AMSC808N/CMSC828V Maria Cameron

**Final exam. Due Monday December 20, 2021, 11:59 PM.**

**Resources:** *You can use any internet resources, textbooks and course materials. You are not allowed to collaborate with your classmates or use anybody's help.*

**Programming:** *You can use any suitable language. A high-level language (e.g. Matlab or Python) is preferable. You are allowed to use built-in functions or available libraries for graph algorithms.*

**Submission:** *You should upload on ELMS a single pdf file with your codes linked to it. For example, you can publish Matlab's code or make a Jupiter notebook and link them to your pdf. Use latex or any other suitable text editor. I will subtract 10% of the maximal score if the file is hand-written.*

This final exam is project-style. You will need to investigate a real-world network: the LastFM Asia Social Network available at http://snap.stanford.edu/data/index.html.

1. Download the data for the LastFM Asia Social Network from http://snap.stanford.edu/data/feather-lastfm-social.html. All you need is the file with edges. Note that node indices start with 0.

2. What is the number of connected components of this network, and what are their sizes?

3. Find the degree distribution for this network. Plot $p_k$ versus $k$ in log-log scale. You will see something like Fig. 4.22 (top right) in Barabasi, "Network Science", chapter 4. Then use log-binning (like Fig. 4.22 (bottom left)) and plot the result in the same figure. For doing log-binning, use bins $b_0$ containing only nodes of degree 1, $b_1$ containing nodes of degrees 2 and 3, ..., $b_n$ containing nodes of degrees $2^n \le k < 2^{n+1}$, etc. The last bin $n_{\max}$ should be such that $n_{\max}$ is the largest integer such that $2^{n_{\max}+1}$ does not exceed the largest degree in the network. The $n$th point in log-binning has coordinates (the mean degree in the bin $b_n$, the mean degree probability in bin $b_n$):
$$\left(\langle k\rangle_n = \frac{\sum_{k=2^n}^{2^{n+1}-1} k p_k}{\sum_{k=2^n}^{2^{n+1}-1} p_k}, \langle p_k\rangle_n = \frac{\sum_{k=2^n}^{2^{n+1}-1} p_k}{2^n}\right). \tag{1}$$

4. Approximate the log-binning data for the degree distribution with the power-law with exponential cut-off:
$$p_k = Ce^{-\alpha k}k^{-\tau}, \tag{2}$$
where $C$, $\alpha$, and $\tau$ are positive constants that you need to find by solving a linear least squares problem. To set it up, take logs of both sides of (2). Plot the approximation to the degree distribution that you find in the same figure as in the previous item. Include legend.

5. Find the average shortest-path length in the actual network. The use of a built-in or a library function for this task is preferable. Now imagine a random graph that has degree distribution (2) with parameters that you have found. For brevity, we will refer to it as *the random graph with the same degree distribution*. Estimate the average shortest-path length in this random graph. *Hint: the paper by Newman, Strogatz and Watts should be very helpful.*

6. Find the clustering coefficient $C$ for the actual network defined as
$$C = \frac{\#(\text{closed paths of lengths 2})}{\#(\text{paths of length 2})}. \tag{3}$$

Now find the clustering coefficient for the random graph with the same degree distribution. Proceed as follows. Let $v$ be an arbitrary node of degree 2 or more. Randomly pick two of its first neighbors $i$ and $j$. Let their *excess degrees* be $k_i$ and $k_j$. The probability that there is a link between $i$ and $j$ is $\frac{k_i k_j}{2m}$ where $2m = n\langle k \rangle$ is twice the expected number of edges, and $n$ is the number of nodes. Then the clustering coefficient is equal to the expectation for $\frac{k_i k_j}{2m}$ taken with respect to the joint excess degree distribution for $i$ and $j$. Due to the independence of excess degree distributions for $i$ and $j$ in the random graph, the joint probability mass function is $q_{k_i} q_{k_j}$. Calculate this expectation and show that it is equal to

$$C_{random} = \frac{1}{n} \frac{\left[\langle k^2 \rangle - \langle k \rangle\right]^2}{\langle k \rangle^3}. \tag{4}$$

7. Comment on the relationships between the shortest-path length and clustering coefficient for the actual network and the random graph with the same degree distribution. Try to explain the discrepancies between them.

8. Consider the spread of an epidemic disease on the actual network. What is the critical transmissibility $T_c$ for this network such that for transmissibility $T > c$ an epidemic occurs and for $T < T_c$ only small outbreaks are possible?

9. Set $T = 0.4$. Run the SIR model on the actual network. Proceed as follows. Randomly select the $T$ fraction of edges. Only these edges will be present in the *transmission graph*. Start with a single infected node. Duration of sickness is one time step. Each sick node infects all its first neighbors in the transmission graph. Repeat 10 times. Plot the number of sick nodes versus time step for each run in the same figure. Whenever you see an epidemic, what fraction of nodes is affected by it?

10. Now consider the random graph with the same degree distribution and a very large number of nodes. Make the following predictions for it using the theory developed in Newman, Strogatz and Watts, 2001, Cohen et al. 2000, and Newman, 2002:

    (a) What is the critical transmissibility?

    (b) What is the expected fraction of nodes affected by an epidemic? *Hint: in order to do it, you will need to solve a nonlinear equation involving polylogarithm. Take real part of polylog to avoid trouble with function evaluation.*

    (c) What is the critical fraction of nodes that need to be vaccinated (i.e. removed from the transmission graph) in order to avoid the epidemic?

    Comment on the relationship between the critical transmissibility and the fraction affected by an epidemic for the actual and random graphs.