Network science and Graph Learning
NET 4103/7431 Homework

Rémi Khoury

1st February 2025

The code for this project is available at https://github.com/remik354/Network-science-and-Graph-Learning.

Contents

1 Intro

This report presents an analysis of network science and graph learning techniques applied to the facebook100 dataset. Various algorithms and metrics are explored, highlighting their significance in understanding complex network structures.

2 Social Network Analysis with the Facebook100 Dataset

(a) (1 point) For these three networks plot the degree distribution for each of the three networks that you downloaded. What are you able to conclude from these degree distributions?

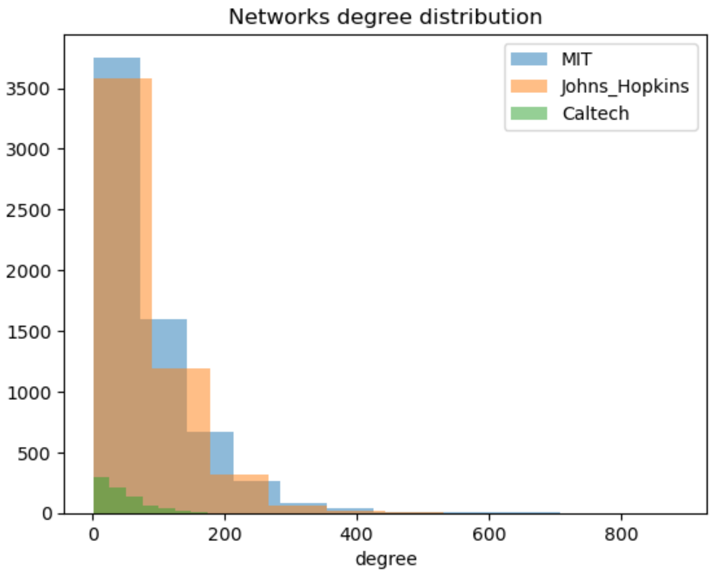

Figure 1: Degree distribution for MIT, John Hopkins and Caltech

The figure above seems logical with a social network. Indeed, most of the people have few friends and connections while few people have a large number of connections (= degree). We can also observe that the number of connection is linked to the size of the network, Caltech being the smalest university network and Mit / John Hopkins bigger ones.

(b) (1 point) Compute the global clustering coefficient and mean local clustering coefficient for each of the 3 networks. In addition compute the edge density of each network. Should either of these networks be construed as sparse? Based on the density information and the clustering information what can you said about the graph topology?

The results présented in the table above show that Caltech stands up compared to John Hopkins and MIT. We cannot call any of these network dense, eventhough the value of the average clustering coefficient may show the existence of several small groups in a quite

University	global clustering coef	average clustering	density
MIT	0.18	0.27	0.01
Johns Hopkins	0.19	0.27	0.01
Caltech	0.29	0.41	0.06

Table 1: Global clustering coefficient, Mean local clustering and Edge density of each network

sparse graph. We can also observe that Caltech has higher values in all 3 metrics above, which can be explained mostly thanks to the fact that Caltech is a much smaller network, which increases the local connectivity and overall density.

(c) (1 point) For each network, also draw a scatter plot of the degree versus local clustering coefficient. 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?

