Network Analysis Project

NGUYEN Hugo - M2 TRIED

January 2025

Repository github of the code (jupyter Notebook) : `https://github.com/Hugo-NGN/Network-Analysis`

Introduction

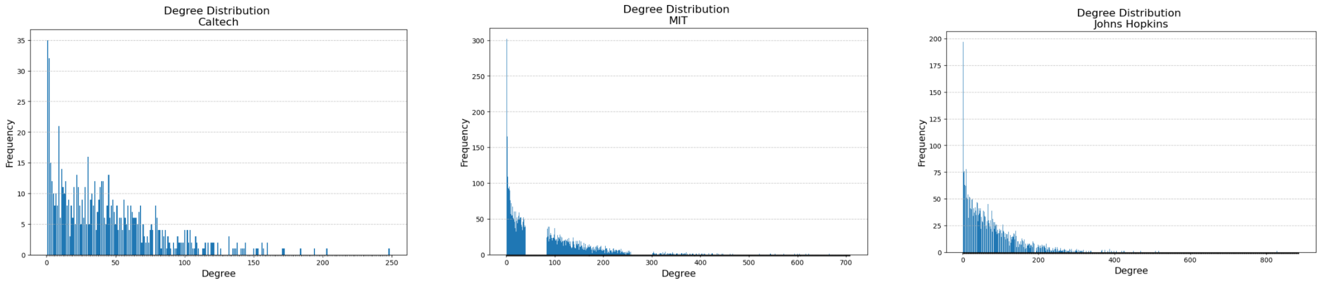
This homework focuses on the analysis of social networks using the $Facebook100$ dataset, which contains friendship connections from 100 US universities in fall 2005. The dataset provides a snapshot of the early stages of Facebook, showing insights of the structure and dynamics of social networks during that period. We begin by reading and understanding key documents that provide background information on the dataset and network analysis techniques. Following this, we perform social network analysis on three specific networks : Caltech, MIT, and Johns Hopkins. We then investigate assortativity patterns for various vertex attributes, such as student/faculty status, major, degree, dorm, and gender, to understand the mixing patterns within the networks. Additionally, we implement and evaluate link prediction metrics to assess their performance in predicting missing edges. We also address the problem of missing labels using label propagation algorithms, evaluating their accuracy in recovering missing attributes. Finally, we formulate a research question about group formation among students and use community detection algorithms to validate our hypothesis.

Question 1

Reading...

Question 2

Question 2.a



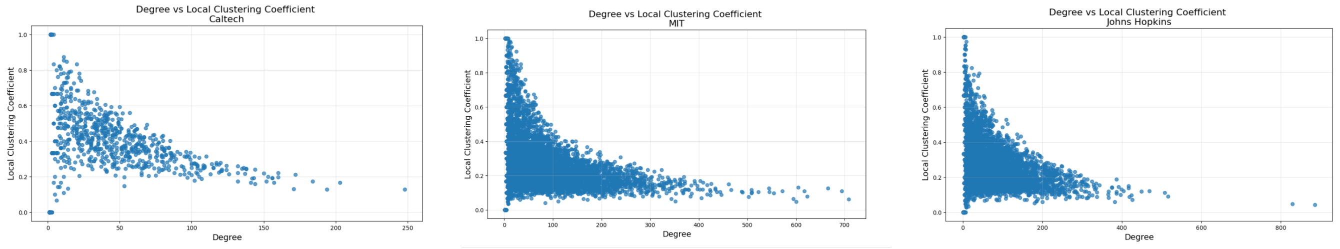
((a)) Caltech ((b)) MIT ((c)) John Hopkins

FIGURE 1 – Degree of distribution

All three distributions exhibit a right-skewed shape, this indicates that most nodes have a low degree, while a few nodes have a very high degree. Which means that most people have few connections and few people have a lot of connections.

