Final Project Report

Guillaume Macquart de Terline

Network Science and Graph Learning

01 Janvier 2026

Contents

Abstract 3

Introduction 3

Context . . . 3

Considered Dataset . . . 3

Question 1: Reading 3

Question 2: Social Network Analysis with the Facebook100 Dataset 4

Question 3: Assortativity Analysis with the Facebook100 Dataset 5

Student/Faculty Status . . . 6

Dorm . . . 6

Degree . . . 7

Major . . . 7

Gender . . . 8

Question 4: Link prediction 8

Question 5: Find missing labels with the label propagation algorithms 10

Question 6. Communities detection with the FB100 datasets 11

Conclusion 12

Abstract

This report includes the answers to my homework for the final project of the Network Science and Graph Learning course. The code used to perform the analysis is available in the GitHub repository:

https://github.com/Gdeterline/Network-Science-and-Graph-Learning.

Introduction

Context

Network Science has emerged as a fundamental discipline for understanding the structure and dynamics of complex systems. From biological interactions and technological infrastructures to social relationships, graph representations allow us to model entities as nodes and their interactions as edges, revealing patterns that are not observable when analyzing individual components in isolation.

In particular, the study of online social networks such as Facebook has gained significant traction. These networks provide a rich testing ground for algorithms in Graph Learning, such as link prediction, label propagation, and community detection. Understanding the topology of social graphs helps in analyzing information diffusion, influence maximization, and the formation of social bubbles.

This project focuses on the analysis of real-world social network data using utilizing various metrics and algorithms from the field of Graph Theory and Machine Learning. By analyzing the Facebook 100 dataset, and applying various algorithms, we aim to uncover the underlying organizational principles of these university-based social communities.

Considered Dataset

The dataset analyzed in this project is the well-known **Facebook 100** dataset. This collection consists of complete friendship networks from 100 American colleges and universities as they existed in September 2005. The anonymized snapshot contains the connections among $n = 1,208,316$ users, and a total of $m = 93,969,074$ edges (unweighted and undirected) between them.

The data consists of 100 graph files in GML (Graph Modeling Language) format, located in the `data/` directory. This includes networks ranging from colleges (e.g., *Caltech* with 762 nodes) to larger universities (e.g., *UCLA* with 20466 nodes, etc.).

Each graph represents a single university's social network:

- **Nodes**: Represent individual persons (students, faculty, or staff).
- **Edges**: Represent "friendship" links between two individuals.
- **Attributes**: Nodes are typically annotated with metadata, including:
  - Person's Status (Undergraduate, Graduate, Summer Student, Faculty, Staff or Alumni)
  - Dormitory/House (if any)
  - Major (if any)
  - Gender (M or F)
  - Graduation Year

Question 1: Reading

The purpose of this question is to read and understand the following three papers:

- *Social structure of Facebook networks* by Traud, A. L., Mucha, P. J. & Porter, M. A. (Physica A: Statistical Mechanics and its Applications 391, 4165 – 4180, 2012). The main goal of this paper is to study the social structure of Facebook networks at 100 American colleges and universities at a single point in time. The authors investigate how user attributes such as gender, major, dormitory, and graduation year influence the formation of friendship links, specifically focusing on assortativity and community structure.
- *Comparing community structure to characteristics in online collegiate social networks* by Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. (SIAM Review 53, 526–543, 2011). This paper explores the community structure within the Facebook networks of five US universities. The study quantifies the correlation between topological communities and metadata, highlighting how different demographic factors drive clustering at different institutions.

- *Assembling thefacebook: Using heterogeneity to understand online social network assembly* by Jacobs, A. Z., Way, S. F., Ugander, J. & Clauset, A. (Proceedings of the ACM Web Science Conference, WebSci '15, 18:1–18:10, 2015). This paper examines the evolution and assembly of the early Facebook network by analyzing the temporal sequence of user sign-ups and link formations. The authors propose a model that leverages node heterogeneity to explain the network's growth patterns and structural properties.

Question 2: Social Network Analysis with the Facebook100 Dataset

Let us consider three networks from the FB100 dataset: `Caltech` (with 762 nodes in the LCC), `MIT` (which has 6402 nodes in the LCC), and `Johns Hopkins` (which has 5157 nodes in the LCC).

The figures 1, 2, and 3 show the degree distributions for each of these three networks on a log-log scale.

