Homework

Luis Enrique Valenzuela Navarro
Institut Polytechnique de Paris/Telecom SudParis

1 Github Link

https://github.com/lvalenzuelana/Network-Analysis-Modeling

2 Question 2: Social Network Analysis with the Facebook100 Dataset

2.1 What are you able to conclude from these degree distributions?

We can see through the graphs that there is a great variety of students with many connections, but these connections are of low degree. This may be because they are new students to the network and may not have as many friends, or they may be new students who do not know anyone and therefore do not have as many connections.

2.2 Should either of these networks be construed as sparse?

Yes, I consider them as sparse. The values obtained in Global Clustering, Mean Local Clustering and Edge Density show us a low interconnection between the nodes. Which makes sense with the statement made in the previous question.

2.3 Based on these calculations as well as your previous ones, are you able to draw any conclusions about any similarities or differences between the tree networks? What other observations can you make?

We can observe in the three graphs that the nodes are interconnected mostly in low degrees. This reaffirms the hypothesis raised in question (a). If we look at MIT and John Hopkins we see that Local Clustering decreases exponentially as we increase the degree. This means that very few nodes in the network had large connections. A similar behavior is observed in Caltech so we could see that the nodes with the highest degree do not belong to a specific cluster. These people must be very popular in the network and that is why they cannot be grouped in a cluster.

3 Question 3: Social Network Analysis with the Facebook100 Dataset

3.1 Briefly discuss the degree to which vertices do or do not exhibit assortative mixing on each attribute, and speculate about what kind of processes or tendencies in the formation of Facebook friendships might produce this kind of pattern.

With the results of the graphs seen above we can infer the following: People within the same careers or faculties tend to relate to each other to a greater extent than those who are living in the same dorm. As human beings we seek that when creating new relationships, the people who will be part of them must have a similar assortment. There are other attributes such as the number of titles or the majors that have a lesser influence when selecting friendships within the network. The distribution of points spans the line of no assortativity, with all the values above 0. This confirms the expected behavior of people who became in-network friends within Facebook at the time the data was collected.

4 Question 4: Link prediction

4.1 Choose a couple of graphs in the facebook100 dataset run and evaluate each link predictor on them, and conclude on the efficiency of the following metrics: common neighbors, jaccard, Adamic/Adar.

The link prediction method, in principle, provides a similarity score for each nonexisting link and for most methods, a higher score means higher likelihood that the link will appear in the future. As we can see, G1 has a higher number of edges than G2. That is why in G2 the metric that best fits is the Common Neighbors. This is because by having fewer links it is likely that the nodes have more neighbors. This means that in Reed98 it was a smaller network and it was more likely to establish friendships because of mutual friends. In G1 we can see that the metric that best fits is Adamic/Adar. Being a sample with a greater number of links, friendships are more viable between two people if the person they have in common does not know as many people. If this condition is met, it is likely that this mutual friend will introduce them. If we wanted a more stable metric on both, I would use Jaccard. This is because we take more into account the similarities between 2 nodes (in this case people) and the more similar they are, the more likely it is to establish a friendship.

5 Question 5: Find missing labels with the label propagation algorithms

5.1 Conclude on the accuracy of the label propagation algorithm for different labels, could you explain why is there such difference in the accuracy between each type of label?

With the data obtained we can see that the algorithm depends on the labels and how much data with labels we have. When the Gender attribute labels were predicted, the algorithm was efficient because only two labels could be predicted: Male or Female. However, it lost precision when it had to predict more labels in other attributes. I can also conclude that as it has more data it becomes more accurate. When the algorithm had more information about the connections between people, it improved its predictions. The mean absolute error and the precision of both decrease as the percentage of missing values increases, giving evidence of the above.