Final Project

Baozhuo Su

Introduction

This project aims at analyzing social networks by modeling them as connected graphs. We will firstly exploit 3 datasets, Caltech36, MIT8, JohonsHopkins55 in terms of degree distribution which will help the following anylysis. We perform transitivity analysis to exhibit the sparsity of the dataset, results are confirmed to be consistent by the following degree-local clustering coefficient. Assortativity is computed and analyzed for each attributes to explore their 'impacts' to nodes connection. Link prediction is performed by removing some of the edges from original dataset. We compare different prediction methods: Common neighbor, Jaccard and Adamic. The prediction precision and recall is performed on choosing the Top-K highest probability edges, different performance of K is explored. Label propagation is performed by removing some fraction of the labels from original dataset, different fraction of each attributes are tested on Caltech dataset.

Finally, I'm interested in how to detect a existing community and more importantly, how this community will evolve in the future. Although this may be a label propagation process because communities can be considered to be a label, we add some intuitive hypothesis which make this question constrained natually. This question is explored by the following path:

- We first give some intuitive hypothesis and defined the conditional connection probability for arbitrary two nodes in the graphs;
- Second, the conditional probability is modeled by a MLP with hypothsis constrained; The possible existed communities are detected by a Kmeans-like algorithm;
- The potential communities are generated by a diffusion model conditioned on the MLP computed probability.

Please refer to the detail in part Q5. The experiments notebook can be found in github:

`https://github.com/BaozhuoSU/graphLearning`

Question 2

The degree distribution of Caltech, MIT and JohnsHop

The degree distributions in Fig. 1 exhibit heavy tailed behavior across all three datasets, with particularly long tails for the MIT and Johns Hopkins networks. This indicates a small fraction of nodes act as hubs with very large numbers of connections, while the majority of students maintain relatively low-degree profiles.

We can observe that Caltech's distribution decays more rapidly, this may because high-degree nodes may serve as bridges across multiple communities and disproportionately shape the global connectivity pattern.

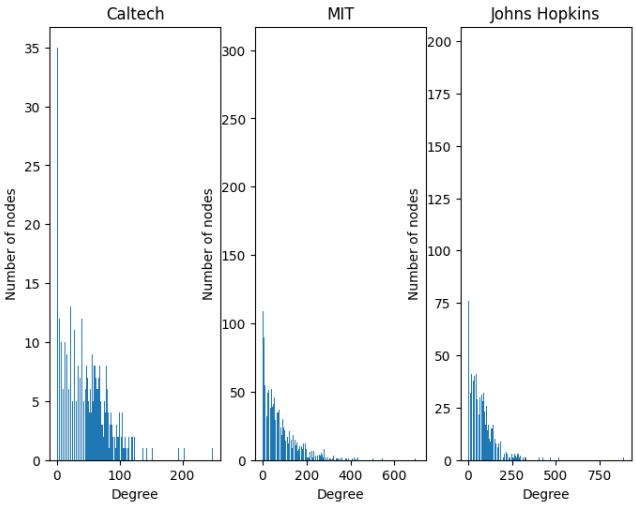


Figure 1: Degree distritbuion of Caltech, MIT and JohnsHopkins. Caltech's tail is shorter than other two.

Transitivity, clustering Coefficient et Density

From table 1, all of the three graphs yield low density. We can concludes that the 3 graphs are sparse, which is consisting with the degree distribution. For Caltech it yields a little bit higher transitivity, so there are more closed triangles in caltech, there are more clusterings in caltech. The graph of caltech is more 'compact', more little groups.

In contrast, the MIT network exhibits the lowest density and lower transitivity, indicating a more globally sparse and less locally clustered structure. The Johns Hopkins (JH) network lies between these two extremes.