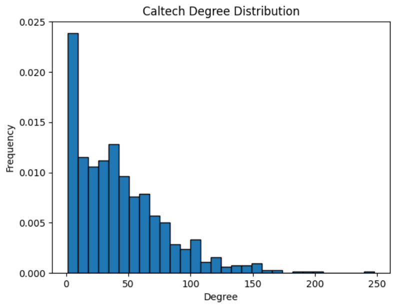

Figure 1: Degree distribution for Caltech network

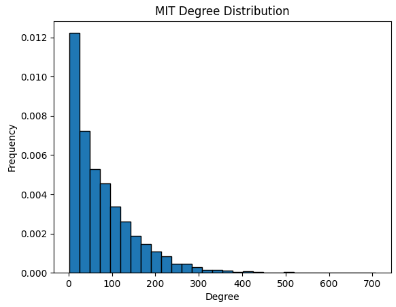

Figure 2: Degree distribution for MIT network

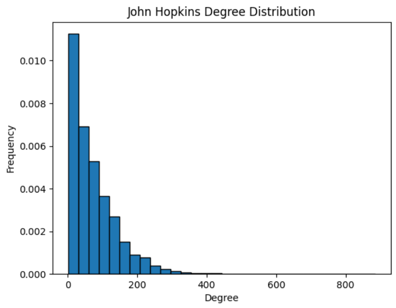

Figure 3: Degree distribution for Johns Hopkins network

The figures 2, and 3 both exhibit quite steep normal distributions, exhibiting that most nodes are of small degree, while a few others are of higher degree (hubs). This is typical of social networks, where most individuals have a limited number of connections, while a few individuals (hubs) have a significantly larger number of connections, facilitating information flow and connectivity within the network. The phenomenon is similar in big universities, where social circles tend to be smaller and more localized. In contrast, the Caltech network displays a noticeably different structure (figure 1). As the smallest network in our sample (762 nodes), it is significantly denser than the others, with a bigger proportion of nodes of higher degrees. This high density suggests that the social environment at Caltech is more cohesive, with students likely knowing a larger fraction of their peers compared to the two larger institutions.

Given these three networks, we now seek to analyze their graph topologies. Thus, we seek to compute the global clustering coefficient and mean local clustering coefficient for each of the 3 networks. In addition, we compute the edge density of each network. The results are summarized in Table 1.

College/University	Global Clustering Coefficient	Mean Local Clustering Coefficient	Edge Density
Caltech	0.291283	0.409294	0.056404
MIT	0.180288	0.271219	0.012118
John Hopkins	0.193161	0.268393	0.013910

Table 1: Clustering Coefficients (Global and Mean Local) and Edge Density for Caltech, MIT, and Johns Hopkins networks

First of all, all three networks should be construed as **sparse** when inspecting the table 1. MIT and Johns Hopkins exhibit very low edge densities (∼1.2% and ∼1.4%), indicating that the vast majority of possible connections are absent. Caltech, while significantly denser (∼5.6%) likely due to its smaller student body and cohesive housing system, still qualifies as sparse in graph-theoretical terms. Then, regarding the clustering coefficients, Caltech stands out with the highest global clustering coefficient (0.291283) and mean local clustering coefficient (0.409294). This suggests that students at Caltech tend to form tightly-knit groups/clusters. In contrast, MIT and Johns Hopkins have lower clustering coefficients, indicating a less pronounced tendency for nodes (and thus students) to cluster together. The graph structures of these larger universities may be more fragmented, with students forming smaller, less interconnected social circles.

To assess the similarities and differences in local structures across these networks, we can scatter plot the degree versus the local clustering coefficient for each network node. The resulting plots are shown in Figures 4, 5, and 6.

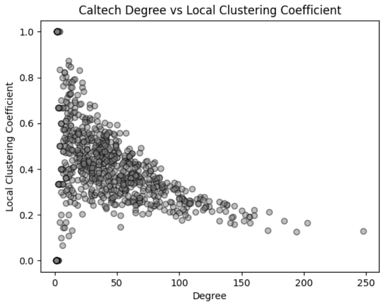

Figure 4: Degree vs Local Clustering Coefficient for Caltech network

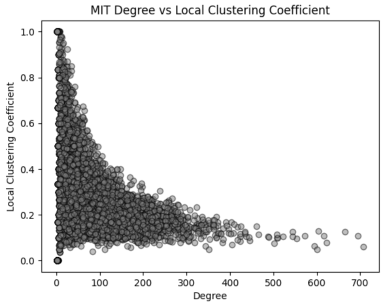

Figure 5: Degree vs Local Clustering Coefficient for MIT network

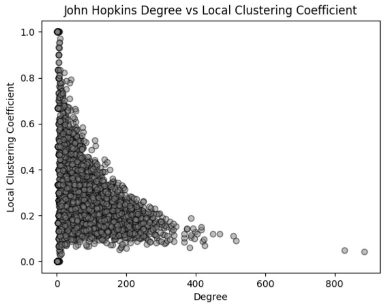

Figure 6: Degree vs Local Clustering Coefficient for Johns Hopkins network

The scatter plots of Degree vs. Local Clustering Coefficient show a consistent pattern across all three networks. In each case, nodes with low degrees tend to have high clustering coefficients, forming tight local groups. In contrast, high-degree nodes ("hubs") exhibit low clustering coefficients. This specific "L-shaped" pattern in the scatter plots may indicate that highly connected individuals serve as bridges between different social groups that are not themselves connected, thereby reducing the local clustering around these hubs.