Moreover, MIT and Johns Hopkins have a wider degree range compared to Caltech (700 and 900 compared to 250 respectively). This suggests that MIT and Johns Hopkins have nodes with higher degrees, indicating the presence of more highly connected individuals.

Question 2.b and 2.c



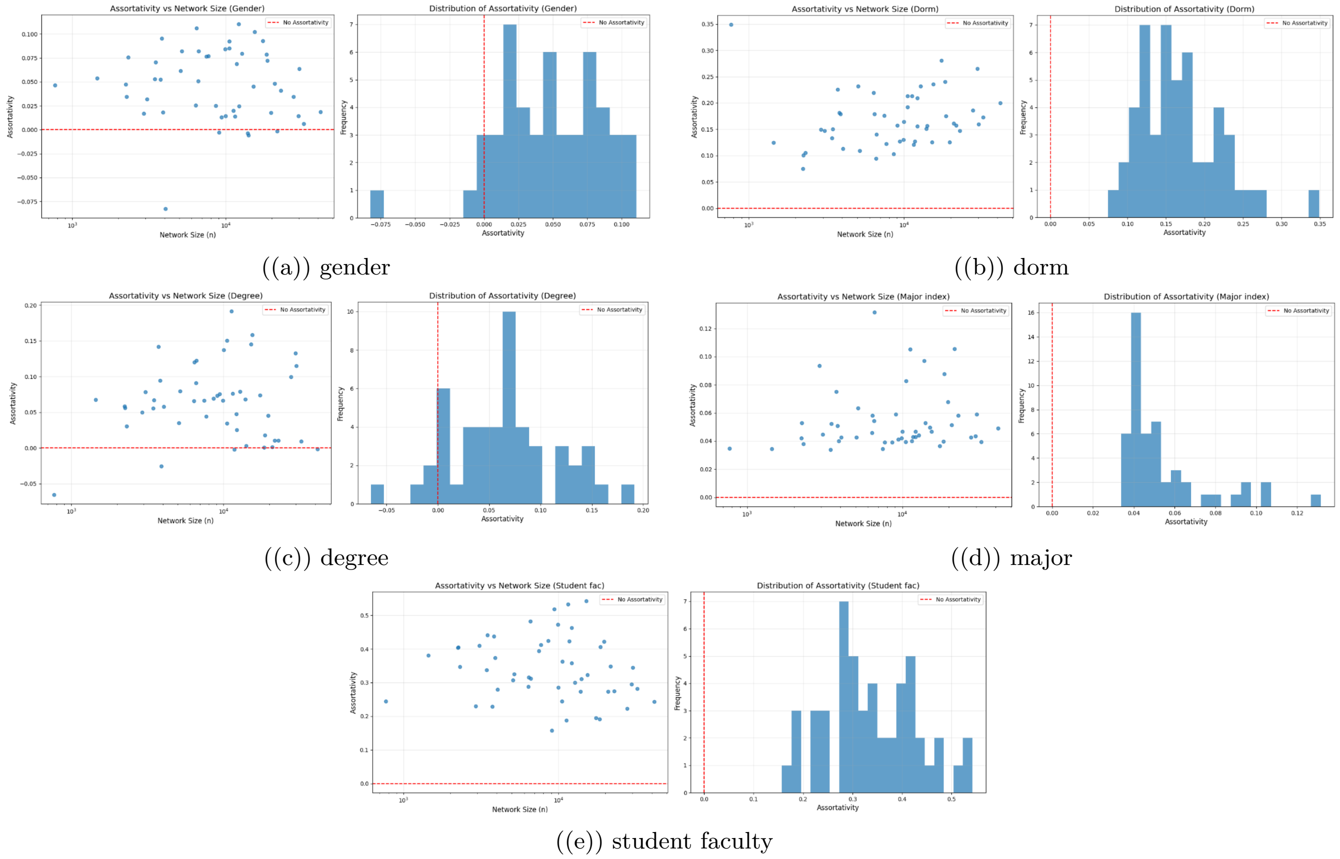
((a)) Caltech ((b)) MIT ((c)) John Hopkins

FIGURE 2 – Degree of distribution vs clustering coefficient

Given the high frequency of low-degree nodes and the relatively low edge density (as inferred from the degree distributions), all three networks can be considered sparse. Which means that the number of edges is much smaller compared to the number of possible edges.

