| **Econ 1033** | Problem set 2 |
|---|---|
| Prof. Ben Golub | Due: **Wed., Sep 16, 2020 11:50pm** |

This problem set is worth 100 points. You should attempt problems totaling at least 100 points, and submit solutions to all parts of those problems. For problems where you use code, attach your code to your solution.

**1. (20 points)** Download `pset2.zip`. It contains a CSV and a MATLAB version of the 20-by-20 adjacency matrix of the social network of our class that we constructed in lecture. Call this matrix $\boldsymbol{G}$. Think of the nodes as the set $N = \{1, 2, \ldots, 20\}$, with $G_{ij}$ telling us whether $i$ and $j$ are adjacent. (The order of rows and columns has been scrambled.)

*Note:* You should download and install MATLAB[1] or a comparable tool (e.g. Python with numpy) for this problem and learn to use it to add, subtract, and multiply matrices. Please schedule an office hour with Yixi if you need help doing this (do this ASAP!). After this problem set it will be assumed that you have a tool to do basic computations with matrices of sizes such as 20-by-20 or a bit bigger.

**a. (4 points)** Compute the total number of walks of length 2 in $\boldsymbol{G}$.

**b. (8 points)** Note that $\boldsymbol{G}$ has all zeroes along the diagonal. Define $\boldsymbol{H} = \boldsymbol{G} + \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix of the appropriate size. Assuming $i \neq j$, prove that there is a walk of length $\ell$ or less from $i$ to $j$ in $\boldsymbol{G}$ if and only if $(\boldsymbol{H}^\ell)_{ij} > 0$.

**c. (4 points)** Find the minimum positive integer $\ell$ so that if $(\boldsymbol{H}^\ell)_{ij} > 0$ then $(\boldsymbol{H}^k)_{ij} > 0$ for all $k > \ell$. Call this $\ell(\boldsymbol{G})$. (In words, this is the power where entries stop going from zero to positive; as long as they've become positive already, they stay positive.)

*Hint*: in MATLAB, running the command `B=(A>0)` returns a matrix $\boldsymbol{B}$ which tells you whether entries of $\boldsymbol{A}$ are positive or not. More precisely, its output is

$$B_{ij} = \begin{cases} 1 & \text{if } A_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

**d. (4 points)** Explain how you can figure out the connected components of $\boldsymbol{G}$ just by looking at $\boldsymbol{H}^{\ell(\boldsymbol{G})}$, without doing any further computations. Write down the connected components (as subsets of the node set $N$).

**2. (20 points)** Read the basic set-up of Problem 1 above. For this problem you need a little more coding experience than Problem 1, so skip it if you're new to coding but read the solution when it comes out.

**a. (4 points)** Compute the average degree of $\boldsymbol{G}$.

**b. (4 points)** Write code in your language of choice for the following function. The input is an adjacency matrix $\boldsymbol{A}$ (of an undirected graph) and two indices $i$ and $j$. The output is an integer. If the distance between the two nodes in the graph is finite (i.e., they are in the same connected component), then that is the output. Otherwise the output is $-1$, or `Inf` if your language has

[1] `downloads.fas.harvard.edu/download`

it.[2] Use this to compute the number of nodes in the largest component of $\boldsymbol{G}$ and the diameter of this component.

**c. (4 points)** Write code to generate a random undirected, unweighted graph by linking every pair of distinct nodes with some probability $p$ (i.e., no self-edges). Choose $p$ so that the expected degree of every node in your graph is equal to your answer in (a). Tell us what $p$ makes this true.

**d. (4 points)** For each random graph you generate in (c), compute the size of the largest connected component. Report the average and standard deviation of this number in a sample of 1000 random graphs.

**e. (4 points)** For each random graph you generate in (c), compute the diameter of the largest connected component (if two components are the same size, pick one of them however you like). Report the average and standard deviation of this number in a sample of 1000 random graphs.

**3. (20 points)** Consider an "infection" process that evolves according to a more general rule than the basic $(k, p)$ model of Easley and Kleinberg: Start with a single infected individual in wave 0. In every wave, each infected individual produces a random number of children (nodes directly infected by it) in the next wave, independently of other individuals. The distribution of number of children, $P$, is a probability mass function with finite mean (a.k.a. expectation) denoted by $R_0$, and $P$ is the same across individuals. Denote by $X_n$ the number of infected individuals at wave $n$.