While this pattern is common across all three networks, **Caltech** still stands out. The "cloud" of points for Caltech seems shifted upwards compared to MIT and Johns Hopkins. Even nodes with moderate-to-high degrees at Caltech maintain a relatively higher clustering coefficient than their counterparts in the larger universities. This observation aligns well with the earlier findings regarding Caltech's higher overall clustering coefficients, confirming that its social structure is not only smaller but also more cohesive, where even "popular students", that would only act as bridges in larger networks, belong to highly interconnected internal groups.

Question 3: Assortativity Analysis with the Facebook100 Dataset

In this section, we compute, then analyze the assortativity patterns of all FB100 networks with respect to 5 different vertex attributes: Student/Faculty Status, Major, Vertex Degree, Dormitory and Gender.

Note: To perform this analysis, we considered the networks as simple graphs (i.e., no self-loops or multiple edges).

The summary statistics for the assortativity attributes across all 100 universities are computed for these five key attributes. The table 2 below summarizes the mean, standard deviation, minimum, and maximum assortativity coefficients for each attribute, accross all FB100 networks.

Statistic	student_fac	major	degree	dorm	gender
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.322738	0.051115	0.062648	0.175106	0.042958
std	0.092508	0.017261	0.052845	0.057733	0.038569
min	0.110200	0.028574	-0.065273	0.074814	-0.082493
25%	0.256311	0.039919	0.025204	0.131742	0.018481
50%	0.316925	0.046796	0.064705	0.172658	0.046710
75%	0.395228	0.054907	0.092888	0.202718	0.073286
max	0.542625	0.131654	0.196892	0.416017	0.124723

Table 2: Summary Statistics for Assortativity Coefficients across FB100 Networks

We can now analyze the assortativity patterns for each attribute in detail. To further illustrate these patterns, we have, for each attribute, created two types of visualizations:

- A scatter plot showing the relationship between network size (number of nodes) and assortativity coefficient.
- A histogram displaying the distribution of assortativity coefficients across all FB100 networks.

Student/Faculty Status

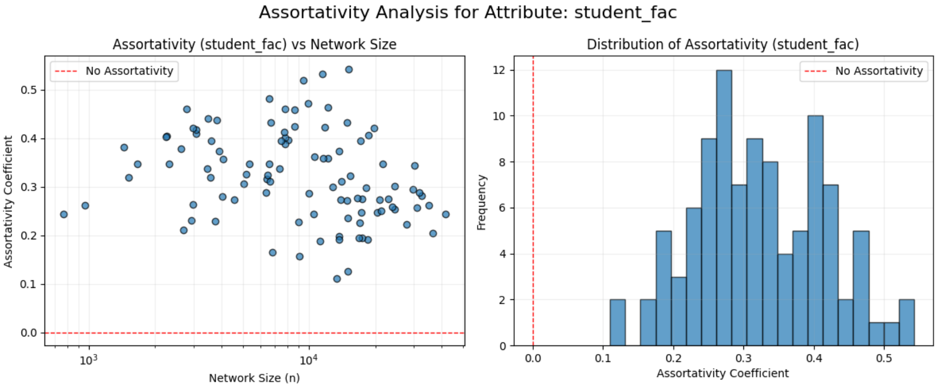

Figure 7: Assortativity by Student/Faculty Status vs Network Size and Assortativity Coefficient Distribution

This attribute exhibits the highest level of assortativity among those measured (mean $\mu \approx 0.32$, with a standard deviation of $\sigma \approx 0.09$). The second plot in figure 7 displays a histogram of the assortativity coefficients across all FB100 networks, showing that clear positive skew. Additionally, the scatter plot shows no significant correlation between network size and assortativity by status, but rather a consistent positive assortativity across all institutions. This indicates a strong tendency for individuals to associate with others of the same status (e.g., undergraduate with undergraduate, staff with staff, etc.). This is expected as social interactions in a university setting are heavily stratified by role; students share classes, lifestyle patterns, etc., while staff interact in distinct professional settings. This structural homophily is most certainly the dominant demographic driver of clustering in these networks.

