CS:4440 Web Mining
Spring 2020
Homework 1

Points**:** 100 (15%)

Due Date: February 27, 2020

Goal: To gain understanding in analyzing networks.

You are given a file of data representing a network. The data represents a network formed from around 25000 users (nodes). These users form a sample of users who posted to the top 500 subreddits during January 2014. There is an edge between two users if they have commented in the same post.

The network data has been provided in file (network.psv) with the following format:

Each row in the file corresponds to a user.

The pipe '|' separates the source node (user) from the destination nodes (users). As an example the entry:

lolnymous|whtisthis, iMantorras

means that there is an edge between user lolnymous and user iMantorras, and there is also an edge between user lolnymous and user whtisthis.

Since the network edges are undirected the file also has entries in the reverse direction, such as:

whtisthis|lolnymous

and

iMantorras|lolnymous

I. Make appropriate measurements to analyze the network. Look over your readings to get ideas. (60 points)

For this question you will be graded on your choice of properties to explore and your ability to analyze results, i.e., go beyond stating numbers.

II. Specific questions

a) Does the power law fit the degree distribution for this network. If so, what is the value for the coefficient and what are the implications of this coefficient value? (20 points)

b) Discuss the small world phenomenon (and any related concepts) in the context of the given network. (20 points)

Note 1: You may use any software of your choice to perform this analysis. What is important is how cogently and concisely you interpret the observations you make.

Note 2: If you have any questions in terms of processing the input file (format questions etc.) please contact the course TA: Osama Khalid.