Table 1: Transitivity and Density for each Graph

Graph	Transitivity	Density
Caltech	0.29	0.06
MIT	0.18	0.01
JH	0.19	0.02

Degree vs. Local clustering Coefficient

From Fig. 2, we observe a clear inverse relationship between node degree and local clustering coefficient across all three networks: low-degree nodes tend to exhibit higher local clustering, while high-degree nodes are associated with lower clustering. This pattern is consistent with the definition of the local clustering coefficient, as nodes with fewer neighbors are more likely to have their neighborhoods fully interconnected.

The JH network is dominated by a large proportion of low-degree nodes, which contributes to the concentration of points in the high-clustering, low-degree region of the plot. In contrast, the Caltech network exhibits a narrower degree range and fewer high-degree nodes, the MIT network shows a broader spread in degree.

Q3 Assortativity

Due to the computation problem, we only use 'Caltech', 'MIT' and 'JH' three datasets to compute assortativity and do analysis. The distribituion is estimated using `seaborn.kdeplot`.

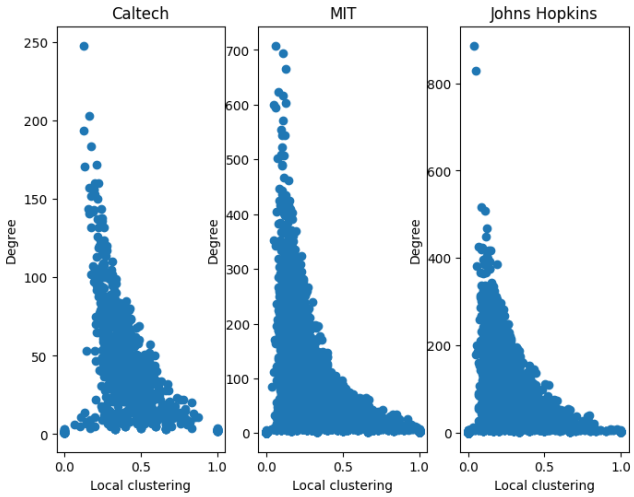


Figure 2: Degree vs. Local Cluster Coefficient

Since we only use 3 datasets, the estimation maybe overly biased.

For each attribute, we examine the assortativity distribution across university Facebook networks. The results show how strongly each social or structural factor influences the tendency of users to connect with others sharing the same attribute.

Dorm. The distribution centers around 0.2, clearly above zero, indicating a strong **homophily** by dormitory. Students living in the same dormitory are much more likely to become friends, which reflects the physical proximity and shared daily environment of residence halls.

Gender. As discussed earlier, gender assortativity averages around 0.02, slightly above zero. This indicates a very weak homophily by gender—people tend to friend those of the same gender slightly more often than expected by chance. However, the effect is small, and a few networks even show slight **heterophily**, possibly due to heteronormative social interactions.

High School. The values cluster around 0.05, suggesting a moderate homophily effect. Students who attended the same high school tend to connect after entering university, maintaining their pre-existing social ties.

Major. The distribution centers around 0.01, only slightly positive. This implies that while students in the same academic major are marginally more likely to connect, the effect is relatively weak—academic specialization is not a major determinant of online friendship structure.

