ST5225: Project

- The due date of the report is on Nov. 13th, 2020.
- This is an INDIVIDUAL project. Discussion allowed but no cheating.
- Please write your answer in the form of a REPORT, not answers to an assignment. I have explained what each section should be about in this problem sheet, please follow it. If the form is incorrect, you may suffer some mark deductions.
- Codes should be submitted together. R/Python are accepted.

A real data example: Political blogs

We are interested in the analysis of political blogs. Now we have a network of blogs about the posts around the time of the 2004 US presidential election. The edges were automatically extracted from a crawl of the front page of the blog. The labels denote whether the blog is liberal (denote by 0) or conservative (denote by 1). Some were labelled manually, which means there might be errors. The network was compiled by Lada A. Adamic and Natalie Glance.

Download the dataset from LumiNUS. Load the data set *polblogs.gml* to R. Based on this data, please generate a report containing the following sections.

1. Introduction section. In this part, please first plot the data and give a short description of this network (including: directed/undirected, multi-graph/simple, connected/unconnected, number of nodes and edges, properties of nodes/edges (if exist), etc.). Simplify the network with the command “simplify(g)”, and then answer the following questions based on the simplified network.

   (a) Check the density of the network, average distance and the diameter.

   (b) Subsection about degree. For the total degree (if the network is directed), find the degree vector and draw the degree distribution. If the distribution has the shape of power-law distribution, then find the corresponding parameter α and comment on it.

   (c) Fill in Table 1 and give some comments.

Methods	In-degree	Out-degree	Closeness	Betweenness
1				
2				
3				
4				
5				
6				

Table 1: The top 6 most important nodes with in-degree centrality, out-degree centrality, closeness centrality, and betweenness centrality.

2. Section about ranking pages. In this section, please rank the pages according to

   - the hub score;
   - the authority score;
   - the PageRank algorithm;
   - the PageRank algorithm with damping parameter 0.85.

   List a table similar as Table 1, which contains the top 10 webpages from the 4 methods. Compare the results and give comments.

3. In Sections 3, 4, 5 and 6, we consider the undirected graph by eliminating the directions of the edges in the network polblogs. It can be done with the command "as.undirected(polblogs)". Note that the current network is the simplified network.

   First, we check the basic properties of this undirected network, including number of edges/nodes in this network, density, average distance and the diameter of the network. Then, we check the following items:

   (a) Subsection about degree. For the total degree (if the network is directed), find the degree vector and draw the degree distribution. If the distribution has the shape of power-law distribution, then find the corresponding parameter α and comment on it.

   (b) Subsection about components. If the network is not connected, please check the number of connected components and corresponding sizes. Find the giant component, denoted as $\tilde{G}$.

   (c) Subsection about coreness. Find the coreness of each node. Say the largest coreness is k, then draw the k-core. How many nodes in k-core?

   (*Note: to erase the names and adjust the magnitude of the node points, you may use the following R command:*)

   ```
   plot(g, vertex.size = 8, vertex.label = NA)
   ```

4. Section about sampling. In this section, apply the following sampling methods to get a subgraph with the same size K:

   - Induced-subgraph sampling
   - Incident subgraph sampling
   - Snowball sampling
   - Respondent-driven sampling.

   Plot the four subgraphs, and for each sample graph, please check

   density of network, connect/unconnect, clustering coefficient,

   and compare with such statistics generated from the original graph.

5. Section about graph partition. Consider the giant component $\tilde{G}$ in Section 3b. We want to apply different graph partition methods to $\tilde{G}$ and compare.

   Note that the network can be divided into 2 groups according to the “value” property of nodes (V(polblogs)$value), denoted as $\ell_i \in \{0,1\}$. We take it as the underlying truth. Suppose with one graph partition method, we got the estimation $\hat{\ell}_i \in \{0,1\}$ for $i \in V$ to denote the group node i belongs to. The error rate is defined as

   $$\text{Error Rate}(\hat{\ell}_i) = \min\left\{\frac{1}{|V|}\sum_{i\in V} 1\{\ell_i \neq \hat{\ell}_i\}, 1 - \frac{1}{|V|}\sum_{i\in V} 1\{\ell_i \neq \hat{\ell}_i\}\right\}.$$

   - Find the graph partition by removing the edges with highest betweenness, given that there are 2 groups. Compare with the truth.
   - Find the graph partition by hierarchical clustering with Euclidean distance and 3 types of linkages. Cut the tree at $k = 2$. Compare with the truth.
   - Find the graph partition by the modularity method given there are 2 communities. Compare the modularity score of it and the truth. Find the corresponding error rate.
   - New graph partition method based on stochastic block model. Suppose the data is generated from a stochastic block model, then the following method is proposed.

     (a) Find the 1st and 2nd eigenvector of the adjacency matrix of $\tilde{G}$, denoted as v_1 and v_2.

     (b) Calculate the entry-wise ratio $v_2(i)/v_1(i)$ for $i = 1, \cdots, n$, denoted the resultant vector as r.

     (c) Apply k-means to r to divide find the labels of the nodes by the command

     ```
     kmeans(r, centers = 2)
     ```

     Realize this new method, and calculate the error rate.

6. In this section, we try to fit stochastic block model to this network. According to the labels of the nodes, please give an estimate of P under the stochastic block model assumptions.