Discovery Thread: Project 2

In this project you will apply the techniques for random graphs, community detection and data embedding to a specific data set.

On the weighted undirected graph dataset either assigned to your project perform the following three tasks:

I. Random graph model testing:

Point Estimation:

1. Under the Erdos-Renyi random graph model, estimate the parameter p. Compute the estimated number of 3-cliques and 4-cliques and compare them to the actual numbers of 3-cliques and 4-cliques in your data set.

2. Under the SSBM random graph model, estimate the parameters a and b based on the number of vertices, edges, and 3-cliques. Compute the estimated number of 4-cliques (under the SSBM model) and compare this predicted number to the actual number of 4-cliques.

Sequence of 4-cliques prediction:

1. Create an ordered sequence of edges accoring to their weight. Specifically, order the edges according to the weight, starting with the largest weight first and then continue in a monotonic decreasing order. To do so, create a data file, say graph.dat, from the data file assigned to your project, that lists the edges in the appropriate order, and has the following format:

   ```
   First line: n m
   Second line: Edge1Vertex1 Edge1Vertex2
   Third line: Edge2Vertex1 Edge2Vertex2
   ...
   m+1st line: EdgemVertex1 EdgemVertex2
   ```

2. For the sequence of edges (and graphs) perform the following computations:

   (a) Under the Erdos-Renyi random graph model, for each graph in the sequence, estimate the parameter p, and compute the expectation of the number of 4-cliques; On the log-log plot, determine the best linear fit, $log(X_4) = a_{ER}log(m) + b_{ER}$, where m is the running number of edges; discard the first values of m when there are no 4-cliques.

   (b) Under the SSBM random graph model, for each graph in the sequence, estimate parameters a and b and compute the expectation of the number of 4-cliques; On the log-log plot, determine the best linear fit, $log(X_4) = a_{SSBM}log(m) + b_{SSBM}$, where m is the running number of edges; discard the first values of m when there are no 4-cliques.

(c) For each graph in the sequence, compute the actual number of 4-cliques, $X_4(m)$, and determine the best linear fit, $log(X_4) = a_0 log(m) + b_0$, where m is the running number of edges; discard the first values of m when there are no 4-cliques.

3. Overlay in the same plot the graphs of $log(X_4)$ and the prediction under Erdos-Renyi and SSBM models of the number of 4-cliques. Print also the parameters a_{ER}, a_{SSBM}, a_0 and b_{ER}, b_{SSBM}, b_0.

II. Community detection:

Implement the six community discovery algorithms (partition algorithms) and run them on your project data set.

Specifically, implement:

1. Spectral methods using: W, Δ, and $\tilde{\Delta}$
2. SDP relaxation algorithms using: W, Δ, and $\tilde{\Delta}$

III. Data Embedding

Implement the Laplacian Eigenmap and the Local Linear Embeding (LLE) algorithms, and run them on your project data set.

Specifically, implement and run:

1. Laplacian Eigenmap data embedding for target dimension $d = 2$;
2. LLE dimension reduction after Laplacian Eigenmap data embedding:
   (a) First run the Laplacian Eigenmap data aembedding algorithm to create a geometric graph $\{x_1, \ldots, x_n\} \subset \mathbb{R}^N$ with $N = 10$;
   (b) Then implement and run the dimension reduction LLE algorithm on the this geometric graph to reduce dimension to $d = 2$; use $K = 2d = 4$.

Regarding LLE: Note the W matrix at step 2.1 is the matrix whose (i, j) elements were computed at 1.5. This is NOT the weight matrix loaded from your data set!