0.5 points).

2.3 Draw a single plot with four distributions and briefly discuss each result (4×0.25 points). Give a printout of the implementation of the preferential attachment model (0.25 points) and a printout of the code used to compute p_k (0.25 points).

3 Node position (1.5 points)

You are given Slovenian highways 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 best utilized to predict the traffic load at each location. You should consider at least three node measures: node degree k_i, clustering coefficient $C_i = \frac{2t_i}{k_i(k_i-1)}$ [WS98] and harmonic mean distance $\ell_i^{-1} = \frac{1}{n-1}\sum_j \frac{1}{d_{ij}}$ [New10].

Possibly the simplest way to achieve this is to compute correlations between the values returned by different node measures and traffic loads. Compute either Pearson or Spearman correlation coefficient for each of the three measures. Are the results expected or surprising? Also, list top ten locations according to your best measure along with its values and actual traffic loads and breifly discuss the results.

What to submit?

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

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.

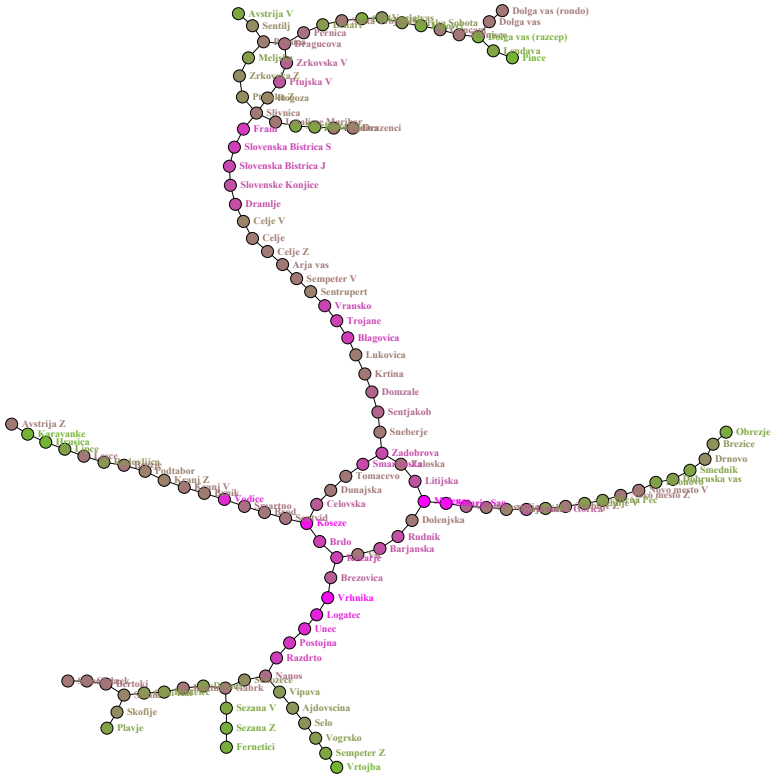

Figure 3: **Slovenian highways network with node colors corresponding to traffic loads**

[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 E. J. 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.