In all three networks, the local clustering coefficient decreases as the degree increases. This indicates that high-degree nodes (hubs) are less likely to have interconnected neighbors.

Question 3



((a)) gender ((b)) dorm

((c)) degree ((d)) major

((e)) student faculty

FIGURE 3 – Assortativity as function of the networks size (scatter plots) and assortativity distribution (histograms)

Globally, we can see in figure 3 that the assortativity patterns in the Facebook100 dataset reveal that vertices exhibit varying degrees of assortative mixing based on different attributes. Physical proximity (dorm), shared academic interests (major), role within the academic community (student/faculty status), and similar levels of connectivity (degree) contribute to higher assortativity values. These findings highlight the multifaceted nature of friendship formation in social networks, influenced by various social, academic, and physical factors.

Question 4

(cf. results on question 4 in the jupyter notebook)

Common Neighbors maintains high precision but suffers from low recall, making it less effective. Jaccard Coefficient shows improved recall compared to Common Neighbors, particularly in the Caltech network, but still struggles with low recall in the other networks. Adamic/Adar demonstrates the best performance, with high precision and significantly better recall across all three networks. Therefore, Adamic/Adar is the most effective metric for link prediction in these cases.

Question 5

Based on the results obtained from the Label Propagation Algorithm (LPA) (cf Jupyter notebook) for recovering missing attributes in the $Facebook100$ dataset, we can deduce several conclusions about the efficiency of the metrics : Common Neighbors, Jaccard Coefficient, and Adamic/Adar.

The accuracy and Mean Absolute Error (MAE) metrics for the attributes "dorm", "major_index", and "gender" across different fractions of removed attributes (10%, 20%, and 30%) provide insights into how well the LPA performs in recovering these attributes. For the "dorm" attribute, the accuracy ranges from approximately 0.41 to 0.64, with MAE values indicating varying levels of error. For the "major_index" attribute, the accuracy is generally lower, ranging from about 0.16 to 0.27, with corresponding MAE values reflecting higher errors. For the "gender" attribute, the accuracy is relatively higher, ranging from about 0.52 to 0.60, with lower MAE values, indicating better performance in recovering this attribute.

These results suggest that the LPA is more effective in recovering certain attributes over others. The "gender" attribute shows the highest accuracy and lowest MAE, indicating that it is the easiest to recover using LPA. This could be due to the more homogeneous distribution of gender labels within the network, so it is easier for the algorithm to propagate the correct labels. On the other hand, the "major_index" attribute shows the lowest accuracy and highest MAE, which suggest that it is the most challenging to recover. This could be attributed to the higher diversity and less homogeneous distribution of major labels within the network.

Overall, the efficiency of the LPA in recovering missing attributes varies significantly depending on the attribute type. The "gender" attribute is recovered with the highest accuracy and lowest error, while the "major_index" attribute is the most challenging to recover. These findings highlight the importance of considering the nature of the attribute when applying label propagation algorithms for missing data imputation in social networks.

Question 6

In this section, we aim to investigate the group formation among students within the FB100 dataset. Especially, we suppose that students tend to form communities based on shared attributes such as dormitory residence, major, or year of graduation.

Our research question is : Do students in the same academic major tend to form tighter communities within the Facebook100 dataset ?

To validate our hypothesis, we will use the Louvain method for community detections. We will apply this algorithm to the friendship networks of Caltech. The Louvain method maximize modularity, a measure of the strength of division of a network into communities.

