Network Science Exam I

November 8th, 2024, Fall Semester 2024/2025 2 hours

The exam can be answered in Portuguese or in English in a separated exam sheet. All 8 questions are equally valued.

I. Consider the harmonic centrality and let $G = (V, E)$ be an undirected sparse graph where V is the set of vertices and $E \subseteq V \times V$ is the set of edges; let $n = |V|$ and $m = |E|$ denote the number of vertices and the number of edges, respectively. Assume that G is provided as input through the list of m edges. Propose a representation for G and an algorithm to determine the harmonic centrality for all vertices $u \in V$, and discuss its computational complexity with respect to both time and space.

Solution hints: Let us load the graph in an adjacency list representation, which allows us to iterate over the neighbors of a given $u \in V$ in time proportional to the size of its neighborhood; an moreover such representation requires $O(n+m)$ space and can be constructed also in $O(n+m)$ time, where $n = |V|$ and $m = |E|$. Then, since we need to compute the sum of the inverse shortest path distances of a node to all other nodes, we can use a conventional shortest path algorithm to compute the shortest path distances from a node to all other nodes; with G unweighted, we can use a BFS from each node $u \in V$, which allows us to compute all shortest path distances from u to all other nodes in $O(n + m)$ time.

II. Define PageRank and Katz centrality, and discuss their properties and differences.

Solution hints: Let $\mathbf{A}$ be the adjacency matrix of a given graph, and $\mathbf{D}$ its degree diagonal matrix. The Katz centrality can be defined as the solution $\mathbf{x}$ of $\mathbf{x} = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{1}$, and the PageRank as the solution $\mathbf{x}$ of $\mathbf{x} = \alpha\mathbf{A}\mathbf{D}^{-1} + \beta\mathbf{1}$, where α and β are positive constants. The Katz centrality has one undesired feature, if a vertex has high Katz centrality then its neighbors also get high Katz centrality, which is often not appropriate, namely when such vertex has millions of neighbors. PageRank addresses this issue by assigning a centrality to a given vertex v that results from summing up the centrality of each neighbor u point to v divided by the out-degree of u.

III. Given a degree distribution for a set of vertices V, propose an algorithm to build a random graph $G = (V, E)$, with edges $E \subseteq V \times V$ randomly chosen, but satisfying the prescribed degree distribution. Explain how and why your algorithm works, and discuss its complexity.

Solution hints: We can construct vertex stubs, i.e., all possible edge linking points for all vertices according to the degree distribution, and then select randomly pairs of stubs to connect. Such an algorithm runs in $O(n + m)$ time, where n is the numbers of vertices and m is the number of edges.

IV. How robust are scale-free networks against targeted attacks? Justify your answer, relating in particular to how such attacks may affect the maximum degree of such networks and the heterogeneity of their degree distribution.

Solution hints: Scale-free networks are not robust against target attacks, being disrupted pretty quickly, and with the maximum degree as well as the degree distribution heterogeneity falling also quickly.

V. Briefly explain i) what is the basic reproduction number R_0, ii) how can we use it to estimate the critical proportion of a population to be vaccinated in order to avoid a disease outbreak described as a SIR model in a well-mixed population, and c) how would your reasoning change if the efficiency of your vaccine is not 100%?

Solution hints: a) The basic reproductive number of an infection is the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. The definition assumes that no other individuals are infected or immunized in the population. b) R_0 values are estimated having a compartment model in mind, and the estimated values are dependent on the model used and values of other parameters. For the SIR model, $R_0 = \frac{\beta\langle k\rangle}{\delta}$ such that if $R_0 < 1$ the outbreak will die out, and if $R_0 > 1$ the outbreak will expand. If a fraction w of the average number of contacts $\langle k\rangle$ is vaccinated, we will get a new reproductive ratio given by $R_{vac} = \frac{\beta\langle k\rangle(1-w)}{\delta} = R_0(1-w)$, such that the disease will die out for $R_{vac} < 1 \rightarrow w > w_c = 1 - 1/R_0$. As an example, an $R_0 = 3.0$ would imply a critical proportion of vaccinated individuals of 66% to halt the disease. c) If only a fraction φ of the vaccines are effective, then $w_c = \left(1 - \frac{1}{R_0}\right)/\varphi$.

VI. Consider the process of vaccinating a population without complete/global knowledge of the contact network topology. Assuming nevertheless that the contact network is scale-free, suggest a vaccination strategy which is likely to perform better than random vaccination. Justify your answer.

Solution hints: We can vaccinate the acquaintances of randomly selected individuals, indirectly targeting the hubs without having to know precisely which individuals are hubs. Example: 1) Choose randomly a fraction p of nodes (Group 0). Select randomly a link for each node in Group 0. Let us call Group 1 to this new set of nodes. Immunize the Group 1 individuals. Proceed like this, creating a group 2 if needed, from the neighbors of the vertices in group 1

VII. Propose a benchmark capable of assessing the accuracy of a community finding algorithm. You may briefly explain the idea of one of the benchmarks we discussed in our classes or propose a new one.

Solution hints: Imagine a simple variant of the Girvan-Newman benchmark. Start with N nodes partitioned into nc communities of equal size. Each node is connected with probability p_{int} to the nodes in its community and with probability p_{ext} to the nodes in the other communities. These probabilities may be adapted to have a fixed average degree or not. The control parameter $\mu = k_{ext}/(k_{ext} + k_{int})$ captures the density differences within and between communities. We expect community finding algorithms to perform well for small μ. The performance of all algorithms should drop for large μ, when the link density within the communities becomes comparable to the link density in the rest of the network. Other possible answer would be to consider the Lancichinetti–Fortunato–Radicchi benchmark. This algorithm generates networks that have a priori known communities. Once again, these networks are used to compare different community detection methods. In this case, the algorithm is able to account for the heterogeneity in the distributions of node degrees and of community sizes. Finally, one can also consider a real-world dataset with known communities, such as the Zachary's karate club.

VIII. Please indicate whether each of the following statements is TRUE or FALSE. (Note: For each wrong answer we discount a correct one.)

a) The Prisoner's dilemma involving two players leads to cooperation dominance.

b) An evolutionary stable strategy, even if adopted by all individuals in a population, can be replaced or invaded by a different strategy through natural selection.

c) Disease spreading in networked populations can be modelled as simple contagion.

d) Betweenness centrality ranks the importance of nodes based on the importance of their peers.

e) In network science we call a community a group of nodes that have a higher likelihood of connecting to nodes within the community than with nodes in other communities.

f) In the model proposed by Watts and Strogatz in 1998 the average path length falls very rapidly with increasing p, where p is the edge rewiring probability.

Solution: False, False, True, False, True, True