Dorm

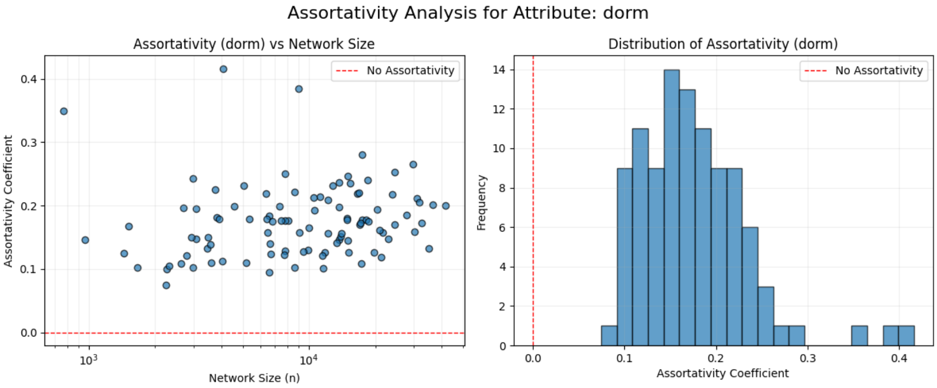

Figure 8: Assortativity by Dormitory vs Network Size and Assortativity Coefficient Distribution

The dormitory residence shows moderate positive assortativity (mean $\mu \approx 0.175$ with a standard deviation of $\sigma \approx 0.058$). The histogram in figure 8 once again shows a clear positive skew, indicating that most universities exhibit positive assortativity by dormitory. The scatter plot shows the same observation as before: no significant correlation between network size and assortativity by dormitory, but rather a consistent positive assortativity across all institutions. This reflects the significant impact of physical proximity on the formation of social ties within university settings: students living in the same housing unit are naturally more likely to interact daily and become friends. The value is significant but lower than status, suggesting that while housing is a strong catalyst for local community formation, students still form substantial ties outside their immediate living environment (e.g., in classes, clubs, or sports).

Degree

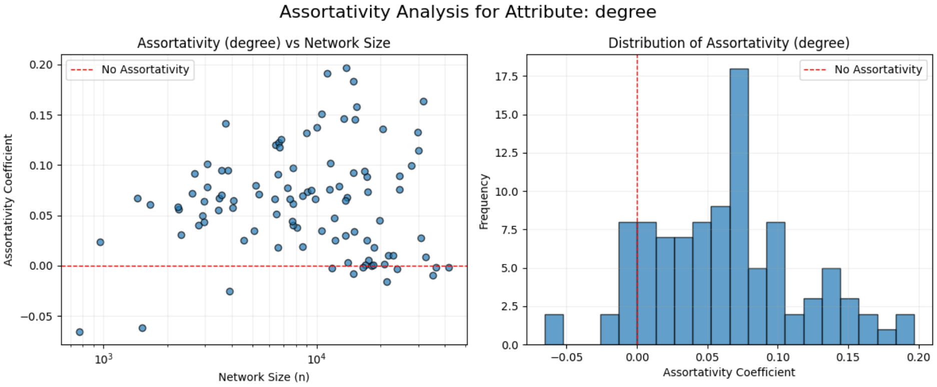

Figure 9: Assortativity by Degree vs Network Size and Assortativity Coefficient Distribution

The degree assortativity is positive but relatively low (mean $\mu \approx 0.06$ with a standard deviation of $\sigma \approx 0.053$). The histogram in figure 9 shows a distribution centered around low positive values, with some networks exhibiting slight disassortativity (negative values). The scatter plot again shows no clear correlation between network size and degree assortativity (or at least a very weak one, that would perhaps require more samples in the lower end to be confirmed). The figure shows that the assortativity coefficients are positive across most institutions, with a few exceptions. The absolute values remain low overall.

This might indicate a slight tendency for high-degree nodes (popular individuals) to connect with other high-degree nodes, and low-degree nodes with other low-degree nodes, though given the low assortativity values, this effect is weak, if not negligible. At least, it suggests that degree-based homophily is not a dominant factor in friendship formation within these collegiate networks.

Major

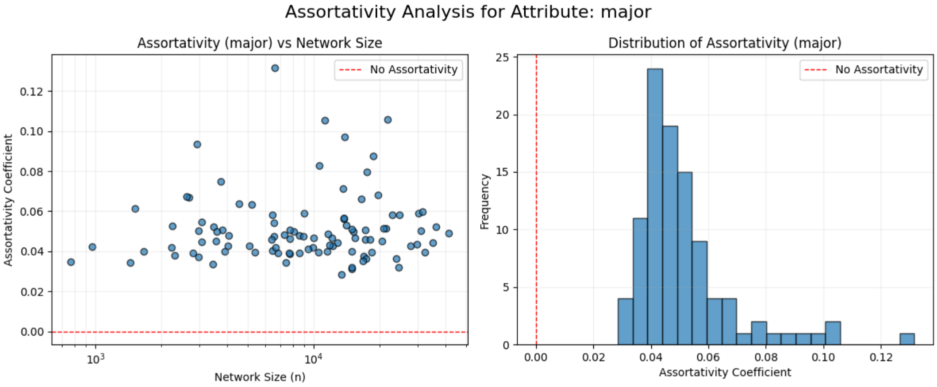

Figure 10: Assortativity by Academic Major vs Network Size and Assortativity Coefficient Distribution

The assortativity by academic major is surprisingly low (mean $\mu \approx 0.05$ with a standard deviation of $\sigma \approx 0.017$). The histogram in figure 10 shows a distribution tightly clustered around low positive values, indicating that most universities exhibit very weak assortativity by major. The scatter plot again shows no significant correlation between network size and assortativity by major. The scatter plot shows consistence, with all institutions exhibiting low but positive assortativity values. This suggests that while shared intellectual interests exist (the assortativity values are low but still positive), they do not dictate the majority of the social structure, with students forming friendships across different fields of study. This result might be the most surprising, as one might expect students to bond over shared majors (and thus classes, projects, etc.). However, it seems that other factors (status, dormitory) play a more significant role in shaping social connections.