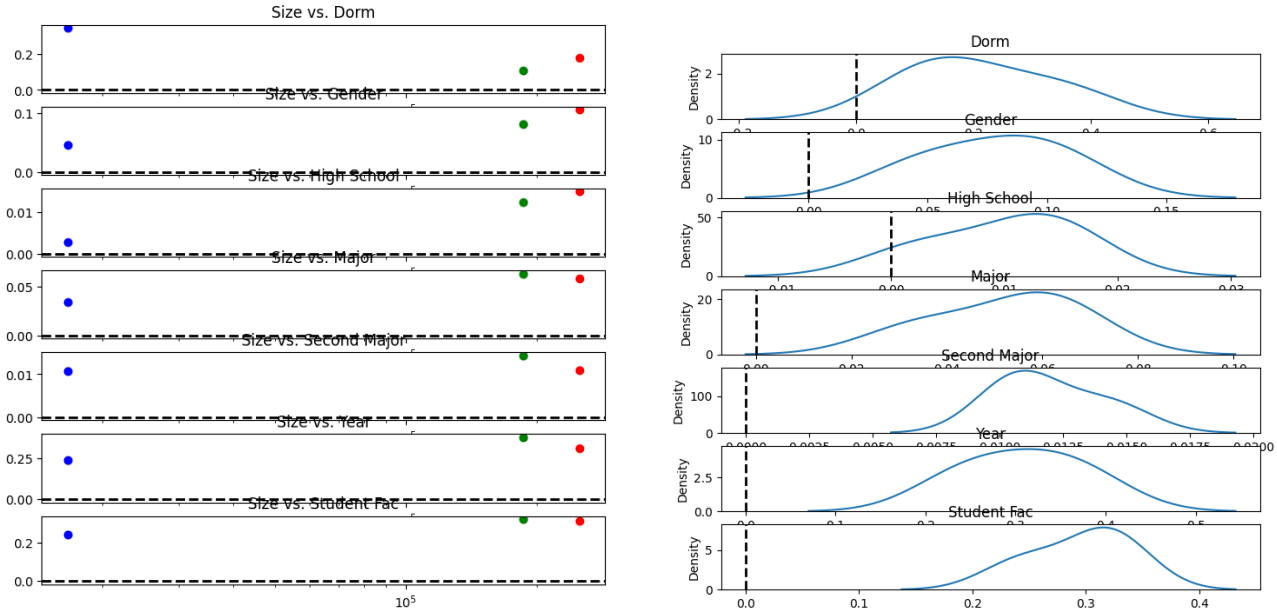
Second Major. Similar to the primary major, second-major assortativity remains weakly positive (around 0.02–0.06). Since fewer students share the same secondary field, these connections are sparser and more variable across schools.

Year. With an average near 0.02, year-based assortativity indicates a small but consistent preference for same-cohort friendships. This reflects the shared courses, activities, and temporal overlap among students of the same entry year.

Student Faculty. The faculty or department attribute exhibits the highest assortativity (around 0.3), indicating strong homophily within academic divisions. Students tend to form

the majority of their friendships within their own department, likely due to shared courses, labs, and physical co-location.

Summary. In summary, dormitory and faculty membership produce the strongest homophily effects, while gender, major, and year show only weak positive assortativity. These findings suggest that structural and spatial proximity are more influential than demographic or academic variables in shaping social connections on university Facebook networks.



(a) Size Vs. Assortativity

(b) Distribution of Assortativity

Figure 3: Assortitavity of three datasets. Blue=Caltech, red=MIT, green=JH

Q4 Link prediction

As shown in Fig. 4, the optimal value of K differs across the three methods. As K increases beyond this optimum, precision decreases from its peak value, while recall continues to increase monotonically. This behavior reflects the classical trade-off between precision and recall: retrieving a larger set of candidate links improves coverage of true positives but also introduces a higher proportion of false positives.

Q5 Label Propagation

As shown in Table 2, removing different node attributes leads to markedly different levels of degradation in link prediction performance, indicating that attributes contribute unequally to the connectivity structure of the Caltech network. In particular, removing the major attribute consistently results in the lowest prediction accuracy across all removal fractions, with performance dropping from 0.29 at a removal rate of 0.1 to 0.18 at 0.4. This suggests that academic major plays a dominant role in shaping the underlying connection patterns of the network.

In contrast, the prediction accuracy remains relatively stable when removing dormitory information, with values staying above 0.71 even at the highest removal fraction. This indicates that dormitory affiliation contributes less discriminative information for link formation in this dataset. The year and gender attributes exhibit intermediate behavior: their removal leads to a moderate but consistent decrease in accuracy, suggesting that they provide complementary, but not primary, signals for predicting connections.

