Network Science and Graph Learning

Homework

January 31, 2025

by

Vuong Quoc Anh VU
vuong-quoc-anh.vu@telecom-sudparis.eu
GitHub: https://github.com/vvquocanh/network-graph

Telecom Sudparis

1 Question 2: Social Network Analysis with the Facebook100 Dataset

a) Figure 1, 2, and 3 display the degree distributions of three social networks, including Caltech, MIT, and John Hopkins. From these three degree distributions, we can conclude that they exhibit a heavy-tailed pattern, suggesting that most nodes have a low degree. In contrast, a few nodes have a very high degree, characteristic of scale-free networks, meaning that they all follow the power law. Therefore, this indicates that well-connected nodes are more likely to gain new connections. However, there are also differences among them. While Caltech has a more scattered degree distribution, indicating a smaller and potentially less connected network, the MIT and Johns Hopkins networks have more data nodes and exhibit a clearer power-law decay. Besides, Johns Hopkins has a broader degree range, with degrees extending close to 10^3 proving that this network has some highly influential nodes.

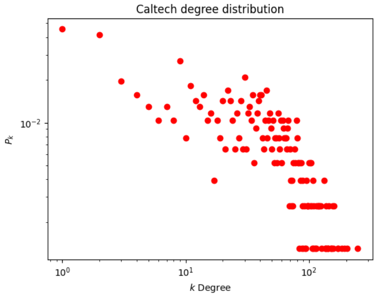

Figure 1: Caltech degree distribution

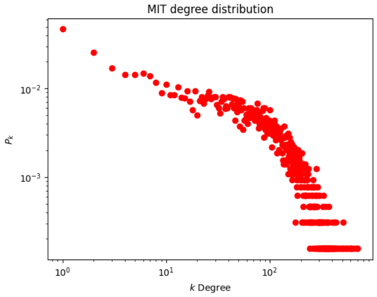

Figure 2: MIT degree distribution

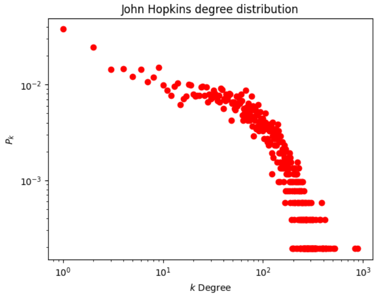

Figure 3: John Hopkins degree distribution

b) Table 1 depicts the global clustering coefficient, mean local clustering coefficient, and the edge density of three networks: Caltech, MIT, and John Hopkins. Based on the computed data, it's safe to say that all three networks are sparse because of their relatively low edge densities. However, MIT and Johns Hopkins are significantly sparser than Caltech when Caltech is nearly 5 times denser compared to them. This structure of the Caltech network is typical of small-world networks, which often have high clustering and short average path lengths. Whereas, MIT and John Hopkins have fewer tightly-knit local groups following the hub-like structures with scale-free network characteristics.

Network	Local Clustering	Global Clustering	Edge Density
Caltech	0.409	0.291	0.056
MIT	0.271	0.180	0.012
Johns Hopkins	0.268	0.193	0.014

Table 1: Clustering coefficients and edge density for different networks.

c) Figure 4, 5, and 6 illustrate the scatter plot of the degree versus local of three networks: Caltech, MIT, and John Hopkins. In general, they have the same pattern when nodes with lower degrees tend to have higher local clustering coefficients and nodes with higher degrees tend to have lower clustering coefficients clustering coefficient. This suggests a common structural property in a social network when a person with just a few close friends (low degree) is likely in a tight-knit friend group where everyone knows each other (high clustering). On the other hand, a celebrity (high degree) has many connections, but their followers are not necessarily connected (low clustering). Besides, there are also dissimilarities among them. It can be seen clearly that the Johns Hopkins and MIT networks have a broader range of degrees compared to the Caltech network. Johns Hopkins, in particular, has nodes with degrees extending beyond 800 when the maximum degree of a node in the Caltech is significantly lower with only around 250. In addition, Caltech shows a relatively higher clustering coefficient for mid-degree nodes meaning the network is more locally clustered and aligns with a higher global clustering coefficient. In contrast, MIT and Johns Hopkins have more hub-and-spoke-like network structures. In those graphs, we can also see a rapid drop in the clustering coefficient for high-degree nodes in MIT and Johns Hopkins with values around 0.05 indicating that they act as bridges connecting different clusters.

