Homework 8

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due April 28, 2016

In this assignment, we will cluster a sample of Wikipedia articles. This dataset was originally collected by Lada Adamic, who used it for her *Social Network Analysis* class.

To complete this assignment, I recommend that you start with the sample `hw08.Rmd` code from the course webpage.

1. Download `wikipedia.gml` and read it into R using the `read.graph` function from the `igraph` package. You will need to specify `format="gml"` when you read in the graph. How many vertices are in the graph? How many edges?

2. Use the `cluster_fast_greedy` algorithm on the entire wikipedia graph to cluster all the vertices in the network. For each of the five largest clusters, report 10–20 of the vertices in the cluster. Do the clusters make sense to you?

3. Choose three vertices in the graph having degrees between 10 and 50. (Hint: you can use the `degree` function to get the vertex degrees). Try to get a range of topics. For each of the vertices, use the `make_ego_graph` and the `plot` function to plot the 1-hop neighborhoods around the vertices.

4. For each of your three ego graphs, produce a plot showing the clusters found by the `cluster_fast_greedy` algorithm (applied to *the entire network*, not just the ego network). Do the clusters seem reasonable in these ego networks?

5. Experiment with some of the other clustering methods listed in the documentation for `communities`. Do the clusterings found by these methods differ that much? (Hint: you can use the `compare` function to measure the differences between the different clusterings.)

6. On the wikipedia graph, which clustering method seems to give the best clusters?