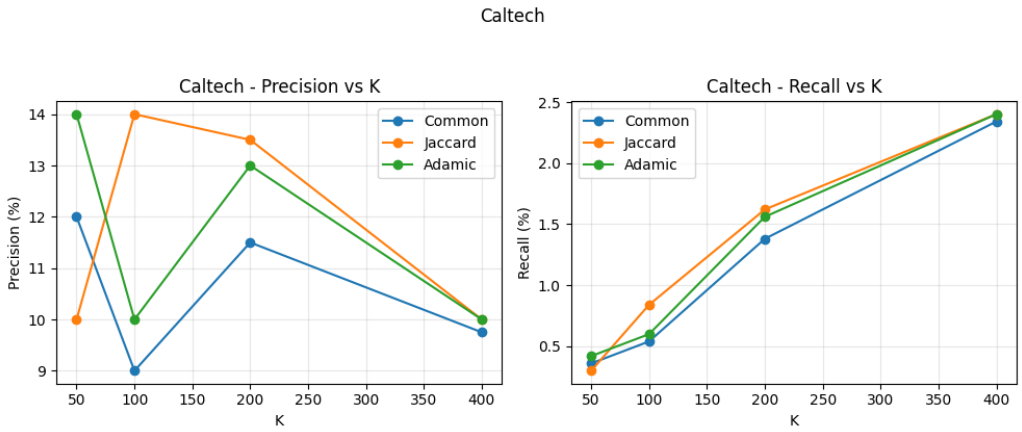


Figure 4: Precision and Recall Vs Top-K of Caltech

Table 2: Prediction accuracy of Caltech dataset for different attributes with different removing fraction. The iteration round is set to be fixed at 50.

Attributes Removed(Caltech)	0.1	0.2	0.3	0.4
Major	0.29	0.24	0.22	0.18
Dorm	0.74	0.75	0.73	0.71
Year	0.66	0.58	0.56	0.57
Gender	0.59	0.59	0.59	0.54

Q5 Communities detection and generation with the FB100 datasets

Research question and hypothesis

Here I'm interested in how the community will evolve in future. More precisely, the future community should show same pattern with the exist community, so the evolution process should be 'stable'. This evolution may be similar with label propagation, as communities can be considered to be a type of label. But here I have added some natural hypothesis so that this question is more more intuitive. The hypothesis is simple:

- Those who have connection will have more probability to form a community;
- The future community yields same pattern with exist community so the evolution will be stationary.
- The probability to be connected can be conditioned on attributes: If two nodes have similar attributes, they tend to be connected;

It is like Google matrix, the non-connection is not absolute, there may be a probability to be connected in the future. In this case, we can formulate the probability of two nodes to be connected in the future :

$$\text{Probability to be connected} = P(\text{Connected} \mid \text{Similairity}) \tag{1}$$

and constrained by above hypothesizes.

Learned Similarity

Then the next question is how to define the attributes similarity between two nodes. Some of the works try to measure the similarity as the Euclidean distance or correlation, these methods

have several drawbacks, such as the Euclidean distance is hard to compute especially if there are both qualitative or quantitative attributes. Since correlation is bounded between [-1,1] it is hard to interpret if two nodes are strongly negatively correlated.

Here, another idea is let a machine learning model to learn how to compute the similarity, for instance, we can use a MLP to model similarity:

$$\text{Similarity} = s_\theta(i,j) = \text{MLP}_\theta(x_i, x_j) \tag{2}$$

where x_i, x_j are the attributes vector of nodes i, j.

If we use sigmoïd to active, then the conditional probalility will be :

$$P(\text{Connected} \mid \text{Similairity}) = P_\theta(\text{Connected} \mid \text{Similairity}) = \sigma(s_\theta(x_i, x_j)) \tag{3}$$

In this case, the objective function will be the cross entropy :

$$\max_\theta \sum \log P_\theta + \sum \log(1 - P_\theta) \tag{4}$$

Community detection and generation

Once we have trained our model, we can use it to compute the connection probability of two arbitrary nodes. We can use it to detect potential community and generate potential community.

Community Detection using MLP.

To detect a community we can do it like a Kmeans algorithm. First initialize some community center the use the trained MLP to compute the probability that the nodes is connected to the center, then the node is distributed to the community with the highest connection probability center; repeat it until converge. Now we have the community label of each node $\hat{z} \in \{0, 1, 2, ...K\}$.

