Discovery Thread: Project 1

Consider a dynamical graph model where the graph growth from a set of isolated vertices to a complete graph by adding one edge at a time. Given two ordered lists of edges the target is to discover which list is more likely to be associated to a percolation growth model rather than a random graph model. Towards this goal, focus on the evolution of four computable features: the number of 3-cliques, the number of 4-cliques, the spectral gap (second smallest eigenvalue of the normalized Laplacian matrix), and the minimum number of edges for the graphs to be connected. The two models (hypotheses) are:

H0: The Random Graph Model: At each step, the next edge is generated randomly with equal probability among the remaining set of edges.

H1: The Percolation Model: Vertices correspond to points in a low dimensional vector space, and the edges are sorted ascendingly acording to their length.

1. Under the random graph hypothesis (H0) for a constant probability p for each edge, derive the Maximum Likelihood Estimator (MLE) p_{MLE} for p when the graph has m edges and n vertices. For the number n of vertices in your dataset, plot p_{MLE} as function of number of edges m, when m varies from 0 to the maximum number of edges your dataset constains.

2. Under H0, compute the expected numbers of 3-cliques and 4-cliques as functions of the number of vertices n and the probability p. Then substitute the MLE estimate obtained at 1. to obtain the expected numbers of 3-cliques and 4-cliques as functions of number of edges. Call these functions $N_3(m)$ and $N_4(m)$. Plot these functions for your data set size (i.e. for the number of vertices n and the number of edges) in both normal and log-log plot.

3. Write a code that counts the number of 3-cliques at each step, for the two datasets (lists). Obtain two sequences, one associated to each list, indexed by the number of edges, and then plot them. Call these functions $E^1(m)$ and $E^2(m)$ respectively.

4. Use the least-squares procedure to estimate the power exponent and the offset for both $log(E^1(m))$ and $log(E^2(m))$ as functions of $log(m)$.

5. Based on the results at 2. and 4. determine which list is more likely to be associated to which model.

6. Use only the first half of datasets to recompute the least-squares estimates at 4. Do you obtain different results? Does the conclusion change?

7. Repeat previous questions for the second half of the datasets.

8. Determine the first index the graphs become connected. Let m_1 denote the number of edges of the first connected graph in the first dataset, and m_2 in the second dataset. Compare these two numbers to the critical threshold under the H0 model.

9. Plot the second smallest eigenvalue of the normalized Laplacian for the two sequence of graphs. Can you compare these plots to what the theory suggests for random graphs (H0 model)?
10. (Optional) Can you repeat 3.,4.,5. for the statistics of 4-cliques instead of the statistics of 3-cliques?