Homework #2

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 **April 13th** at 2:00pm, while late days expire on **April 17th** at 1:00pm. 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: ____

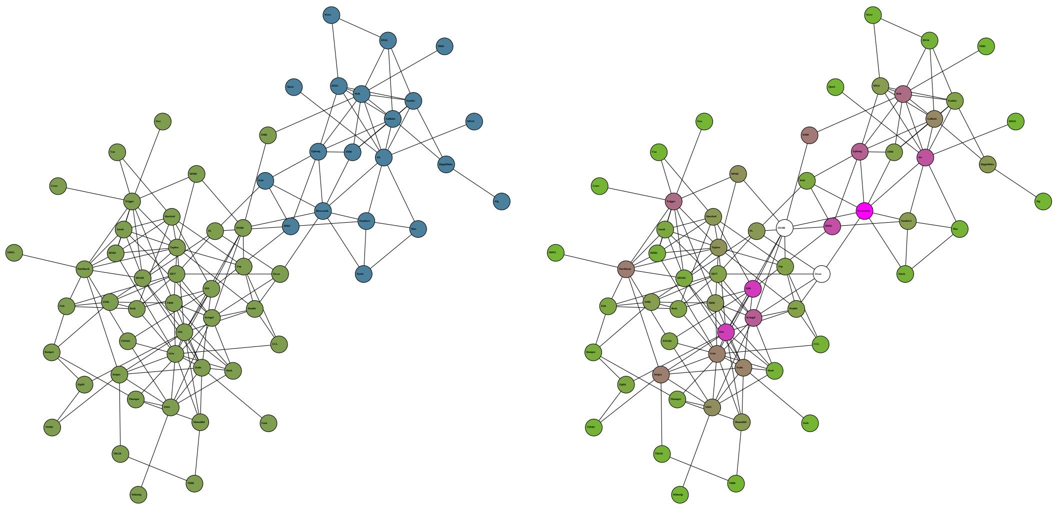
Date & time: ______________________

I acknowledge and accept the honor code.

Signature: ______________________

1 Where's SN100? (1 point)

Figure 1 shows social network of bottlenose dolphins famously studied by Lusseau [LSB+03]. After the disappearance of a particular dolphin named SN100 during the course of experiment [LN04], the rest split into two groups shown by different colors in Figure 1a. Your task is to analyze whether different measures of node position could be utilized to detect this obviously important role of SN100. You should consider at least three measures of node importance: one spectral analysis approach such as degree centrality, one link analysis algorithm such as PageRank algorithm [BP98] and one measure based on geodesic paths between the nodes such as betweenness [Fre77] or bridgeness [JMK+15] centrality. State the rank of SN100 for each measure of node importance and briefly discuss the results.



(a) Division into **two groups** (b) **Betweenness centrality**

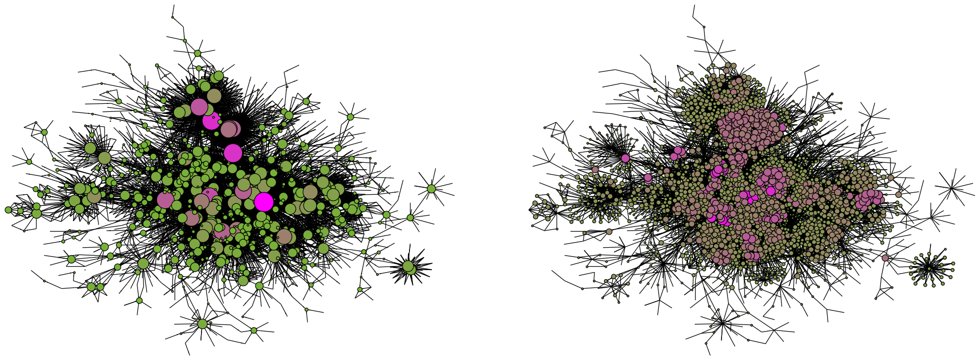
Figure 1: **Lusseau bottlenose dolphins network**

What to submit?

State the rank of SN100 for each measure and briefly discuss the results (3×0.25 points). Give a printout of all the code used to solve the exercise (0.25 points).

