Please keep in mind the standard collaboration policy. You may use standard mathematical software (Perl, Maple, Matlab, etc.) and you may write programs to help you determine answers. For problems where numerical answers are required, please CLEARLY give your numerical answers at the beginning of your solution. This will make it easier to grade your paper, ensuring a higher score. (In general, please also keep in mind that the longer it takes me to read your solutions, the lower your score is likely to be; so show your relevant work, but feel free to organize it so I don’t actually have to read it if you are correct.)

Most of these problems can be solved in a variety of ways, including numerically by simulation. Any reasonable approach is fine, although you should always explain your attack.

Please turn this in using Canvas. Each problem is worth the same (10 points), regardless of actual difficulty, because it makes grading easier. If you need an extension, please ask; I’m more pleasant if you ask in advance, however, rather than the last minute.

1. The generalized Fibonacci numbers are defined as follows:

\[ F_d(j) = 0 \text{ for } j < 0; \quad F_d(0) = 1; \quad F_d(j) = \sum_{k=j-d}^{j-1} F_d(k) \text{ for } j > 0. \]

In English, this means for the Fibonacci sequence \( F_d \), you get the next number by adding up the \( d \) previous numbers. \( F_2 \) is just the standard Fibonacci sequence 1, 1, 2, 3, 5, 8, 13, …. The sequence \( F_3 \) is 1, 1, 2, 4, 7, 13, 24, …. For \( d = 3 \), \( d = 4 \), and \( d = 5 \), find the matrix form of the Fibonacci equation and determine its largest (real) eigenvalue. From this determine the asymptotic rate of growth for the generalized Fibonacci numbers for \( d = 3 \) and \( d = 4 \).

2. Consider the transition matrix for the following Markov chain on states a, b, c, and d:

\[
\begin{array}{cccc}
   & a & b & c & d \\
 a & 0.3 & 0.0 & 0.5 & 0.2 \\
b & 0.0 & 0.4 & 0.3 & 0.3 \\
c & 0.3 & 0.2 & 0.0 & 0.5 \\
d & 0.4 & 0.1 & 0.5 & 0.0 \\
\end{array}
\]

Note that in the matrix the entry in the \( i \)th row and \( j \)th column represents the probability of going from the \( i \)th state to the \( j \)th state. So the probability of going from state \( a \) to state \( d \) is 0.2, and the probability of going from state \( b \) to state \( a \) is 0.0. First, find the stationary distribution of the Markov chain.

Suppose we start with the chain in state \( a \). Find the probability distribution after 10 steps. (In other words, find the probability of being at each state after 10 steps, starting from \( a \).) What about 100 steps?

The total variation distance of the chain from a stationary distribution \( \pi \) at time \( t \) when starting at state \( y \) is given by

\[ \frac{1}{2} \sum_x |P^t_{yx} - \pi(x)|. \]

Here the probability \( P^t_{yx} \) is the probability of being in state \( x \) after \( t \) steps when starting from \( y \). How many steps \( t \) do you need to take before the total variation distance when starting from \( d \) is first less than 0.01? What about the number of steps until it is first less than 0.001, and 0.0001? Does the total variation distance always decrease when you calculate it, or does it ever increase at some step?
3. Determine Pagerank scores and Hub and Authority scores for the graph given by the following transition matrix. The transition matrix tells you what edges are in the graph. For example, from the first row we can see that there is a directed edge from node a to nodes c, d, and e. In calculating the PageRank, assume that at each step the probability of jumping to a random page is 0.15. Also, be sure to check that you are normalizing scores appropriately (particularly for Hub/Authority).

\[
\begin{array}{cccccc}
& a & b & c & d & e & f \\
a & 0 & 0 & 1 & 1 & 1 & 0 \\
b & 0 & 0 & 1 & 0 & 1 & 0 \\
c & 0 & 0 & 0 & 1 & 1 & 0 \\
d & 0 & 1 & 1 & 0 & 0 & 0 \\
e & 0 & 0 & 0 & 0 & 0 & 1 \\
f & 1 & 0 & 0 & 0 & 0 & 0
\end{array}
\]