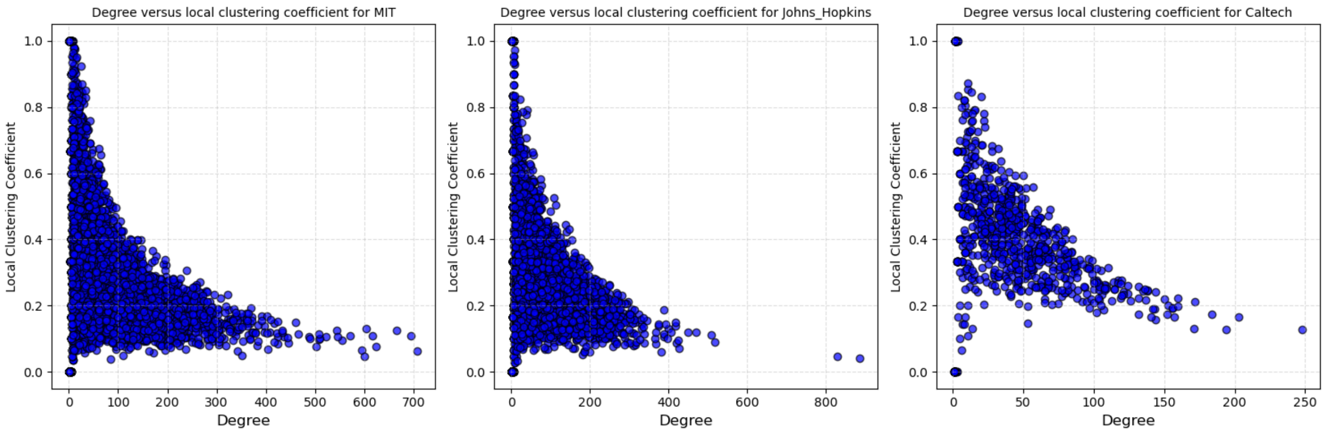

Figure 2: Degree distribution versus local clustering for MIT, John Hopkins and Caltech

According to the figues above, we can see that MIT and John hopkins are very similar in terms of the organization and distribution of the graph. In contrast, the figure representing Caltech enables us to draw some observations : there are fewer people who have both a low degree and a low local clustering coefficient, and the average clustering coefficient is higher. This shows that there are fewer people who are not included in the community as a whole and that the students at Caltech are all somehow part of a group.

3 Assortativity Analysis with the Facebook100 Dataset

(0) (2 points) Of the FB100 networks, investigate the assortativity patterns for five vertex attributes: (i) student/faculty status, (ii) major, (iii) vertex degree, and (iiii) dorm, (iiiii)

gender. 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.

The figure is available at the end of the report, I was not able to fit it here properly.

Except for the degree and the gender (where the assortativity is less relevant), it seems that the major, dorm and student fac are very assortative factors. These 3 factors leads the students to create links more than gender or degree (= being popular or unpopular will not help you connecting with popular or resp. unpopular students).

4 Link Prediction

We implemented severall link prediction metrics (CommonNeighbors, Jaccard and AdamicAdar).
As an example, the followwing table presents the results for the MIT network, with the Common Neighbor metric, with 20% of all nodes removed.

Top-k	**Top@k Size**	**Precision**	**Recall**
50	43	0.8600	0.0009
100	90	0.9000	0.0018
150	132	0.8800	0.0026
200	176	0.8800	0.0035
250	216	0.8640	0.0043
300	257	0.8567	0.0051
350	299	0.8543	0.0060

Table 2: Performance metrics for different Top-k values

To find out which metric would be best suited, we took the first 10 networks of the dataset and computed the metrics on each of them, removing some fractions of their networks.