Gender

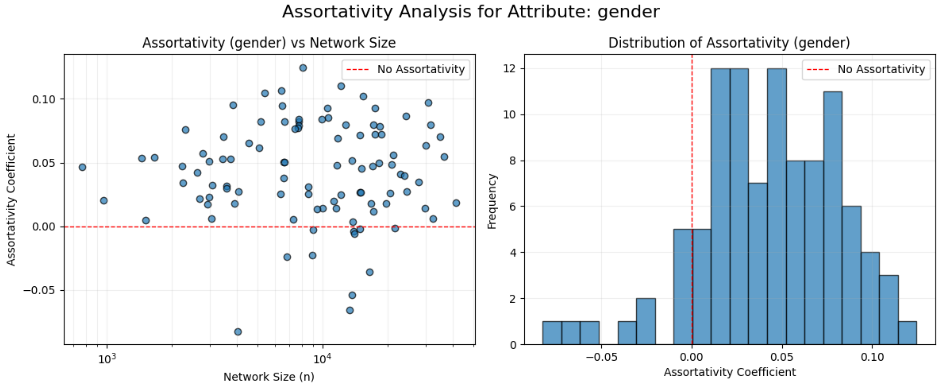

Figure 11: Assortativity by Gender vs Network Size and Assortativity Coefficient Distribution

Gender exhibits the lowest assortativity (mean $\mu \approx 0.04$ with a standard deviation of $\sigma \approx 0.039$). The histogram in figure 11 shows a distribution slightly skewed towards positive values, but with a significant number of networks exhibiting near-zero or even negative assortativity. If the scatter plot does not show any clear correlation between network size and gender assortativity, we can still observe that all negative values (still very low in absolute terms) are found in larger networks (4000 nodes and more), suggesting a slight trend. Most scatter points remain clustered around low positive values. This indicates a nearly neutral mixing in college/university networks, where gender is not a significant barrier or exclusive driver for friendship formation.

Overall, the assortativity analysis across the FB100 networks reveals that **status** and **dormitory** are the most influential attributes driving homophily in these social networks, while **major**, **degree**, and especially gender play much smaller roles. This suggests that social stratification by role and physical proximity are the primary factors shaping friendship patterns in collegiate environments.

Question 4: Link prediction

We will now perform link prediction on a subset of 10 graphs from the FB100 dataset. Due to computational constraints, we have selected the 10 smallest networks in terms of number of file size, to ensure that the link prediction algorithms can run efficiently, and on a significant number of networks. These are the following networks: `Caltech`, `Reed`, `Simmons`, `Haverford`, `Swarthmore`, `USFCA`, `Bowdoin`, `Mich`, `Amherst`, and `Oberlin`.

To proceed, we have first read the article *The link prediction problem for social networks* by Liben-Nowell, D. & Kleinberg, J. (Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03, 556–559, 2003). Without entering into too much detail, this article presents a comprehensive study of various link prediction algorithms applied to social networks. The authors evaluate several local and global similarity measures to predict future links based on the existing network structure.

Based on the insights from this article, we implemented three local link prediction metrics: Common Neighbors, Jaccard's Coefficient, and Adamic/Adar. The implementations are available in the `main.ipynb` notebook of the GitHub repository, and the implementation inherits from the `LinkPrediction` class defined in the `link_prediction.py` file.

First, we will compute the Precision@k and Recall@k for one given network (Caltech), accross different fractions of removed edges (5%, 10%, 15%, and 20%), and accross different values of k (50, 100, 200, 300, and 400). The results are summarized in the tables 3 and 4 below.