4. Suppose that I suggest the following alternative method for obtaining Hub and Authority scores, based on random walks. To obtain Authority scores, we consider the following random walk: from a page \( p_1 \), a step consists of first following back a random inlink from the current page to a page \( p_2 \), and then following forward a random outlink from this page to a page \( p_3 \). The step of the Markov chain takes us from \( p_1 \) to \( p_3 \). Similarly, Hub scores are computed as follows: from a page \( p_1 \), a step consists of first following forward a random outlink from the current page to a page \( p_2 \), and then following backward a random inlink from this page to a page \( p_3 \). The step of the Markov chain takes us from \( p_1 \) to \( p_3 \). We assume that these Markov chains are finite, irreducible, and aperiodic and hence have a unique stationary distribution.

Prove that, using this approach, the Hub score for a page is simply proportional to the number of outlinks and the Authority score is proportional to the number of inlinks.

5. Consider the following sequences:

\[
\text{q w e r t y u i o p a s d f g h j k l z x c v b n m} \\
\text{e t a o n i h s r d l u m w c f g y p b v k q x j z}
\]

(The first is keyboard order; the second is English usage frequency.) Compute the footrule distance and the Kendall-tau distance between these two sequences. (Please give the UNSCALED values – integers, not a fraction between 0 and 1. You might note the scaling formulas in the paper are actually a bit off...)

Now (different subproblem) consider the following six ranking sequences, which would be returned by say a meta-search engine.

\[
\text{A B C D, D F E C, B E F A, B C D E, C F A B, A E F D.}
\]

Determine the scores for the pages A, B, C, D, E, and F under the chains MC3 and MC4 of the rank aggregation paper (either by finding the exact stationary distribution directly or by simulation).

6. In the rank aggregation paper, it is stated that the Kendall-tau distance is the “bubble sort” distance; that is, it is the minimum number of pairwise adjacent transpositions needed to transform one list to the other. Assuming this is true, prove another fact mentioned in the paper: \( F(\sigma, \tau) \leq 2K(\sigma, \tau) \), where \( F \) is the footrule distance (unscaled) and \( K \) is the Kendall distance. One approach: use induction on \( K \), or more to the point on the number of transpositions needed. (Be careful, the proof is a little subtle.)
7. In the standard $G_{n,m}$ model of random graphs, there are $n$ vertices and $m$ edges; the $m$ edges are chosen by uniformly choosing a subset from the $\binom{n^2}{m}$ possible sets of $m$ edges. In practice, a graph from $G_{n,m}$ can be obtained by sequentially choosing edges randomly one at a time, starting from the empty graph, and throwing out repeats until $m$ edges have been chosen.

Consider the following alternative random graph model, which might better model “social network” behavior. We again choose $m$ edges sequentially starting from an empty graph on $n$ vertices, but at each step we want to choose an edge $(x,y)$ not already in the graph with probability proportional to $(\text{deg}(x) + 1) \cdot (\text{deg}(y) + 1)$, where $\text{deg}$ is the degree of the vertex. (The “plus 1” ensures that when we start, each edge can be chosen!) Hence, after the first edge $(x,y)$ is chosen, edges that neighbor $x$ and $y$ are roughly twice as likely to be chosen as other edges. Call this the $Z_{n,m}$ family of random graphs.

Explain how you would write code to construct a random graph from $Z_{n,m}$. Then, perform the following experiment. Consider graphs with 100 vertices and 250 edges. Create 100 such graphs using the $G_{n,m}$ model and the $Z_{n,m}$ model, and compute the average number of triangles in each of the graphs. Compare.

Bonus: if you can figure out how to find triangles efficiently, explain how you do so, and boost up your experiments to 1000 vertices and 5000 edges.