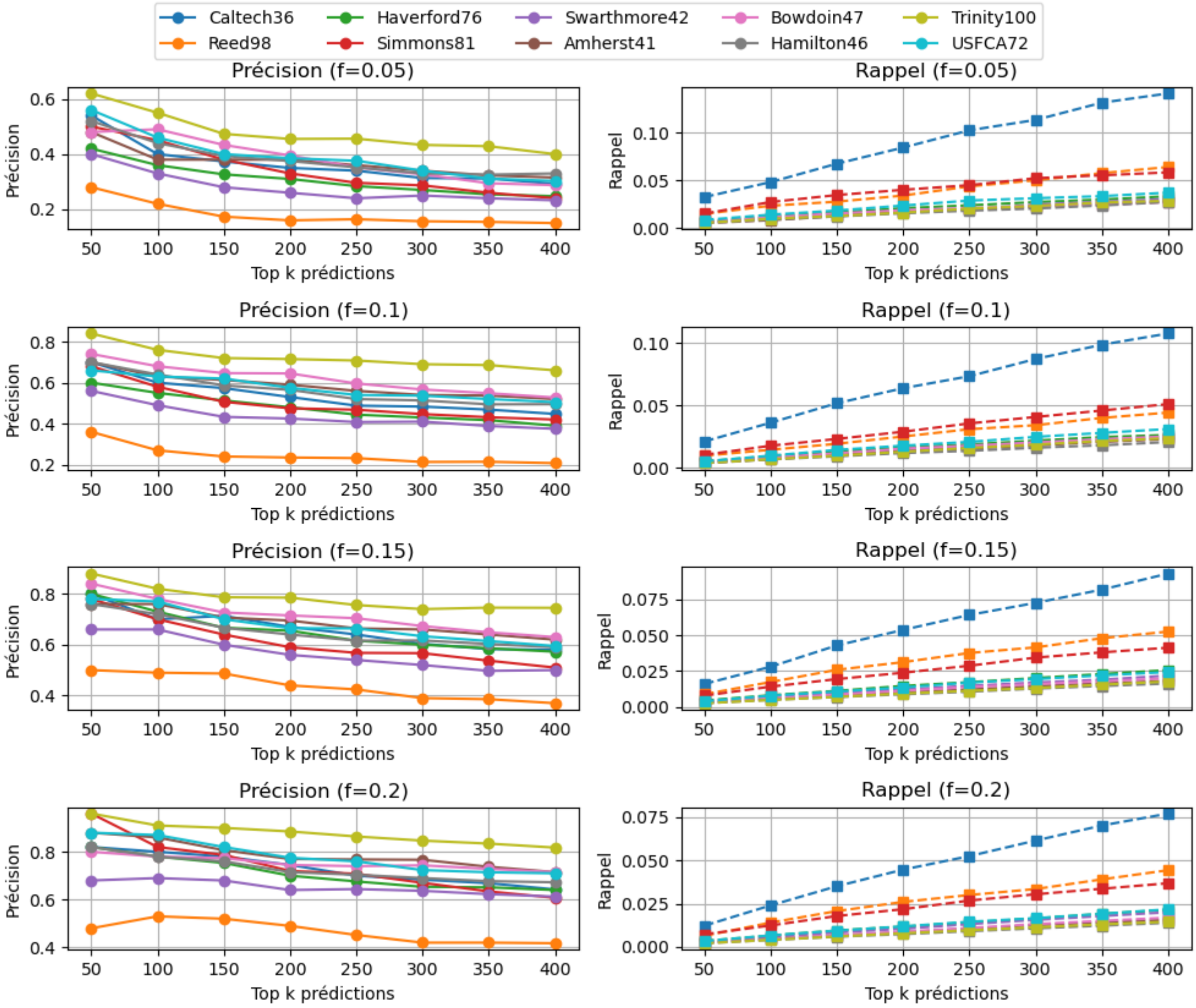

Figure 3: Precision report for Common Neighbors metric, on several fraction of each networks

