CS 6220 - Fall 2015 Olga Vitek

Homework 9

Each part of the problems 5 points
Due on Blackboard by **5pm on Thursday December 10**.

1. We will explore the gene regulatory network that we briefly saw during the class. The dataset and the example code are posted on the course website.

   (a) For each node, calculate its degree, betweenness centrality, and pageRank. Plot the histograms of these three features.

   (b) Implement the HITS algorithm described on p. 104 of the notes, and calculate the 'hub' and the 'authority' scores for each node. Plot the histograms of these two features. *Hint:* Use `as.matrix(get.adjacency(g.100))` to extract the adjacency matrix from a graph structure. See `https://www.ccs.neu.edu/course/cs6220f15/ssl/8-graphs.R` for an example of R code with matrix multiplication.

   (c) Make pairwise scatterplots that compare the five features above (i.e., make plots, where x axis is one feature, y axis is another feature, and a point is a node in the graph). Provide one example of a node where at least two metrics disagree. Plot a subnetwork that contains that node, and its direct neighbors.

2. We will explore the running time required to work with networks of different size. For the questions below, use the function `sample_gnp` in `https://www.ccs.neu.edu/course/cs6220f15/ssl/8-graphs.R` to generate random graphs. (*Note:* the runtime may depend on your computer. You can adjust the total number of nodes and the length of the graph sequence, such that you can perform the computation and can see the difference between the methods and/or between the graphs).

   (a) Fix the probability of an edge to 1/20. Create a sequence of graphs, where the number of nodes is `seq(from=10, to=210, by=20)`. Make a graph where the a axis is the number of nodes, and the y axis is the runtime of pageRank on each of the graph. Overlay a similar plot, where the y axis is the runtime of the HITS algorithm that you implemented above.

   (b) Repeat (a), by changing the probability of an edge to 1/3.

   (c) Discuss the reasons for the differences in runtime between the graphs and/or the analysis methods.