Readings

Newman Chapter 11 (but it's ok to skip 11.6).

Problem 1.

Find two posts on Campuswire from Short Essay 2 that seem like cool things that you might enjoy working on for your course project. For each one, write a long comment (one substantial paragraph) in which you briefly discuss the following items:

- What is something that you find especially interesting about the topic proposed?
- What is one question that you suggest the original author think more about?
- What is one possible risk that you can see that could prevent the project from being fully successful?
- What is one topic related to the theory or data science of networks that you would need to learn more about in order to work on this project?

Take screencaps of each of your two posts and include them in your submission to Gradescope.

Problem 2.

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

$$p(n) = \frac{f(n)}{n} \tag{1}$$

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 probability of any k-cycle existing on the nodes $(i_1, i_2, \ldots, i_k)$?

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)

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. This can be done with a standard probability inequality. Conclude that, at any node i, as n grows large, it becomes very unlikely that a k-cycle exists on 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 3.

A *tree* is a connected graph that contains no cycles.

Part (a)

Let T be a tree with n nodes. Determine the number of edges in T, and rigorously support your result.

Part (b)

A *path* is a walk that does not contain any repeated vertices. Prove that, between any two nodes i and j in a tree, there exists exactly one path.

Part (c)

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. We'll also discuss this model in class a bit.

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 (d)

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.

Problem 4.

Newman, Exercise 11.1

Problem 5. (2 points)

Yes, this problem is worth the equivalent of **two** standard problems. There's still no partial credit though – you need to complete the entire problem to a high standard in order to earn the two points.

You may complete this problem in a programming language of your choice. I have only tested it using Python and the NetworkX package.

First, write a function called `ER_comparison`, which accepts an undirected graph `G` as an argument. Here's what this function should do:

i. Generate an Erdős-Rényi random graph, which you can call `ER`, with the same number of nodes and expected mean degree as `G`.

ii. Compute each of the following four for both `G` and `ER`:

- The mean degree for each graph.
- The transitivity (global clustering coefficient) for each graph.
- The number of connected components for each graph.
- The betweenness heterogeneity for each graph. This number is

$$\frac{\text{standard deviation of betweenness centrality}}{\text{mean betweenness centrality}} . \tag{2}$$

  The standard deviation and mean are both computed over nodes. A larger value of the heterogeneity indicates that there is a small number of very important nodes, while a lower value indicates that most nodes are almost equally important.

iii. Finally, print the numbers at two significant figures in a readable table.

On the next page, I've included an example of expected output for the *Les Miserables* network that is included with NetworkX.

Second, locate three undirected graph data sets from online sources. Some useful sources are listed here. It's best to choose data sets that aren't too large; 100-1,000 nodes is a good size. For each network, describe briefly where you found it and what it means. Show the results of applying your function `ER_comparison` to each of these three networks. Briefly comment on your results in each case.

Third, include both your output and code in your submission to Gradescope.

Hint. You do NOT need to implement algorithms for transitivity, connected components, or betweenness centrality. Packages like NetworkX have functions for these built in! Search the docs and find them.

Measure	Original Graph	ER
Mean degree	6.60	6.13
Transitivity	0.50	0.05
Betweenness Heterogeneity	3.30	0.71
Connected Components	1.00	1.00