Fraction Removed	Link Predictor	Precision@50	Precision@100	Precision@200	Precision@400
	Common Neighbors	0.5000	0.4600	0.3850	0.2850
0.05	Jaccard	0.3000	0.2700	0.2250	0.1775
	Adamic/Adar	0.3800	0.3500	0.3200	0.2475
—	—	—	—	—	—
	Common Neighbors	0.7000	0.5900	0.5450	0.4525
0.1	Jaccard	0.4400	0.4300	0.3800	0.3600
	Adamic/Adar	0.5800	0.5700	0.5300	0.4125
—	—	—	—	—	—
	Common Neighbors	0.7600	0.7000	0.6800	0.5700
0.15	Jaccard	0.4000	0.5000	0.4100	0.4450

Fraction Removed	Link Predictor	Precision@50	Precision@100	Precision@200	Precision@400
	Adamic/Adar	0.7400	0.6900	0.6700	0.5600
—	—	—	—	—	—
0.2	Common Neighbors	0.6800	0.7600	0.7550	0.6175
	Jaccard	0.5200	0.5500	0.5100	0.4925
	Adamic/Adar	0.8000	0.8100	0.7500	0.6900

Table 3: Link Prediction Results for Caltech Network: Precision@k

Fraction Removed	Link Predictor	Recall@50	Recall@100	Recall@200	Recall@400
0.05	Common Neighbors	0.0300	0.0553	0.0925	0.1370
	Jaccard	0.0180	0.0325	0.0541	0.0853
	Adamic/Adar	0.0228	0.0421	0.0769	0.1190
—	—	—	—	—	—
0.1	Common Neighbors	0.0210	0.0354	0.0655	0.1087
	Jaccard	0.0132	0.0258	0.0456	0.0865
	Adamic/Adar	0.0174	0.0342	0.0637	0.0991
—	—	—	—	—	—
0.15	Common Neighbors	0.0152	0.0280	0.0544	0.0913
	Jaccard	0.0080	0.0200	0.0328	0.0713
	Adamic/Adar	0.0148	0.0276	0.0536	0.0897
—	—	—	—	—	—
0.2	Common Neighbors	0.0102	0.0228	0.0453	0.0742
	Jaccard	0.0078	0.0165	0.0306	0.0591
	Adamic/Adar	0.0120	0.0243	0.0450	0.0829

Table 4: Link Prediction Results for Caltech Network: Recall@k

The results for the Caltech network indicate that the Common Neighbors and Adamic/Adar predictors generally outperform Jaccard's Coefficient in terms of Precision@k across all fractions of removed edges. This suggests that these two predictors are more effective at identifying true positive links among the top k predictions. Additionally, as the fraction of removed edges increases, the Precision@k values tend to improve for all predictors, for a given k. But as the value of k increases, the Precision@k values tend to decrease for all predictors, for a given fraction of removed edges. This is expected, as taking higher values of k means taking more predictions into account, which increases the likelihood of including false positives. On the opposite, the Recall@k values tend to decrease as the fraction of removed edges increases, for a given k. This is also expected, as removing more edges makes it harder for the predictors to recover the missing links. However, as the value of k increases, the Recall@k values tend to increase for all predictors, for a given fraction of removed edges. This is also expected, as taking higher values of k means considering more predictions, which increases the likelihood of capturing true positives.

Then, we compute the average Precision@k and Recall@k for the 10 chosen networks, accross different fractions of removed edges, and accross different values of k. The results are summarized in the tables 5 and 6 below.

Fraction Removed	Link Predictor	Precision@50	Precision@100	Precision@200	Precision@300	Precision@400
0.05	Common Neighbors	0.4620	0.4020	0.3245	0.2873	0.2627
	Jaccard	0.3360	0.3550	0.3290	0.3077	0.2907
	Adamic/Adar	0.4460	0.3950	0.3230	0.2967	0.2682
—	—	—	—	—	—	—
0.1	Common Neighbors	0.6320	0.5880	0.5100	0.4710	0.4460
	Jaccard	0.4560	0.4690	0.4730	0.4537	0.4450
	Adamic/Adar	0.6700	0.6160	0.5415	0.4990	0.4635
—	—	—	—	—	—	—
0.15	Common Neighbors	0.7520	0.7120	0.6470	0.5967	0.5582
	Jaccard	0.5080	0.5310	0.5490	0.5393	0.5353
	Adamic/Adar	0.7440	0.7140	0.6565	0.6137	0.5795
—	—	—	—	—	—	—
0.2	Common Neighbors	0.8200	0.7650	0.6985	0.6627	0.6283
	Jaccard	0.5400	0.5540	0.5745	0.5760	0.5755

Fraction Removed	Link Predictor	Precision@50	Precision@100	Precision@200	Precision@300	Precision@400
	Adamic/Adar	0.7740	0.7580	0.7220	0.6843	0.6550

Table 5: Link Prediction Results: Average Precision@k for the Different Link Predictors and Fractions of Removed Edges