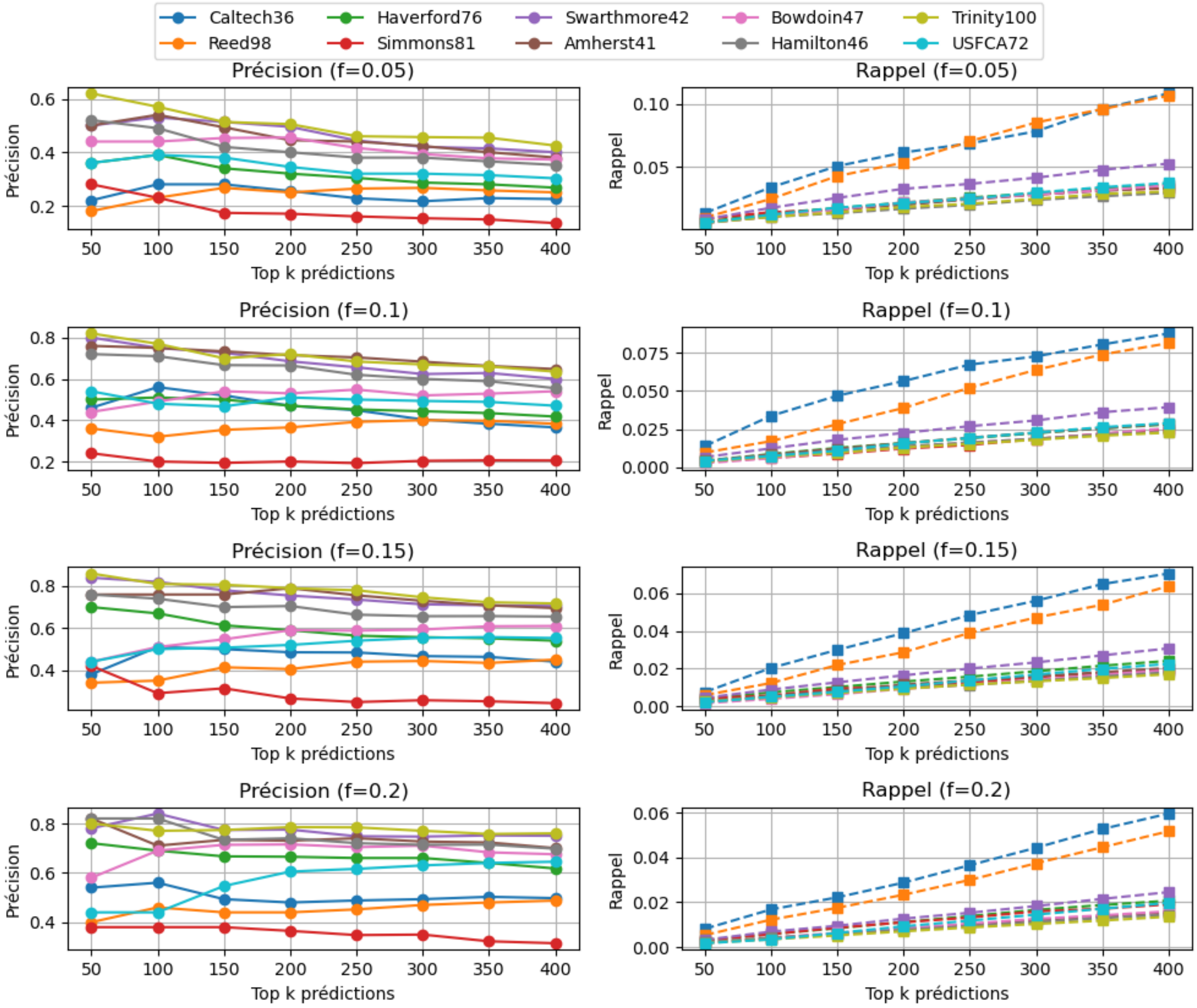

Figure 4: Precision report for Jaccard metric, on several fraction of each networks

