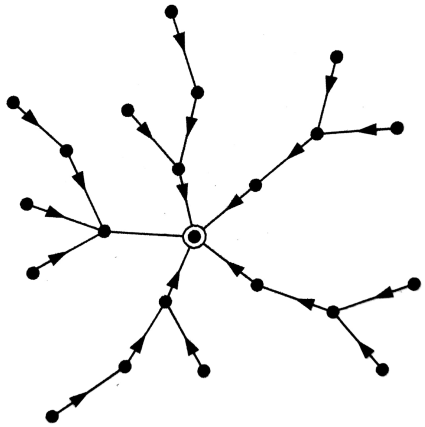
Problem 1. (PageRank (Newman 7.5))

This problem requires you to review Newman's discussion of PageRank, which we covered only briefly in class.

Suppose that a directed network takes the form of a tree with all edges pointing inward towards a central node:



What is the PageRank centrality of the central node in terms of the single parameter α appearing in the definition of PageRank and the distances d_i from each node i to the central node?

Note: The PageRank centrality satisfies the equation

$$x_i = \alpha \sum_{j=1}^{N} A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

for some arbitrary choice of $\alpha, \beta > 0$.

Problem 2. (Katz centrality and walks)

Let's consider a "new" centrality measure that we'll call *weighted walk centrality*. The idea is that a node i is important when there are lots of short walks in the graph that lead to i.

We know from lecture that the number of walks of length k leading to each node is $\mathbf{A}^k\mathbf{1}$, where $\mathbf{A}$ is the adjacency matrix and $\mathbf{1}$ is the vector of ones. If we want to prioritize shorter walks in our centrality score, we'll add a *discount factor* α so that the weighted number of walks to each node is $\mathbf{c} = \sum_{k=1}^{\infty} \alpha^k \mathbf{A}^k \mathbf{1} = \left(\sum_{k=1}^{\infty} (\alpha\mathbf{A})^k\right) \mathbf{1}$.

i. For our method to make sense, we need to guarantee that the series converges. A helpful theorem is as follows: A series $\sum_{k=1}^{\infty} \mathbf{B}^k = \mathbf{W}$ for some matrix $\mathbf{W}$ if and only if $|\lambda| < 1$ for all eigenvalues λ of $\mathbf{B}$.

   Using this theorem, determine conditions on our discount factor α to guarantee that the sum defining $\mathbf{c}$ exists.

ii. Under the conditions from part (A), prove that the weighted walk centrality measure is given by

$$\mathbf{c} = \left((\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I}\right) \mathbf{1}\,,$$

   where $\mathbf{I}$ is the identity matrix.

   Hint: there is a similarity here to the geometric series formula $\sum_{k=1}^{\infty} \alpha^k = \frac{1}{1-\alpha} - 1$. *You may like to follow this structure.*

   - Consider the expression $(\mathbf{I} - \alpha\mathbf{A})\mathbf{S}_m$, where $\mathbf{S}_m$ is the sum $\mathbf{I} + \alpha\mathbf{A} + \cdots + (\alpha\mathbf{A})^m$.
   - Because we have chosen α sufficiently small, we are guaranteed that $\lim_{m\to\infty}(\alpha\mathbf{A})^m \to \mathbf{0}$.
   - Don't forget to prove that $\mathbf{I} - \alpha\mathbf{A}$ is invertible.

iii. Explain why this centrality measure can be considered equivalent to Katz centrality as defined in the lecture notes.

Problem 3. (Katz centrality and walks (computational))

This problem is a companion to the previous problem, but you can do either one separately.

Write code to demonstrate that the weighted walk centrality for walks of length up to m, given by

$$\mathbf{c}_m = \sum_{k=1}^{m} \alpha^k \mathbf{A}^k \mathbf{1} \tag{1}$$

converges to the limiting value

$$\mathbf{c} = \left((\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I}\right)\mathbf{1} \tag{2}$$

as predicted by the analysis in the previous problem. To do this:

- Implement a function to compute the limiting value.
- Implement a function to compute the mth partial sum $\mathbf{c}_m$.
- Make a plot showing that $\mathbf{c}_m$ converges entrywise to $\mathbf{c}$.

Hint: You may want to use `np.linalg.matrix_power()` to make sure matrix multiplication is done correctly in Python.

Problem 4. (Computational Cost of Generating $G(n, p)$)

Suppose that we'd like to generate an Erdős-Rényi random graph $G(n, p)$. Assume that we incur a small constant cost each time we want to generate a random number. You can think of this cost as computation time. So, a computation that requires $10,000$ random numbers takes twice the time of a computation that requires $5,000$ random numbers. You can assume that the cost of a random number is the same whether we need to flip a weighted coin, sample from a Poisson distribution, or choose a random element from a discrete and finite set (perhaps surprisingly, this is approximately true).

i. What is the cost, in terms of quantity of random numbers needed, to generate $G(n, p)$ using the method specified in its definition? That is, we check each pair of edges and flipping a weighted coin with success probability p.

ii. Suppose we are generating a sparse Erdős-Rényi model, in which $p = c/(n-1)$. Consider now the following alternative algorithm.