Fraction Removed	Link Predictor	Recall@50	Recall@100	Recall@200	Recall@300	Recall@400
0.05	Common Neighbors	0.0105	0.0184	0.0295	0.0384	0.0464
	Jaccard	0.0067	0.0148	0.0271	0.0383	0.0482
	Adamic/Adar	0.0097	0.0175	0.0279	0.0382	0.0461
—	—	—	—	—	—	—
0.1	Common Neighbors	0.0072	0.0135	0.0229	0.0316	0.0399
	Jaccard	0.0048	0.0099	0.0201	0.0292	0.0376
	Adamic/Adar	0.0076	0.0139	0.0241	0.0331	0.0411
—	—	—	—	—	—	—
0.15	Common Neighbors	0.0058	0.0110	0.0195	0.0268	0.0330
	Jaccard	0.0035	0.0076	0.0156	0.0233	0.0307
	Adamic/Adar	0.0056	0.0109	0.0200	0.0277	0.0350
—	—	—	—	—	—	—
0.2	Common Neighbors	0.0048	0.0089	0.0160	0.0225	0.0284
	Jaccard	0.0031	0.0064	0.0126	0.0191	0.0254
	Adamic/Adar	0.0049	0.0091	0.0165	0.0232	0.0295

Table 6: Link Prediction Results: Average Recall@k for the Different Link Predictors and Fractions of Removed Edges

The overall trends observed in the average results across the 10 networks are consistent with those seen in the Caltech network specifically. Common Neighbors and Adamic/Adar generally outperform Jaccard's Coefficient in terms of Precision@k across all fractions of removed edges. As the fraction of removed edges increases, the Precision@k values tend to improve for all predictors, while the Recall@k values tend to decrease. Additionally, as the value of k increases, the Precision@k values tend to decrease, while the Recall@k values tend to increase for all predictors. These consistent patterns reinforce the effectiveness of Common Neighbors and Adamic/Adar as link prediction methods in social networks, particularly in the context of the FB100 dataset. Should we have to choose one predictor among the three, Adamic/Adar would probably be the best choice, as it often outperforms Common Neighbors by a small (to negligible) margin, while being conceptually more robust. Nevertheless, the choice of predictor may still depend on the specific characteristics of the network being analyzed, as well as the computational resources available: Common Neighbors is computationally less intensive than Adamic/Adar, which may be a consideration for very large networks, or when computational efficiency is a constraint/priority.

Question 5: Find missing labels with the label propagation algorithms

In this section, we will implement and evaluate a label prediction algorithm, on the same subset of 10 graphs from the FB100 dataset as in Question 4.

To proceed, we first started by reading the article *Node Classification in Social Networks* by Bhagat, S., Cormode, G. & Muthukrishnan, S. (In Social Network Data Analytics, 115–148, Springer US, 2011). The article provides a comprehensive overview of various node classification techniques in social networks, including label propagation algorithms. The authors consider methods based on iterative application of traditional classifiers using graph information as features, and methods which propagate the existing labels via random walks. Based on the insights from this article, and on the class material, we implemented a semi-supervised label propagation algorithm, that leverages the graph structure to infer missing labels based on the known labels of neighboring nodes. The implementation is available in the `src/LabelPropagation.py` file of the GitHub repository.

We evaluated the performance of our label propagation algorithm using three metrics: Accuracy, F1-Score. We first considered a single network (MIT), and evaluated the performance of the label propagation algorithm accross different fractions of removed labels (10%, 20%, 30%) and accross different node attributes ("dorm", "major", and "gender"). The results are summarized in the table 7 below.

attribute	fraction_removed	accuracy	f1_score
	0.1	0.545	0.453
dorm	0.2	0.515	0.428
	0.3	0.404	0.265
—	—	—	—
	0.1	0.247	0.068
major	0.2	0.235	0.070
	0.3	0.236	0.054
—	—	—	—
	0.1	0.627	0.441
gender	0.2	0.625	0.447
	0.3	0.626	0.437

Table 7: Label Propagation Results for MIT Network

The results for the MIT network indicate that the label propagation algorithm performs best when predicting the "gender" attribute, achieving the highest accuracy and F1-score across all fractions of removed labels. The "dorm" attribute also shows moderate performance, while the "major" attribute yields the lowest accuracy and F1-score. This suggests that certain attributes are perhaps better suited for label propagation based on the underlying social structure of the network. Additionally, as the fraction of removed labels increases, both accuracy and F1-score tend to decrease for all attributes, which is expected as more missing labels make the prediction task more challenging.

Then, we computed the average accuracy and F1-score accross the 10 chosen networks, accross the same fractions of removed labels, and accross the same node attributes. The results are summarized in the table 8 below.

attribute	fraction_removed	average_accuracy	average_f1_score
	0.1	0.563	0.257
dorm	0.2	0.552	0.232
	0.3	0.533	0.22
—	—	—	—
	0.1	0.273	0.092
major	0.2	0.229	0.053
	0.3	0.204	0.033
—	—	—	—
	0.1	0.612	0.357
gender	0.2	0.623	0.346
	0.3	0.607	0.32

Table 8: Label Propagation Results: Average Accuracy and F1-Score for the Different Attributes and Fractions of Removed Labels

The overall trends observed in the average results across the 10 networks are consistent with those seen in the MIT network specifically. The label propagation algorithm performs best when predicting the "gender" attribute, achieving the highest average accuracy and F1-score across all fractions of removed labels. The "dorm" attribute also shows moderate performance, while the "major" attribute delivers the lowest average accuracy and F1-score. These patterns reinforce the effectiveness of label propagation for certain attributes in social networks, particularly in the context of the FB100 dataset.