The community detection results for the selected university are visualized in the following figure 4. Each color represents a different community detected by the Louvain method. The visualization itself do not distinctly reveal the various community clusters. However, upon examining the attributes within individual communities (example in figure 5), it becomes evident that a significant proportion of members within each group share the same attribute value (such as "major" in figure 5). This observation suggests that individuals with shared attributes tend to form closer connections compared to those who do not share these attributes.

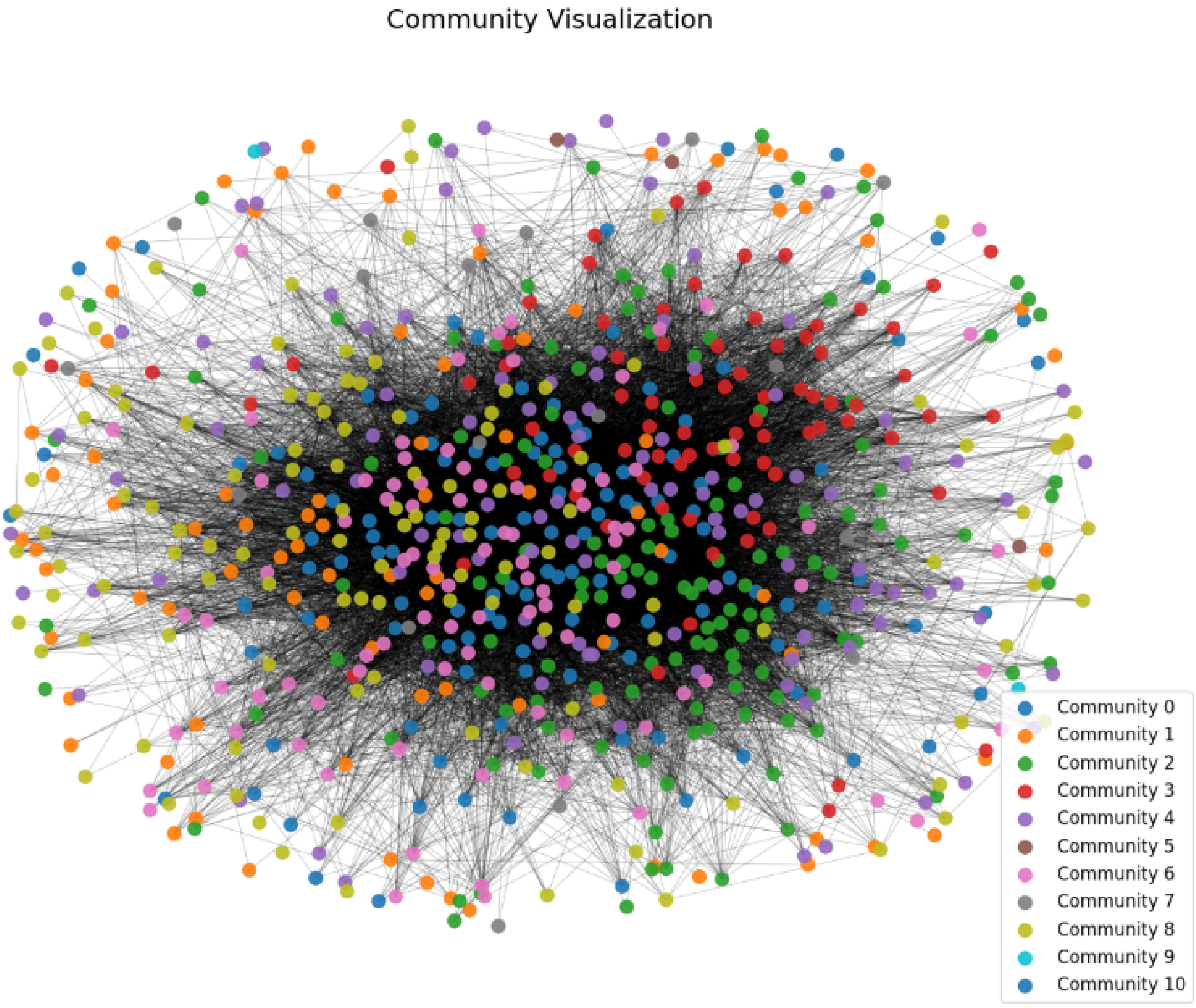