2 Is software scale-free? (2 points)

Consider two software class dependency networks [ŠB11] representing java namespace of Java language and Lucene search engine library (see Figure 2). These are directed networks where node i links to node j if software class represented by i depends on or *uses* class represented by j. Compute degree distributions p_k of both networks and plot them separately on log-log plots by representing each distinct (k, p_k) with a single dot. *(Transformation to doubly logarithmic axes should be done by your plotting software.)* Next, compute also in-degree distributions $p_{k^{in}}$ and out-degree distributions $p_{k^{out}}$ of both networks, and superimpose them on the same plots as before using dots of different color. Compare all three degree distributions $p_{k^\cdot}$ of each network and highlight the differences among them. Do the distributions appear to be scale-free



(a) Colors show **node in-degree** (b) Colors show **node out-degree**

Figure 2: **Lucene class dependency network**

following a power-law $k_\cdot^{-\gamma_\cdot}$? For the distributions not appearing like power-laws, reason why. For the distributions that seem to follow a power-law, reason why and compute the power-law exponents $\gamma_\cdot$ using the maximum likelihood formula in Eq. (1) [Bar16], where $k_i^\cdot$ is $\cdot$-degree of 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.

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

What to submit?

Draw two plots with three distributions each and briefly comment on the results (3×0.25 points). Briefly reason whether distributions are scale-free and why, and state the power-law exponents of seemingly scale-free distributions (3 × 0.25 points). Give a printout of all the code used to solve the exercise (0.5 points).

3 Errors and attacks on the Internet (2 points)

Consider the Internet overlay map represented by an undirected graph. Routers, switches and hubs on the Internet fail from time to time. Your task is to study how this affects the Internet's ability to stay connected. Simulate such failure by random removal of nodes and measure the size of the largest connected component in the resulting network. Next, consider also a malicious individual attacking the Internet. Simulate the attack by removing most highly linked nodes or hubs and again measure the size of the largest connected component. For each of the two scenarios, plot the fraction of nodes in the largest connected component after removing 0%, 10%, 20%, 30%, 40% and 50% of randomly selected nodes or hubs. Finally, repeat both experiments also on a random graph [ER59] with the same number of nodes and links as the Internet network. Briefly discuss the results by comparing the robustness of the Internet and a random graph. Could you claim that the Internet is robust to random failures? Could you claim that the Internet is robust to malicious attacks? What about a random graph? Could you explain your reasoning by some structural property of the Internet and/or a random graph?

What to submit?

Draw two plots showing the robustness of the Internet network and the robustness of a random graph (3×0.25 points). Briefly comment on the results and give answers to all four questions (4×0.25 points). Give a printout of all the code used to solve the exercise (0.25 points).

4 HIV and network sampling (2 points)

When a patient is diagnosed with COVID-19 or say HIV, in most Western countries, she or her will be questioned about recent sexual contacts. The authorities would then make an effort to track down those contacts and test them for HIV. The process is repeated with anyone who also tests positive, tracing her or his contacts as well, until all leads have been exhausted. This process is called contact tracing. Notice that contact tracing also gives a (biased) sample of the underlying social network. Assuming that one gets HIV from a random sexual contact, contact tracing can be simulated by a simple random walk [LF06]. Simulate a random walk on a small social network until you sample 10% of the nodes and take an induced graph on the sampled nodes for your sampled network. Is the original social network small-world [WS98] and/or seemingly scale-free [BA99]? Is the sampled network small-world and/or seemingly scale-free? Could you reason why? Support your answers with appropriate computations.

What to submit?

Give brief answers to all three questions and provide necessary results to support your reasoning (3×0.5 points). Give a printout of all the code used to solve the exercise (0.5 points).

5 Who to vaccinate? (0.5 points)

Probability that an individual would spread the disease through her or 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 randomly select the same number of individuals, but then rather vaccinate a random acquaintance of theirs. Which of the two schemes do you expect to provide better immunization? Why?

What to submit?

Give brief answers to both questions (2×0.25 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, 2016.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN*, 30(1-7):107–117, 1998.

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

[Fre77] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[JMK+15] Pablo Jensen, Matteo Morini, Marton Karsai, Tommaso Venturini, Alessandro Vespignani, Mathieu Jacomy, Jean-Philippe Cointet, Pierre Merckle, and Eric Fleury. Detecting global bridges in networks. *e-print arXiv:150908295v2*, 2015.

[LF06] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2006.

[LN04] David Lusseau and M. E. J. Newman. Identifying the role that animals play in their social networks. *Proc. Biol. Sci.*, 271:S477–S481, 2004.

[LSB+03] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? *Behav. Ecol. Sociobiol.*, 54(4):396–405, 2003.

[New10] Mark 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.

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