Wissam OUALI

Homework - NET 4103

Outline

Readme

The following report discusses some of the features of the Facebook100 graph dataset. You will find attached to the following link the notebook containing the whole code and answers.
URL : https://github.com/wouali/NET4103_project

Question 2: Social Network Analysis with the Facebook100 Dataset

Degree distribution

Each of these degree distributions seems to follow a power-law. There are few nodes with high degrees and many nodes with low degrees which means that having much more friends than average is rare. It can be consistent with the fact that we are studying a Social Network.

Since Caltech Network has fewer nodes than John Hopkins and MIT Networks, we can observe that its degree distribution is more skewed towards low degrees compared to the other two.

Finally, we can conclude on that statement : these distributions indicate to us that the three graphs have few hubs.

Clustering coefficient

The global clustering coefficient measures the degree to which nodes in a network tend to cluster together, while the mean local clustering coefficient measures the clustering around individual nodes. The edge density measures the proportion of possible edges that are present in the network.

Based on the density information, we can say that John Hopkins and MIT Networks are sparse (respectively 0.014 and 0.012 edge density) while Caltech Network is denser (with 0.056 edge density). The global clustering coefficient of the latter is also higher than the two others (0.41 vs 0.27 for the two), indicating that nodes in this network tend to cluster together more tightly.

Degree vs. Local clustering coefficient

These 3 graphs show the correlation between Degree and Local clustering coefficient for the Networks we are studying. We can observe that the less a node has connections with other nodes (low degree) the more its clustering coefficient can be high. In other words, high degree nodes tend to have lower clustering coefficients. This is consistent with the **small-world** phenomenon, which suggests that networks tend to have highly clustered regions while still allowing for relatively short paths between nodes.

In terms of differences, we can see that the Caltech network has a higher average clustering coefficient compared to the MIT and Johns Hopkins networks. As we already said, this suggests that the Caltech network may be more tightly connected than the other two networks. On the other hand, the MIT network has a wider range of degrees compared to the other two networks, with more nodes having higher degrees.

Overall, the degree distributions, clustering coefficients, and density information all suggest that the three networks exhibit properties of small-world networks. The differences observed between the networks may reflect underlying structural differences in the social relationships of the individuals represented in each network.

Question 3: Assortativity Analysis with the Facebook100 Dataset

Assortativity is a measure of the tendency of nodes in a network to connect to other nodes with similar or dissimilar attributes.

Assortativity - Student status attribute

Reminder : Student status attribute --> undergraduate, graduate student, summer student, faculty, staff, or alumni.

We can observe that every network shows a positive assortativity, and most of them are significant. In this sense, people are more likely to make friends with others of similar status, which is understandable and plausible.

By printing the networks with the highests and lowests values of assortativity, we can possibly speculate a little bit more on the process of the formation of friendships. In this case, maybe the students in the schools with higher reputation than average (like Princeton and Harvard) take more into account the student status because they have been in a more competitive environment and are therefore more likely to recognize and relate to others with at least the same student status. Indeed this last statement is only speculation.

Assortativity - Major attribute

Reminder : Major --> main focus of a degree's student

At first sight, it seems that we observe more or less the same thing as for the attribute "Student status". However, if we look at the values of the coefficients, we can see that even if they are all positive, they are mostly very close to zero, the maximum being equal to 0.13.

In this sense, we can conclude that there is no strong tendency for this attribute. At best, this tendency is very slightly positive. In other words, people can be interested in making friends with people of the same major but in a very slight way.

Assortativity - Degree

Reminder : Degree --> Number of connections with other nodes

This scatter plot is more sparse than the first two. There are mainly positive values but also few negative values. Hence, the Degree is slightly assortative in these social networks. People with the same social behavior may have a tendency to bound themselves. For example, shy people may link with few people that are probably shy as well, and the same goes for extroverts. Indeed this is not always the case, but the tendency is strong enough to be noted

However, I have no interpretation regarding the Networks with the top and bot values.

Assortativity - Dorm attribute

The Dorm attribute is clearly assortative. This is consistent and understandable, people are indeed more likely to befriend each other if they rest in the same dorm. So this friendship is transposed from reality to the social network. However, sleeping in different dorms does not prevent the students from linking with each other.

It would be interesting to see if the networks with highest values of assortativity for the Dorm attribute are the same as the networks with highest values for the gender attribute (maybe in case of mixed dorms). Let's observe that below.

Assortativity - Gender attribute

cf project statement :

" The distribution of points spans
the line of no assortativity, with some values nearly as far below 0 as others are
above 0. However, the gender attributes do appear to be slightly assortative in
these social networks: although all values are within 6% in either direction of 0,
the mean assortativity is 0.02, which is slightly above 0. This suggests a slight
amount of homophily by gender (like links with like) in the way people friend
each other on Facebook, although the tendency is very weak. In some schools,
we see a slight tendency for heterophily (like links with dislike), as one might
expect if the networks reflected heteronormative dating relationships. "

Unfortunately, there is no apparent link between the Network with highest values for Dorm attribute and the others with highest values for Gender attribute.

Conclusion

To conclude on the assortativity of each vertex's attributes, we can say that every attribute is **at least** slightly assortative and some of them are more (in a descending order : Student status, Dorm, Degree, Major, Gender).

In most of the cases, we can make the following statement : two linked vertices are more likely to have similarities in their attributes.

Question 4: Link prediction

Tests

We just want to ensure that our implementation computes good values by comparing them with NetworkX methods.

Conclusion

WARNING : Normally it would be necessary to compute the predictions on a large number of graphs. Unfortunately, I couldn't due to my own device performance. Even if I tried to load on a single graph, because of its size, the computing time was too long and my PC would crash and reboot no matter what. It may be due to a wrong implementation of the predictors, but I didn't find the issue. I found out that I finally could compute with one of the smallest graphs (Caltech). In this sense, the following interpretation will be incomplete.

It is hard to have a relevant analysis since we couldn't compute our predictions on a large scale of graphs.
NB : We should have Adamic/Adar and Jaccar predictions outperforming Common Neighbors prediction.

Question 5: Find missing labels with the label propagation algorithm

It appears that the accuracy of the label propagation algorithm is higher for the "dorm" and the "gender" attributes compared to the "major" attribute, across all fractions. This means that the algorithm is better at recovering missing values for "dorm" and "gender" attributes than for the "major" attribute. It is noticeable that these results fit with the table given in the project statement for Duke university.

The reason for this difference in accuracy between attributes could be due to a variety of factors, such as the density of connections between nodes for each attribute.

More specifically, knowing our assortativity results, the gap between "dorm" and "major" accuracy may be explained by the fact that the assortativity of the "dorm" attribute is much more significant than the "major" attribute. Hence, the "dorm" label is more likely to be spread by the LPA over the node connections.

If we deal with the case of the "gender" attribute which has the highest LPA accuracy, we can't really explain that assortativity as its coefficient is not that important. In this sense, this high value may be due to the number of possible "gender" values : 2 in this case. So we approach a value of 1/2 accuracy, as a purely random algorithm should also achieve. Nevertheless, we can notice that the "gender" accuracy value is below 0.5. This is consistent with its non significant assortativity value. In other words, if that attribute would be more assortative, the LPA accuracy should be higher than 0.5 since the label spread would be easier.

Question 6: Communities detection with the Facebook100 Dataset

Nothing asked, cf code