These considerations are perhaps correlated to the assortativity analysis performed in Question 3: attributes with higher assortativity (e.g., gender, dorm) tend to yield better label propagation performance, as nodes with similar attributes are more likely to be connected, facilitating the spread of labels through the network. Conversely, attributes with lower assortativity (e.g., major) may not benefit as much from the network structure, leading to poorer prediction performance.

Question 6. Communities detection with the FB100 datasets

In this section, we seek to detect communities in the FB100 networks. Indeed, here we ask ourselves whether there are meaningful communities based on the topology of the graphs in these social networks, and if so, how they relate to the known attributes of the nodes (students), such as dormitory, or faculty status.

To proceed, we implemented two community detection algorithms:

- The Greedy Modularity method, which is a hierarchical agglomerative algorithm that builds communities by iteratively merging pairs of communities that result in the largest increase in modularity.
- The Louvain method for community detection, which is a popular algorithm for identifying communities in large networks by optimizing modularity (measure of the density of links inside communities compared to links between communities).

The implementation is available in the `src/CommunityDetection.py` file of the GitHub repository.

Here, we make the hypothesis that communities detected via the Louvain method will correspond to meaningful social groupings within the university networks, such as students living in the same dormitory or belonging to the same faculty status (undergraduate, graduate, staff, etc.). Should this hypothesis hold true, it would suggest that the structural properties of the networks reflect real-world social dynamics. On the contrary, if no such correspondence is found, it may indicate that other factors (not captured by the network structure) play a more significant role in shaping social groupings.

To evaluate this hypothesis, we applied both community detection algorithms to two graphs: the Caltech and Johns Hopkins networks. We chose these two networks due to their contrasting sizes and structures, which may yield different insights into community formation.

For each detected community, we computed the NMI (Normalized Mutual Information) and ARI (Adjusted Rand Index) scores between the detected communities and the known attributes of the nodes (dormitory and faculty status). The results are summarized in the table 9 below.

Network	Attribute	Algorithm	NMI Score	ARI Score
Caltech	dorm	Louvain	0.699	0.703
Caltech	dorm	Greedy Modularity	0.412	0.272
Caltech	faculty_status	Louvain	0.046	0.007
Caltech	faculty_status	Greedy Modularity	0.062	0.044
Johns Hopkins	dorm	Louvain	0.260	0.087
Johns Hopkins	dorm	Greedy Modularity	0.104	0.015
Johns Hopkins	faculty_status	Louvain	0.110	0.039
Johns Hopkins	faculty_status	Greedy Modularity	0.063	0.036

Table 9: Community Detection Results: NMI and ARI Scores for Caltech and Johns Hopkins Networks

First of all, the results seem to indicate that the Louvain method generally outperforms the Greedy Modularity method in terms of both NMI and ARI scores across all attributes and networks, though the margin of improvement varies. For instance, in the Caltech network, the Louvain method achieves significantly higher NMI and ARI scores for the "dorm" attribute compared to the Greedy Modularity method, while it is much less pronounced for the "faculty status" attribute. This suggests that the Louvain method may be more effective at capturing community structures that align with certain attributes. Secondly, the "dorm" attribute consistently returns higher NMI and ARI scores compared to the "faculty status" attribute across both networks and algorithms. This indicates that communities detected based on network topology are more closely related to dormitory residence than to faculty status. This finding aligns with the earlier assortativity analysis, which showed higher assortativity for dormitory compared to faculty status. Additionnally, regarding the dorm attribute, the Caltech network exhibits notably higher NMI and ARI scores compared to the Johns Hopkins network, particularly with the Louvain method. This suggests that the community structures in the Caltech network are more strongly aligned with dormitory residence than in the Johns Hopkins network. We had previously observed that Caltech had a more cohesive social structure, which may contribute (or at least be correlated) to this stronger alignment between detected communities and dormitory residence.

Conclusion

In this project, we had the opportunity to explore various aspects of social network analysis using the FB100 dataset. We began by examining fundamental network properties, such as degree distributions and clustering coefficients, revealing insights into the structural characteristics of university social networks. Our analysis highlighted the unique social dynamics at Caltech compared to larger institutions like MIT and Johns Hopkins. We then dove deeper into assortativity patterns, uncovering how attributes like status and dormitory residence significantly influence social connections. Our link prediction experiments demonstrated the effectiveness of local similarity measures, particularly Common Neighbors and Adamic/Adar, in recovering missing links. Finally, we explored label propagation techniques for inferring missing node attributes and employed community detection algorithms to identify meaningful social groupings within the networks. Overall, this project was the opportunity for me to apply a range of network analysis techniques discussed in class and gain deeper knowledge in that matter.