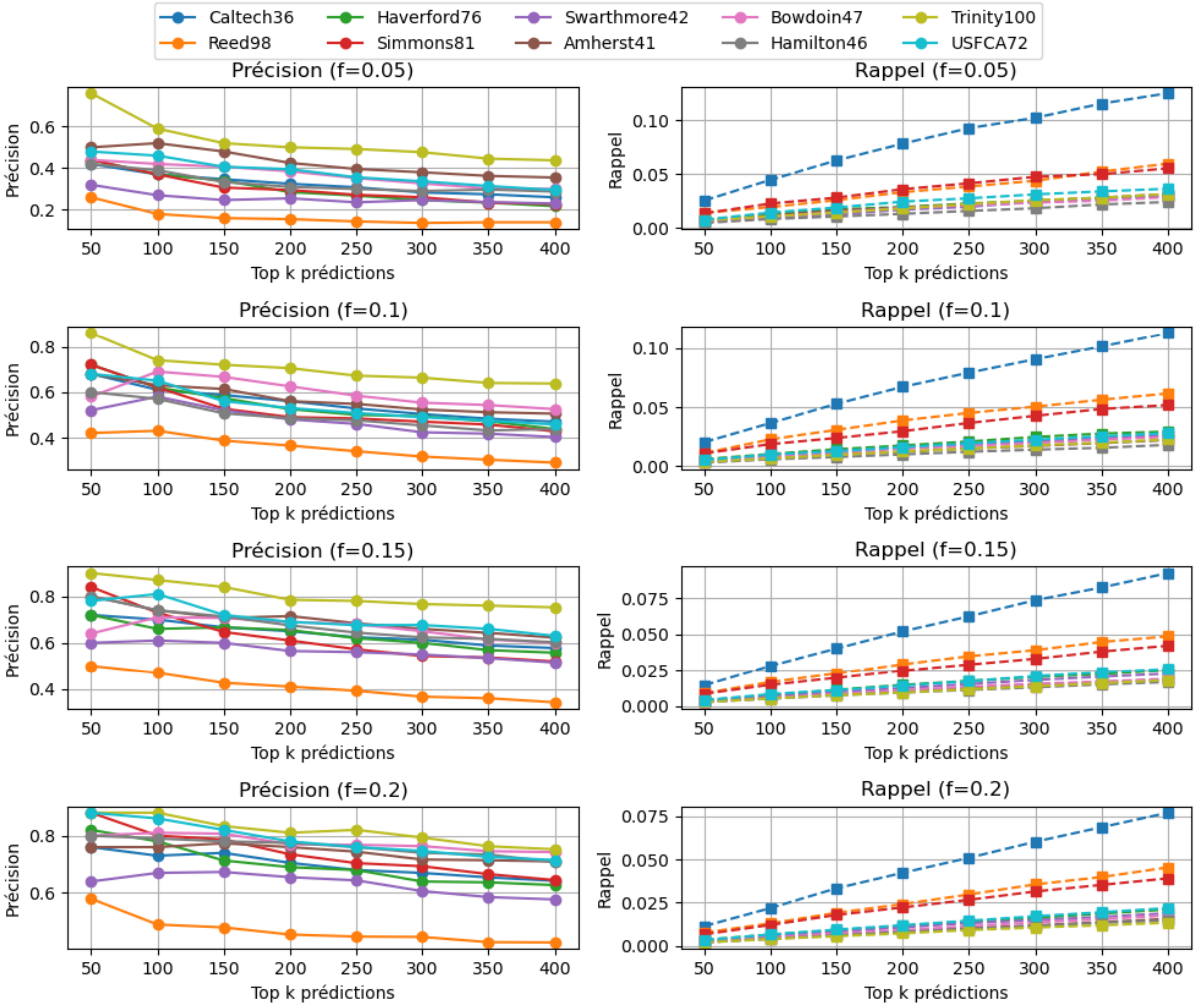

Figure 5: Precision report for Adamic Adar metric, on several fraction of each networks

On the figure above, the performance of each metrics are very similar. However, it seems like the AdamicAdar metric performs the best overall, even with a large number of nodes removed.

5 Find missing labels with the label propagation algorithms

(0) Compute the accuracy and the MAE for the label propagation algorithm (1) (1 point) 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 ?

We computed the Label Propagation algorithm on the MIT network. The table below presents the results of the algorithm, for each of the fraction of the labels removed. The first table presents the accuracy and the second table the MAE score.

Attribute	**10% removed**	**20% removed**	**30% removed**
major_index	0.92	0.84	0.76
dorm	0.96	0.90	0.81
gender	0.96	0.92	0.87

Table 3: LPA accuracy results for different levels of removed labels

Attribute	**10% removed**	**20% removed**	**30% removed**
major_index	0.54	1.11	1.67
dorm	6.88	19.5	39.80
gender	0.04	0.084	0.139

Table 4: LPA error rate results for different levels of removed labels

The Label Propagation Algorithm performs best for gender (highest accuracy) and slightly worse for dorm, while major index has the lowest accuracy. This difference is due to network structure and gender connections are typically stronger and more clustered, making propagation more effective. Dorm assignments also form tight communities, though with some inter-dorm connections reducing accuracy. On the contrary, academic majors are more diffuse across the network, leading to weaker label propagation.

We can imagine that extra curricular activities play an important role in community creation, and so the study field is not prominent. Also, looking at the data, the number of majors may also bring some bias into the model, meaning that with a large number of major (not linked in clusters, like for example 'science majors') the mae and accuracy may be affected (while there is only 2 gender to take into account).