Community generation using Diffusion model.

The probability can also be used to generate the community with generative model, such as Diffusion model. In this case the Diffusion mode will be conditioned on the connection probability. The idea is simple, it is like a procedure of label prediction, we predict the community label $\mathbf{z}$ of each node. Next we will discuss how the Diffusion model work precisely in each stage.

Forward noise adding. In the forward process, we add noise to z_0, the label distributed by the community detection stage, we try to mess up the distributed label :

$$z_t \sim q(z_t \mid \hat{z})$$

Backward denoising. Intuitively, we can add the conditional probability P_θ into the denoising network,then nodes with higher connection probability will tend to be distributed with same community label:

$$p_\phi = (\hat{z} \mid z_t, P_\theta, x, t)$$

Objective function. In this case, our objective function will be cross entropy:

$$L_{diff} = \mathbb{E}[-\log(p_\theta(\hat{z}) \mid z, P_\phi, x, t)]$$

Results

We evaluate both community detection and generation using the sorted conditional probability matrix, as shown in Fig. 5. We analyze three datasets: Caltech, MIT, and Johns Hopkins (JH). Since the MIT and JH networks contain a large number of nodes and edges, we apply a Breadth-First Search (BFS)–based sub-sampling strategy to extract induced subgraphs while preserving the local connectivity structure.

From the MLP-based results, we observe that the diagonal blocks of the sorted probability matrices exhibit consistently higher values than the off-diagonal regions. This indicates that node pairs assigned to the same community are predicted to have a higher conditional probability of being connected. Since the MLP is trained to model attribute-based similarity between node pairs, this observation supports our hypothesis that attribute similarity is strongly associated with community formation and intra-community connectivity. Among the three datasets, the JH network shows the strongest contrast between intra- and inter-community probabilities, suggesting a more pronounced community structure under the learned connectivity model.

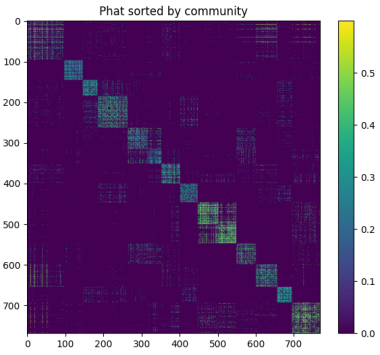
The probability matrices generated by the diffusion model exhibit a similar block-diagonal pattern, indicating that the generated community assignments are structurally consistent with those detected by the MLP-based model. This suggests that the diffusion process is able to generate plausible community configurations that preserve the learned connectivity patterns, rather than merely reproducing random partitions.

The graph-level visualizations further highlight qualitative differences across datasets. In particular, the communities in the Caltech network tend to occupy a larger spatial extent in the layout, visually suggesting a more dispersed intra-community structure, whereas the MIT and JH communities appear more compact and tightly clustered. This observation is consistent with the weaker contrast in the corresponding probability matrices for Caltech.

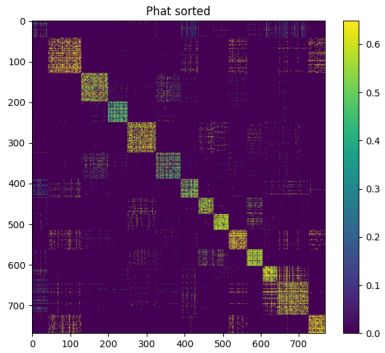
These results indicate that the proposed method not only learns a meaningful attributes conditioned connection probability, but also indicates that the diffusion model generated communities are consistent with exist communities.