- Sample a random number $M \sim \text{Binomial}(t, q)$, where t is the number of trials and q is the success probability of the binomial distribution.
- Then, choose M pairs of nodes **without replacement**, and place an edge between each of those pairs.

Prove that the graph generated by this algorithm is $G(n, c/(n-1))$. To do this, you should:

- Determine the correct value of the binomial parameters t and q.
- Prove that the presence of an edge on each pair of nodes is independent of the presence of an edge on any other pair.
- Prove that the probability of an edge being present on a pair of nodes is $p = c/(n-1)$, as needed.

iii. What is the expected cost (in random numbers) of the alternative algorithm? How does that compare to the cost you calculated in Part (a)?

iv. Suppose that $p = Cn^r$ for some constants C and r. Determine the possible values of r for which the generation scheme in Part B would be faster than the generation scheme in Part A as n grows large.

Problem 5. (Presence of Cycles in $G(n,p)$, 2 points)

Consider the Erdős-Rényi random graph $G(n,p)$ with the connection probability a function of n. In particular, we'll let

$$p(n) = \frac{f(n)}{n} \tag{3}$$

for some function f that we won't specify yet.

A **cycle** of length k (also called a k-cycle) is a walk of length k that begins and ends at the same node, without repeating any nodes or edges. Triangles are examples of cycles of length 3. Some of the approximations that we'll discuss during lecture, as well as arguments in Newman such as that connected to Fig. 11.3 in Newman, depend on the idea that *cycles are rare* in $G(n,p)$. In this problem, you'll prove some results related to this idea.

Part (a)

Fix k nodes $(i_1, i_2, \ldots, i_k)$. What is the probability that all the edges

$$(i_1, i_2)\ ,\ (i_2, i_3)\ , \ldots ,\ (i_{k-1}, i_k)\ ,\ (i_k, i_1)$$

exist? This is an example of *one* possible cycle on these nodes.

Part (b)

What is the expected number of *any* k-cycles existing on the nodes $(i_1, i_2, \ldots, i_k)$, in any order?

Part (c)

Let the random variable $X_k(i)$ denote the total number of cycles of length k involving node i. Compute $\mathbb{E}[X_k(i)]$.

Part (d)

Using your answer from above, determine the function $g(n)$ that makes the following statement true:

Theorem 1 (Cycles are rare when...). *For any k, as $n \to \infty$, $\mathbb{E}[X_k(i)] \to 0$ iff $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.*

Part (e)

Markov's inequality says that, if X is a nonnegative random variable, then for any $a > 0$,

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}. \tag{4}$$

Using Markov's inequality, prove that, if $f(n)/g(n) \to 0$ as in the previous part, then $\mathbb{P}(X_k(i) > 0) \to 0$ as well. Conclude that, at any node i, as n grows large, it becomes very unlikely that a k-cycle exists on node i.

Part (f)

Consider the case $f(n) = c$ for some constant c independent of n. Do the results of Parts (d) and (e) apply in this case? This corresponds to an Erdős-Rényi with constant degree that does not depend on n. Are cycles rare in this graph?

Problem 6. (Branching Process)

Consider the following random model of a tree.

Fix a discrete probability distribution $p = (p(0), p(1), p(2), p(3), \ldots, p(\ell))$. We require $p(i) \geq 0$ and $\sum_i p(i) = 1$. Start with a single node, which we'll call v_1. Node v_1 starts out "active."

Repeat the following process until there are no "active" nodes:

- For each "active" node j:
  - Create X new nodes, where X is a a random variable with probability distribution p.
  - Connect each of these new nodes to j.
  - Each of the X new nodes become "active," but j now becomes "inactive."

This is a *Galton-Watson branching process* with *offspring distribution* p. For a visual on how a tree generated by this process might look, check Newman's Figure 11.3.

Part (a)

Let Y_k be the total number of new nodes added in timestep k. Prove rigorously that $\mathbb{E}[Y_k] = \mu^k$, where $\mu = \sum_i ip(i)$.

Hint. You might wish to use and cite Wald's Theorem.

Part (b)

Consider the number $N = 1 + \sum_{k=1}^{\infty} Y_k$, the total number of nodes in the tree generated by this process (including the initial node). N is a random number. Determine in terms of μ a necessary and sufficient condition for $\mathbb{E}[N] < \infty$ (i.e. $\mathbb{E}[N]$ exists and is equal to a finite number).

Note. You might find it useful to compute $\mathbb{E}[N] = 1 + \sum_{k=1}^{\infty} \mathbb{E}[Y_k]$. Technically speaking, it's not guaranteed that you can distribute expectations over infinite sums. You may assume here that this is allowed without further justification.