ELEC 573: Network Science and Analytics **Fall 2021**

Homework 3

Hand-out Date: 10/25/21 **Due Date: 11/09/21**

This homework focuses on network models. There are four problems combining theoretical and coding tasks. Please hand in all of your work, including executable code.
The grading is over 100 points distributed as indicated throughout the document. There is an exercise that carries 5 extra-credit points.

3.1) Erdős Rényi phase transitions [24 points + 5 extra credit]. We begin by exploring numerically some of the properties of Erdős Rényi (ER) graphs. For this we will create ER graphs with $n = 5,000$ nodes, and increasing probability p.
a) *[18 points]* Create a number of ER graphs with increasing probability p. Your probabilities should cover the range $p = [10^{-5}, 10^{-2}]$ with 'logspace' (e.g., `np.logspace` in python). Plot the size of the largest connected component relative to the number of nodes n (i.e., if the giant component consists of the whole graph its size is 1), as a function of p. For each p you should create 20 graphs and plot the mean value plus / minus the standard deviation of the giant component size. (hint: consider using the functions `fast_gnp_random_graph` and `connected_components` in networkx).
b) *[6 points]* What do you observe for $p \approx 1/n$? What happens for $p \approx \ln(n)/n$? Provide a brief description of these phenomena in terms of the threshold properties that we studied in class.
c) *[5 extra credit]* Let A denote the event that node 1 has at least $l > 1$ neighbors in an ER graph $G_{n,p}$. Consider a family of functions $t(n) = r/n$ for any $r \in \mathbb{R}^+$. Show that $P(A) \to 0$ if $\frac{p(n)}{t(n)} \to 0$ and $P(A) \to 1$ if $\frac{p(n)}{t(n)} \to \infty$ for increasing n and any $r \in \mathbb{R}^+$.

3.2) Fitting power-law exponents [29 points]. [adapted from Kolaczyk 4.1, credit Gonzalo Mateos] Consider the problem of estimating the exponent $\alpha > 0$ of a power-law distribution with probability density function (PDF) proportional to $x^{-\alpha}$.
a) *[5 points]* A random variable X is said to follow a Pareto distribution if it has a PDF

$$f(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}, \quad \text{for } x \geq x_0 > 0.$$

Verify that $f(x)$ is a valid PDF, and determine the complementary cumulative distribution (CCDF) function $\bar{F}(x) = P[X \geq x]$.
b) *[7 points]* Generate a synthetic dataset of $n = 100,000$ independent and identically distributed samples from a Pareto distribution, for $\alpha = 1$ and $x_0 = 1$. Samples may be drawn from a Pareto distribution by using the inverse transform method. Specifically, having generated random numbers $u_1, u_2, \ldots, u_n$ in $[0, 1]$, the values $x_i = \frac{x_0}{u_i^{1/\alpha}}$, $i = 1, \ldots, n$ will follow the desired Pareto distribution (Can you argue why is this true?). Plot on a log-log scale the empirical estimate of $P[X = x]$ from your data. To obtain this estimate, you may round each point to the nearest integer and compute a normalized histogram over integer values. Use 'points' instead of 'bars' for your histogram plot in log-log scale; and superimpose a plot of the actual PDF (i.e., a line of slope $-\alpha - 1$) to verify that you generated the data correctly.
c) *[6 points]* Show that under the Pareto model the log-likelihood function is given by

$$\ell_n(\alpha) = \sum_{i=1}^{n} [\log \alpha + \alpha \log x_0 - (\alpha + 1) \log x_i]$$

and it is maximized by the Hill estimator that we discussed in class.
d) *[11 points]* For $n = 100,000$, $\alpha = 1$, and $x_0 = 1$, compute estimates of the exponent α via (i) least-squares regression on the log-log empirical distribution (histogram) $\log P[X = x]$; (ii) least-squares regression on the log-log empirical CCDF $\log \bar{F}(x)$; and (iii) maximum likelihood as derived in c). Repeat the data-generation

process and estimation 100 times, and report the sample mean and standard deviations for each method (i)-(iii). Based on these values, which method gives the best estimate?

3.3) Small-world clustering coefficient [27 points]. Consider the following variation on the small-world model studied in class. Just like in the original model, we start with a circular lattice with n nodes in which each node is connected to the $2k$ nodes that are k positions or less away in the lattice. For each edge in this lattice, with probability p we add an edge between two nodes nodes chosen uniformly at random *but we do not delete the original edge*. That is, instead of rewiring edges, we keep the original lattice graph and add a few shortcut edges on top of it.
a) *[13 points]* Show that when $p = 0$, the overall clustering coefficient of this graph is given by

$$C = \frac{3k-3}{4k-2}.$$

b) *[14 points]* Show that for small $p > 0$ (so that terms with p^2 can be ignored), the expected clustering coefficient when $n \to \infty$ is given by

$$C = \frac{3k-3}{4k-2+8kp}.$$

3.4) Uniform attachment model [20 points]. [adapted from MIT 1.022] In this exercise, we will implement the mean-field approximation idea that we saw in class to a non-preferential attachment setting. Consider a system where nodes are born over time and form edges to existing nodes at the time of their birth. The network formation process is as follows:

- Nodes are born over time and indexed by their time of birth, i.e., node i is born at time instant I for $i = 1, 2, \ldots$.
- The network is initialized with m nodes (born at times $1, \ldots, m$), all connected to one another. The first newborn node is thus the one born at time $m + 1$.
- Each newborn node selects m nodes uniformly at random among the existing set of nodes and links to them.

Let $k_i(t)$ be the degree of node i at time t. Let's use a continuous-time mean-field analysis (just like we did in class on Oct 25, 2021) to track the evolution of the expected degrees of nodes.
a) *[2 point]* What is $k_i(i)$ for $i > m$?
b) *[9 point]* Find an expression for $dk_i(t)/dt$ in the mean-field approximation and verify that $k_i(t) = m + m\log(t/i)$ for $t \geq i$ is a solution to that differential equation.
c) *[9 point]* Let $p(d)$ for $d \geq m$ be the probability of a node having degree d in the uniform attachment mode. Derive an exptession for $p(d)$ from the CCDF as done for the preferential attachment model in class.