**a. (5 points)** Describe the relationship between $X_{n+1}$ and $X_n$ by expressing $X_{n+1}$ as a sum of $X_n$-many random variables. Your description can be in words as long as it is clear and correct.

**b. (5 points)** Given $X_n$, compute the conditional expectation $\mathbb{E}[X_{n+1} \mid X_n]$.

**c. (5 points)** Explain why the ratio $\mathbb{E}[X_{n+1}]/\,\mathbb{E}[X_n]$ is a deterministic constant, and say what this ratio is. Write $\mathbb{E}[X_n]$ in a simple way using this insight.

**d. (5 points)** What is $C_n$, the *cumulative* number of infections up to and including wave $n$? What is the expectation of this random variable? Write your answer as simply as possible, and justify it.

**4. (20 points)**

**a. (10 points)** In our reading and class discussion, we have been taking a branching process perspective on the growth of a virus, by studying the growth of the (random) "family tree" branching out from a "patient zero." In [J-HN] Chapter 3, contagions are studied by thinking about components of a large (random) graph. Can you harmonize these two perspectives? How are they related to each other? Write about 200-300 words.

**b. (10 points)** Take the $(k, p)$ model from [EK] 21.2/21.8.A, with one tweak: because the individuals involved are reacting to the epidemic in real time, $p$ actually depends on the cumulative number of cases to date. Recalling the notation of Problem 2(d), assume that the probability that individuals in wave $n+1$ get infected, $p_{n+1}$, is equal to $\mathcal{P}(C_n)$, where $\mathcal{P}$ is a decreasing function of its argument. You can think of this as a simple or "reduced-form" way of capturing rational

[2] Feel free to use a pre-made tool, such as the distances function in MATLAB.

agents' response to the epidemic: they protect themselves from exposure to infection when the infection is more widespread. Give reasonable conditions on $\mathcal{P}$ ensuring that (with probability 1) the epidemic is eventually stopped. Explain what your condition means practically.

**5. (20 points)** Recall the following function from [EK 21.8.A]:

$$f(x) = 1 - (1 - px)^k, \quad x \in [0, 1].$$

Assume $0 < p < 1$ and $k \geq 1$. If you need to make further (reasonable) assumptions to establish the statements below, make the assumptions clear.

**a. (2 points)** Show that $f(0) = 0$ and $f(1) < 1$, and compute $f'(0)$ in terms of $k$ and $p$.

**b. (2 points)** Show that $f$ is increasing.

**c. (4 points)** Show that $f$ is strictly concave.

**d. (4 points)** When is there is a *positive* $q^*$ such that $q^* = f(q^*)$? Explain your answer. Again, the condition you give will involve $k$ and $p$.

**e. (4 points)** What happens to $q^*$ if you hold $p$ fixed and increase $k$ to some other $k'$? Justify your answer. Also give a clear intuitive explanation of why this happens.

**f. (4 points)** What happens to $q^*$ if you hold $k$ fixed and increase $p$ to some other $p'$? Justify your answer. Also give a clear intuitive explanation of why this happens.

**6. (20 points)** This exercise is for those who know some basic programming (in Python, MATLAB, or any other reasonable tool for quantitative simulation) and enjoy understanding things via simulations. Consider the branching process studied in Easley and Kleinberg 21.2/21.8.A. Recall from Problem 3 that $X_n$ is the number of newly infected individuals at wave $n$.

**a. (5 points)** We know that $\mathbb{E}[X_n] = (R_0)^n$. This gives us a prediction of the mean number of infected, but nothing about its variability. Choose five combinations $(k, p)$ with $kp = 1.5$ and compute $X_{20}$. Do at least 10,000 simulations for each combination and report the empirical standard deviation of $X_{20}$ in your simulations.

**b. (5 points)** Imagine that you see one realization of a branching process, and observe that in this realization $X_{20} = 30000$. Discuss how you could use your work in (a) to statistically test (and possibly reject) the null hypothesis that $k = 3$ and $p = 0.5$.

**c. (5 points)** Do the simulations of (a) again but keep only the simulations where $X_{10} \geq 30$. For those simulations, what is the fraction of cases in which $X_{20} \geq 3000$? How does this compare to the unconditional frequency of the same event (i.e. the frequency of $X_{20} \geq 3000$ without throwing away any simulations)?

**d. (5 points)** Explain what is going on in (c).

**7. (20 points)** Write a 400-word essay about something that the theory of branching processes and random graphs (from our reading and lectures so far) has helped you understand better, which was *not* discussed in the reading or class discussions.