CS-558: Data Mining Spring 2018

Empirical Study 4: Topological Mining of Graphs

Instructor: Mehmet Koyutürk

Presentation Date

Friday, May 4, 2018.

Datasets

For this exercise, we will use network data from the Stanford graph database at `https://snap.stanford.edu/data/`. Please select at least three networks from

Questions

1. For each of the networks, plot the degree distribution of the network (using log-scale for both the degree and relative frequency might produce more meaningful results). Do you see a pattern?

    - For comparison purposes, generate the following random graphs with number of vertices and edges equal to each network, and plot the degree distribution for these random networks as well. Which model generates degree distributions that are more similar to that of the original networks? Why?
    - $G(n,p)$: There are n vertices in the network and there is an edge between any pair of nodes witrh probability p. Set n to number of nodes in the original network, set $p = m/\binom{n}{2}$, where m is the number of edges in the original network.
    - *Preferential attachment:* Start with two nodes connected by an edge. Iteratively add $n-2$ nodes such that the probability of adding an edge between a new node and an already existing node is proportional to the number of existing edges of the already existing node. Normalize the probabilities in such a way that the expected number of edges in the resulting graph will be m.

2. For each of the networks, plot clustering coefficient as a function of degree (you may want to bin degrees logarithmically). Is there any relationship between clustering coefficient and degree?

    - Repeat this analysis for the two random network models above. Also generate a "permuted network" by repeatedly swapping edges in the original network. Repeat this analysis for the permuted network. Do you observe a relationship between degree and clustering coefficient? Which random network model generates networks that are similar to the original networks in terms of the relationship between degree and clustering coefficient?

3. Define a criterion to assess the "assortativity" of a given network, where a network is said to be assortative if high-degree nodes are more likely to be connected to each other as compared to low degree nodes (i.e., the "traffic" in the network goes through hubs, for example, airline route maps are examples of highly assortative networks). Be careful, high degree nodes are already likely to be connected to each other since they are more likely to be connected to any other node; so your criterion should take this effect into account. Evaluate the assortativeness of the original networks based on your criterion, repeat the same for the random/permuted networks (you may want to repeat this multiple times to generate a distribution), and evaluate the assortativeness of the original networks in comparison to the assortativeness of the random networks.

4. Define a criterion to assess the "clusteredness" of a given network, where a network is said to be clustered if the network is composed of "communities" that tend to be loosely connected within the community, but tend to be loosely connected with the rest of the network. Again, provide an analysis of the clusteredness of the original networks in comparison to the random/permuted networks you have generated above.

Please choose from these questions to put together a coherent presentation that will last around 40 minutes. Feel free to alter the questions as you see fit. Visualize your results in a way that will allow effective interpretation of results. For all questions, derive conclusions that you can defend, identify results that are surprising, confusing, or not clear to you, and/or suggest questions for further investigation.