Conclusion

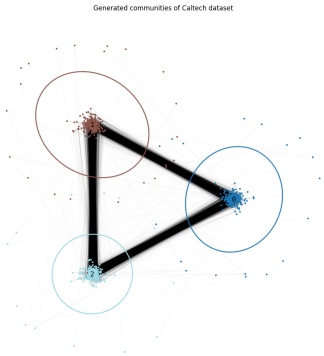
In this project, we investigated the structural properties and attribute drived mechanisms underlying social network formation using Facebook datasets. By analyzing degree distributions, transitivity, clustering, and assortativity, we characterized the sparsity, heterogeneity, and sources of homophily that shape connectivity patterns. We then developed an attribute conditioned link prediction framework based on a learned similarity model and demonstrated how it can be extended to community detection and diffusion based community generation. The experimental results show that the proposed approach captures meaningful intra-community connectivity patterns and produces generated communities that are structurally consistent with those detected in the original networks. This project highlights the interplay between network structure, node attributes, and generative modeling in understanding and simulating the evolution of social communities.



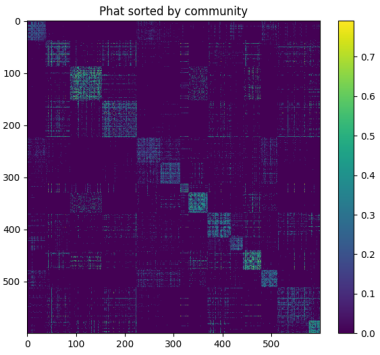
(a) Sorted probability matrix using MLP of Caltech dataset



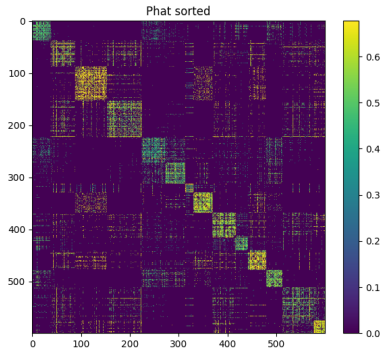
(b) Sorted probability matrix using Diffusion model of Caltech dataset



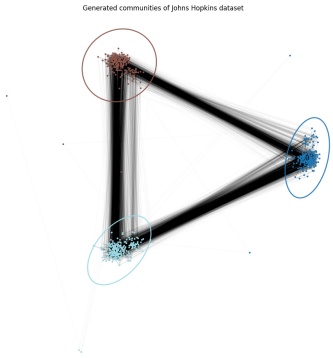
(c) Community generated by Diffusion model of Caltech



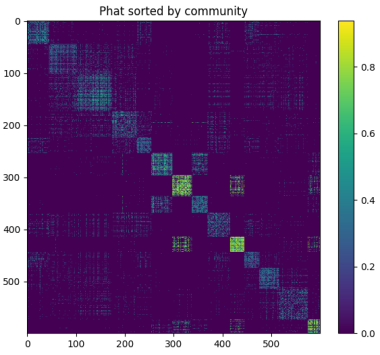
(d) Sorted probability matrix using MLP of MIT dataset



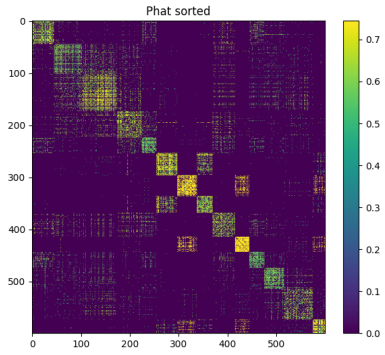
(e) Sorted probability matrix using Diffusion model of MIT dataset



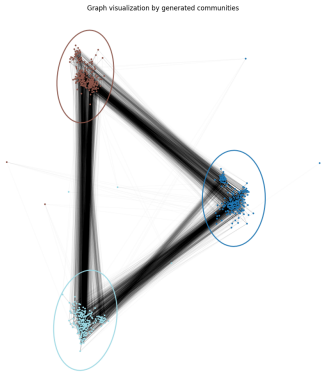
(f) Community generated by Diffusion model of MIT



(g) Sorted probability matrix using MLP of JH dataset



(h) Sorted probability matrix using Diffusion model of JH dataset



(i) Community generated by Diffusion model of JH

Figure 5