CMPSCI514: Algorithms for Data Science

Mini Exercise 2

Instructor: Barna Saha Posted: Apr 8th, Due: April 22nd at 11:55 pm

Do not look up materials on the Web. You can consult the reference books mentioned on the course website, and also the class slides for solving the homework problems. You may work in a group of size at most 2. *Submit one homework solution per group. No late homework will be accepted.*

For programming assignments, submit your code with a detailed readme file that contains instruction for running it. Also include any test dataset that you have used and results obtained to show correctness of your implementation.

Total Points: 50

Exercise 1. *In this assignment, the goal is to implement a set of simple Map Reduce tasks. Please include the Python scripts used in the submitted report. You will work with a collection of e-mail data downloadable from:* `https://snap.stanford.edu/data/email-EuAll.txt.gz` *The data forms a graph G of e-mails between users, with each line being of the form* ***sender receiver****. Compute the following on G:*

- *Number of nodes in the graph*
- *Average (and median) indegree and out degree*
- *Average (and median) number of nodes reachable in two hops [*
- *Number of nodes with indegree* > 100