COMP 6730 Fall 2020

Homework 2

Due 11:59pm Thursday 10/22/20 in Blackboard

General Instructions

Submission instructions: You must **submit a print-out of your programs** and the output for each subproblem. The *majority of points will be given to the programs.* You will lose points if they are not clear or readable.

Questions

1 Network Characteristics [35 points]

One of the goals of network analysis is to find mathematical models that characterize real-world networks and that can then be used to generate new networks with similar properties. In this problem, we will explore two famous models—Erdős-Rényi and Small World—and compare them to real-world data from an academic collaboration network. Note that in this problem all networks are *undirected.* You may use the starter code `q1-starter.py` for this problem in the homework folder.

- *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. Write code to construct instances of this model, i.e., do not call a SNAP function.

- *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. Finally, randomly select 4000 pairs of nodes not yet connected and add an edge between them. In total, this will make $m = 5242 \cdot 2 + 4000 = 14484$ edges. Write code to construct instances of this model, i.e., do not call a SNAP function.

- *Real-World Collaboration Network:* Download this undirected network from homework folder ca-GrQc.txt.gz. Nodes in this network represent authors of research papers on the arXiv in the General Relativity and Quantum Cosmology section. There is an edge between two authors if they have co-authored at least one paper together. Note that some edges may appear twice in the data, once for each direction. Ignoring repeats and self-edges, there are 5242 nodes and 14484 edges. (Note: Repeats are automatically ignored when loading an (un)directed graph with SNAP's `LoadEdgeList` function).

1.1 Degree Distribution [10 points]

Generate a random graph from both the Erdős-Rényi (i.e., $G(n, m)$) and Small-World models and read in the collaboration network. Delete all of the self-edges in the collaboration network (there should be 14,484 total edges remaining).

Plot the degree distribution of all three networks *in the same plot* on a log-log scale. In other words, generate a plot with the horizontal axis representing node degrees and the vertical axis representing the proportion of nodes with a given degree (by "log-log scale" we mean that both the horizontal and vertical axis must be in logarithmic scale). 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.

1.2 Excess Degree Distribution [15 points]

An important concept in network analysis is the *excess degree distribution*, denoted as q_k, for $k \geq 0$. Intuitively, q_k gives the probability that a randomly chosen edge goes to a node of degree $k + 1$. Excess degree can be calculated as follows:

$$q_k = \frac{q'_k}{\sum_i q'_i}, \qquad q'_k = \sum_{i \in V} \sum_{(i,j) \in E} I_{[k_j = k+1]},$$

where $I_{\text{condition}} = 1$ when condition is true and 0 otherwise. V denotes the set of nodes, E the set of edges and k_j the number of neighbors of node j (equivalently, the degree of node j). Additionally, the *expected excess degree* is $\sum_{k \geq 0} k \cdot q_k$, and the *expected degree* is $\sum_{k \geq 0} k \cdot p_k$, where p_k is the proportion of nodes having degree exactly k.

1.2 (a) [**5 points**] Show how to compute the excess degree distribution $\{q_k\}$ given only the degree distribution $\{p_k\}$.

1.2 (b) [**10 points**] Plot the excess degree distributions of all three networks in the same plot on a log-log scale. In one to two sentences, describe one key difference between the degree distribution and the excess degree distribution of the collaboration network. Then compute and report the expected degree and the expected excess degree for each network.

1.3 Clustering Coefficient [10 points]

Recall that the local clustering coefficient for a node v_i was defined in class as

$$C_i = \begin{cases} \frac{2|e_i|}{k_i \cdot (k_i - 1)} & k_i \geq 2 \\ 0 & \text{otherwise,} \end{cases}$$

where k_i is the degree of node v_i and e_i is the number of edges between the neighbors of v_i. The *average clustering coefficient* is defined as

$$C = \frac{1}{|V|} \sum_{i \in V} C_i.$$

Compute and report the average clustering coefficient of the three networks. For this question, write your own implementation to compute the clustering coefficient, instead of using a built-in SNAP function.

Which network has the largest clustering coefficient? In one to two sentences, explain. Think about the underlying process that generated the network.

What to submit: All the programs and the following results from your program:

1.1:
- Log-log degree distribution plot for all three networks (in same plot)
- One to two sentence description of a difference between the collaboration network's degree distribution and the degree distributions from the random graph model.

1.2:
- (part a) Log-log excess degree distribution plot for all three networks (in same plot)
- (part a) One to two sentence description of the difference in the distribution of the degree and excess degree distributions for the collaboration network.
- (part a) Expected degree and expected excess degree for each network.
- (part b) Short proof showing how to calculate $\{q_k\}$ in terms of $\{p_k\}$.

1.3:
- Average clustering coefficient for each network.
- Network that has the largest average clustering coefficient.
- One to two sentences explaining why this network has the largest average clustering coefficient.

2 Bowtie Structure of Non-Web Networks [20 points]

In this problem, we will explore the structure of a directed social network, namely the Epinions Social Network (dataset and more information available at http://snap.stanford.edu/data/soc-Epinions1.html) and a communication network, namely the EU Email Communication Network (dataset and more information available at http://snap.stanford.edu/data/email-EuAll.html). We will use methods similar to the ones Broder et al. employed when they determined that the web graph is structured like a bowtie (which we discussed in class).

2.1 Node Position [4 points]

Consider the nodes with IDs 9809 and 1952 in the Epinions network, and 189587 and 675 in the Email network. Use forward (i.e., following the outwards links) and backward (i.e., following the inwards links) BFS starting at these nodes to determine whether they belong to the SCC, IN, or OUT components. (*Hint*: You may want to use the SNAP function `GetBfsTree`.)

2.2 Random-start BFS [8 points]

For each of the two networks, choose 100 nodes at random and do one forward and one backward BFS traversal for each node. How many nodes can you reach each time, for each of the two traversals? What behavior do these traversals exhibit and what can you infer from them about the graph structure? Plot the cumulative distributions of the nodes covered in these BFS runs, like in the paper by Broder et al. (Figure 1 below). Create one figure for the forward BFS and one for the backward BFS.

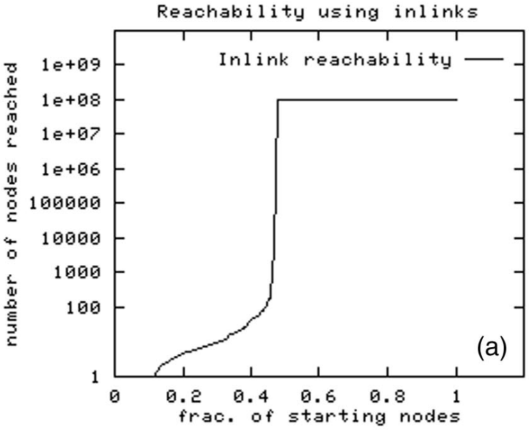

Figure 1: Cumulative distribution on the number of nodes reached by backward BFS started from randomly chosen nodes.

2.3 Size of Bowtie Regions [8 points]

Determine the sizes of the regions of the two networks using the data obtained by running the BFS experiments in the previous question. How many nodes are in the SCC, IN, OUT, TENDRILS, and DISCONNECTED regions of each of the two networks?