CSE416A

HOMEWORK 3

M. Neumann

Due: TUE 23 OCT 2018 10AM

This homework consists of 3 problems.

SUBMISSION INSTRUCTIONS

- **written work**
  - needs to be submitted electronically in *pdf format* via GRADESCOPE
  - start every problem on a *new page*
  - we prefer *typed submissions*, e.g., using LaTeX (if we cannot read your handwriting we cannot give you credit)
- **code**
  - needs to be submitted in form of a submission to the corresponding GRADESCOPE programming assignment (intructions can be found on the course webpage)
  - make sure to not change the *file name(s)* and follow the *formatting instructions* (otherwise we cannot grade your submission)
- **group work** (up to 2 students)
  - make a **goup submission** via GRADESCOPE (one submission per team) for both written work and code submission

PIAZZA

If you haven't done so already, sign up to Piazza (`https://piazza.com/wustl/fall2018/cse416a`). All course and homework related announcements will be made there. Ask **all questions** on Piazza using the appropriate tags.

GRADING RESULTS AND REGRADES

Grades will be uploaded to Canvas and detailed grading comments will be provided via GRADESCOPE . You will be notified viaGRADESCOPE when the grades are published. All regrade requests need to made via GRADESCOPE **within one week** of this announcement.

PROBLEM 1: Network Models (30%)

One of the goals of complex network analysis is to find mathematical models that characterize real-world networks. Those models can then be used to generate new networks in a controlled way. Further, we can compare network measures of real-world network to those of our mathematical models. In this problem, we will explore two famous models—Erdős-Rényi and Small World—and compare them to the statistics of the real-world academic collaboration network computed in the last homework. Note that in this problem all networks are *undirected*.

- *Erdős-Rényi Random graph ($G(n, m)$ random network):* Generate a random instance of this model by using $n = 5242$ nodes and picking $m = 14484$ edges at random.
- *Small-World Random Network:* Generate an instance from this model as follows: begin with $n = 5242$ nodes arranged as a ring, i.e., imagine the nodes form a circle and each node is connected to its two direct neighbors (e.g., node 399 is connected to nodes 398 and 400), giving us 5242 edges. Next, connect each node to the neighbors of its neighbors (e.g., node 399 is also connected to nodes 397 and 401). This gives us another 5242 edges. Randomly Finally, randomly select 4000 pairs of nodes not yet connected and add an edge between them. In total, this will make $m = 5242 \times 2 + 4000 = 14484$ edges. (You will have to write your own code to construct instances of this model.)
- *Real-World Collaboration Network:* Reuse your computations from the last homework.

1.1 **Graph Generation**
Generate a random graph from both the Erdős-Rényi (i.e., $G(n, m)$) and Small-World models and read in the collaboration network.

HINT: You may use NetworkX to generate the Erdős-Rényi graph. However, you need to write a routine to generate a random instance of the Small-World model as described above. You may use the NetworkX function in the generation process, but be careful to set the parameters correctly. Note that there is **no** rewiring in our small-world network.

Comment all your code to receive maximum credit. Add the code for generating the ER and SM graph instance to your written submission.

1.2 **Degree Distribution**
Let p_k be the *degree distribution* of a network. p_k gives us the probability that a randomly chosen vertex has degree k.

(i) Plot the degree distribution of all three networks in the same plot on a log-log scale. Add the plot to your written submission. No code submission required.

(ii) In one to two sentences, describe one key difference between the degree distribution of the collaboration network and the degree distributions of the random graph models.

1.3 **Excess Degree Distribution**
Let q_k be the *excess degree distribution* of a network. q_k gives us the probability that a rondomly chosen edge goes to a node of degree $k + 1$.

(i) Plot the excess degree distributions of all three networks in the same plot on a log-log scale. Add the plot to your written submission. No code submission required.

(ii) In one to two sentences, describe one key difference between the degree distribution and the excess degree distribution of the collaboration network.

(iii) Compare the expected degree and the expected excess degree of the collaboration network to the ones of the random networks. Is the average number of collaborator and the average number of collaborators of collaborators surprising? No code submission required.

1.4 **Clustering Coefficient** The clustering coefficient gives us an idea of how tightly knit local groups of nodes are. We can use it to measure local structure and investigate triadic closure.

(i) Compute and report the average clustering coefficient of the two random networks and

(ii) Compare the average clustering coefficients of the random networks and the collaboration network. Which network has the largest clustering coefficient? In one to two sentences, explain.

1.5 Briefly discuss why having null models like the ER and SM graph models is beneficial for comparing network measures computed for real-world networks.

PROBLEM 2: Power Laws - Mathematical Analysis (20%)

In this problem, you will analyze the power-law distribution mathematically. **Note**: Use *continuous distributions* for all derivations as they tend to be simpler than those for discrete distributions. So, it is common to approximate discrete power-law behavior with its continuous counterpart for the sake of mathematical convenience.

2.1 Derive the **probability density function** (PDF) of the power law distribution. HINT: you will have to derive the *normalization constant* Z for $P(k) = Zk^{-\alpha}$.

Note that $P(X = k)$ diverges for $k \to 0$, so we set a minimum value for $k_{min} \geq 1$ and we look at the power-law distribution in $[k_{min}, \infty]$.

2.2 Derive the **expected value** of a power-law random variable. Note that the power-law degree exponent for real world networks is typically $2 < \alpha < 3$. What is the expected value for $\alpha \leq 2$?

2.3 The **variance** of a power-law random variable is given by $Var(k) = k_{min}^2 \left(\frac{\alpha-1}{\alpha-3}\right)$. Comment on the variance for the typical values of α ($2 < \alpha < 3$). What does that mean for the expected value?

2.4 When studying the degree distributions for large networks, we are interested in the the *tail distribution*. The tail distribution is measured by the **complementary cumulative distribution function** (CCDF). Show that the CCDF of the power-law distribution is given by $P(X \geq k) = \left(\frac{k}{k_{\min}}\right)^{-(\alpha-1)}$.

Note that the CCDF is itself a power-law distribution with exponent $\alpha' = \alpha - 1$. So, you can estimate α from fitting a the power-law distribution on CCDF (cf. Problem 3.4).

PROBLEM 3: Power Laws - Empirical Study (50%)

See `hw3_problem3.ipynb` for instructions.