Social Network Analysis
Spring 2019
MS(DS)
Assignment 2
Submission Date: March 8, 2019

In this Assignment, we explore the structure of a directed social network, namely the Epinions Social Network (file attached). This is a who-trust-whom online social network of a general consumer review site Epinions.com. Members of the site can decide whether to "trust" each other. All the trust relationships interact and form the Web of Trust which is then combined with review ratings to determine which reviews are shown to the user.

Problem 1

In this problem, we will try to inquire the similar structure the ones Broder et al. employed in their seminal paper where they determined that the web graph is structured like a bowtie. The authors discovered that the web graph had a large strongly connected component (SCC) which could be reached from any node in IN, and could go to any node of OUT. There were also TENDRILS hanging off IN and OUT, containing nodes reachable from portions of IN or nodes going to portions of OUT. TENDRILS going from IN to OUT without touching SCC formed TUBES. There are also some DISCONNECTED components isolated from the rest of the graph.

We will try to determine whether bowtie structure exists in Epinions network or not. For that we will compute the sizes of each region in the Epinions network. How many nodes are in the SCC, IN, OUT, TENDRILS+TUBES (referring to TENDRILS and TUBES combined), and DISCONNECTED regions? Write a python program to calculate the size of each of the above components in Epinions network(both number of nodes and fraction of total nodes) (Hint: You may want to use the SNAP functions GetMxWcc and GetMxScc along with BFS.)

Broder et al. found in their paper that given a pair of randomly chosen start and finish webpages, one can get from the start page to the finish page by traversing links only approximately 25% of the time. For the Epinions network, what is the probability that a path exists between two nodes chosen uniformly from the graph? (Hint: One way you can approach this question is to first sample many node pairs at random and then report the fraction of times pairs were reachable. You may find the SNAP function GetShortPath useful here.) Start with 10 pairs and then double the node pairs. Plot a graph that has number of pairs on the x-axis and fraction of reachable pairs on the y-axis. As the number of node pairs sampled grows larger, what would you expect the fraction of reachable pairs to converge to?

Problem 2

As you know that page rank algorithm can be used to determine the ranks/importance of nodes in the network. Apply page rank algorithm to the Epinions network. Do not use GetPageRank function of SNAP to compute the results. Device your own function for this problem. Take value of β (teleportation probability), and ϵ (termination criteria) from the user to compute your results. Use google algorithm's approach to solve dead end problem (this means do not modify the matrix M and uniformly divide all the leaked page rank to each node in the network).

Report

Submit a report that includes all your findings. The report must include the following:

- Number of nodes in SCC, IN, OUT, TENDRILS+TUBES, and DISCONNECTED regions of Epinions network.
- Graph of number of pairs Vs. fraction of reachable nodes
- What is your answer to probability of path existence when a pair of nodes is selected uniformly at random based on your experiment?
- Top 10 ranked nodes along with their ranks in Epinions network when $\beta = 0.8$ and $\epsilon = 0.001$
- Also state the following for each of the top 10 ranked nodes in the network. Let x be the node:
    - Number of incoming edges (indegree of x)
    - Ranks of all the source pages having hyperlinks toward x

Important Note:

Submit your code files at google classroom as one single file. If you have more than one files compress them and submit them as single file. The name of file must your complete roll number. Your code must be properly commented.

This is an individual assignment and the code submitted must be your own contribution. Any sort of plagiarism will be dealt seriously and may lead to severe consequences including negative marking.