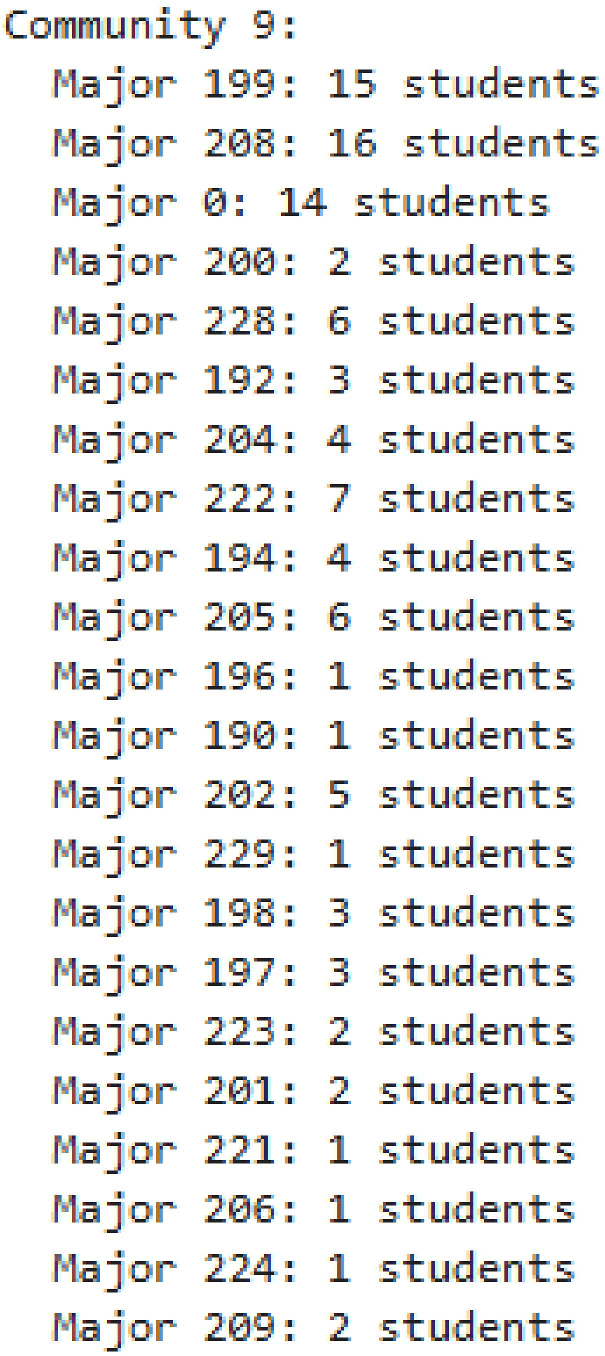


Figure 4 – Community visualization

Example of a community composition :



```
Community 9:
  Major 199: 15 students
  Major 208: 16 students
  Major 0: 14 students
  Major 200: 2 students
  Major 228: 6 students
  Major 192: 3 students
  Major 204: 4 students
  Major 222: 7 students
  Major 194: 4 students
  Major 205: 6 students
  Major 196: 1 students
  Major 190: 1 students
  Major 202: 5 students
  Major 229: 1 students
  Major 198: 3 students
  Major 197: 3 students
  Major 223: 2 students
  Major 201: 2 students
  Major 221: 1 students
  Major 206: 1 students
  Major 224: 1 students
  Major 209: 2 students
```

Figure 5 – Example of community composition