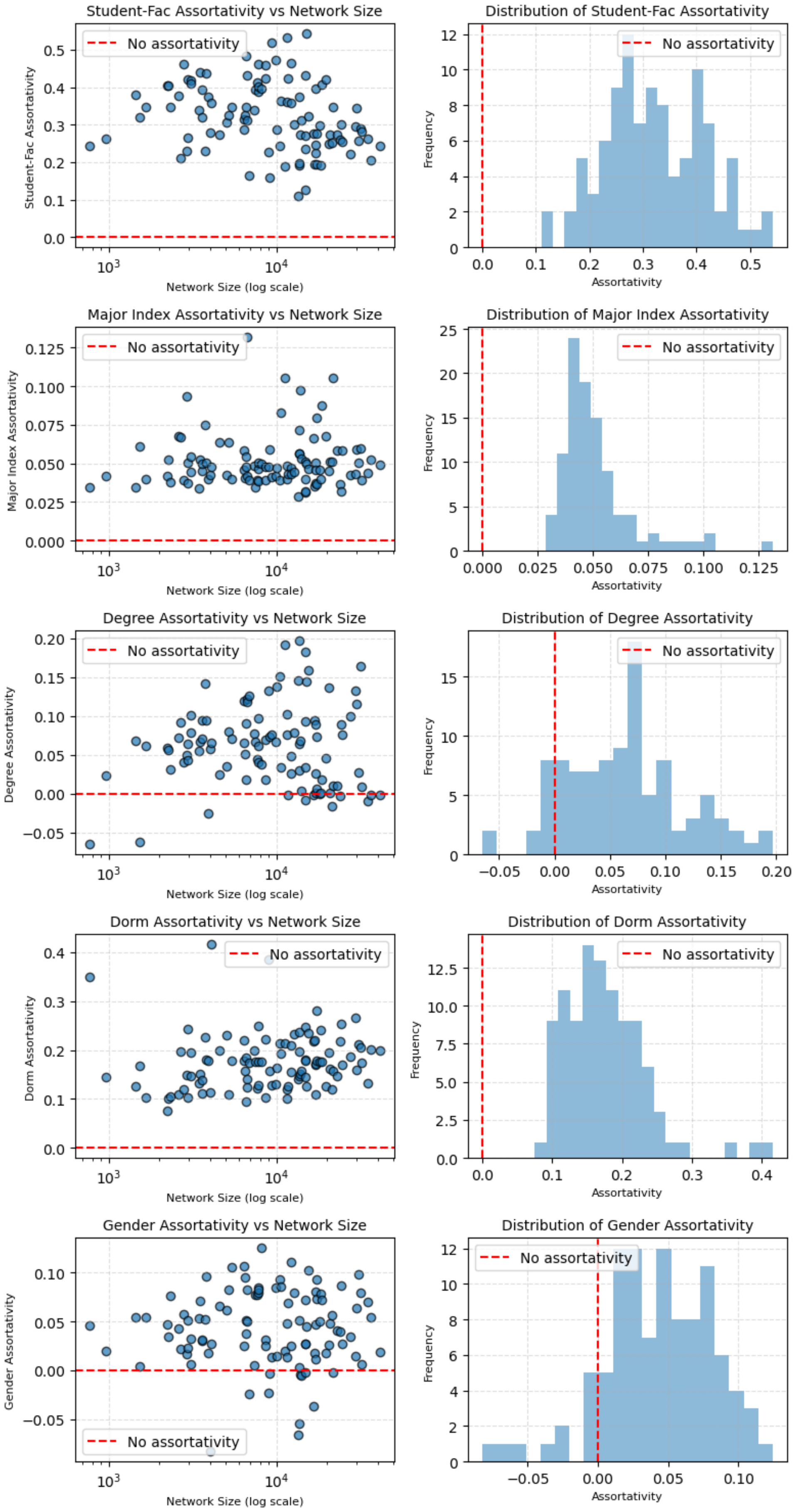

Figure 6: Assortativity patterns for five vertex attributes, computed on the facebook100 dataset