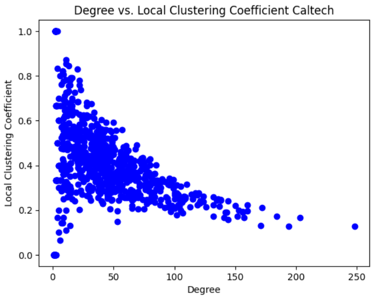

Figure 4: Caltech degree versus local clustering coefficient

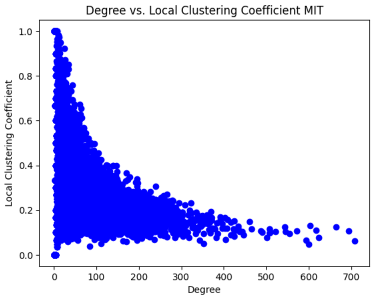

Figure 5: MIT degree versus local clustering coefficient

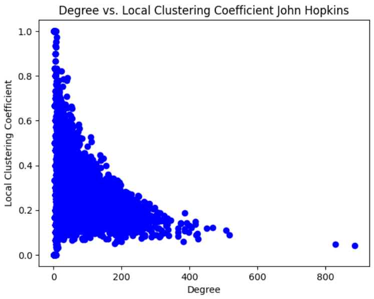

Figure 6: John Hopkins degree versus local clustering coefficient

2 Question 3: Assortativity Analysis with the Facebook100 Dataset

Figure 7 and 8 demonstrate the assortativity on student/faculty status of universities' members. The assortativity of this attribute is strong with most of the values lie between 0.2 to 0.5 with the mean value around 0.35. This strong positive assortativity indicates people tend to make friends with others who share a similar role to them such as student with student, staff with staff, or professor with professor. This reflects a the natural academic social when there are professional boundaries between these groups.

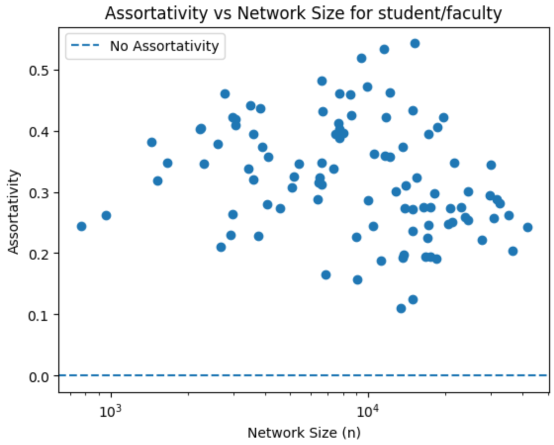

Figure 7: Student/Faculty assortativity versus network size

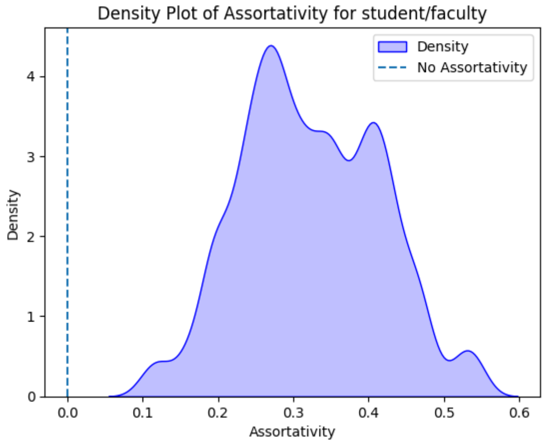

Figure 8: Student/Faculty distribution of assortativity

Figure 9 and 10 demonstrate the assortativity on the major attribute. The distribution is primarily positive with the values falling between 0.038 to 0.06, slightly assortative with the mean around 0.05. The numbers suggest a weak but consistent tendency for students of the same major to connect. This indicates some tendencies based on academic fields and the time students from the same field are likely to spend with each other.

Figure 11 and 12 show the assortativity of the vertex degree. On average, the mean value of the assortativity is about 0.08 showing students with higher social activity are more likely to form connections with other highly active students. Similarly, students with fewer connections may form bonds with others who are similarly less involved. This pattern suggests that, within the social structure of a university, students tend to form groups based on their similarity of engagement levels, interests, or extracurricular activities. However, a few universities have a negative value of assortativity, or they are disassortative. This can be

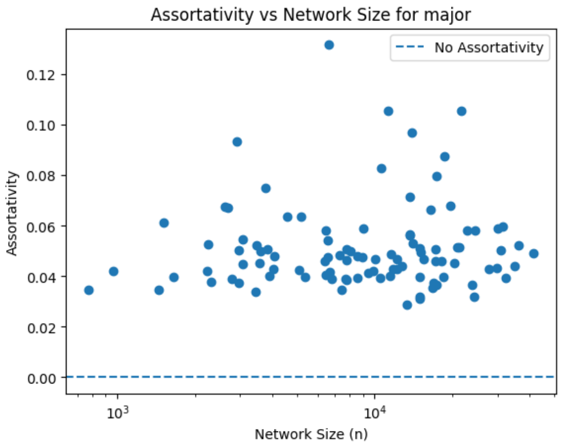

Figure 9: Major assortativity versus network size

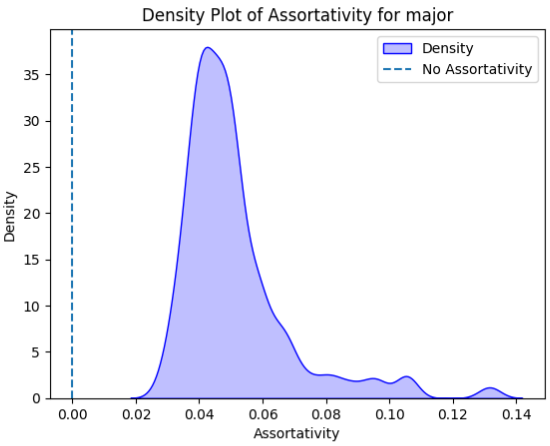

Figure 10: Major distribution of assortativity

understood by the reason that those universities are structured in a way where active students are more likely to help or guide those who are less involved, potentially through mentoring or social integration efforts.

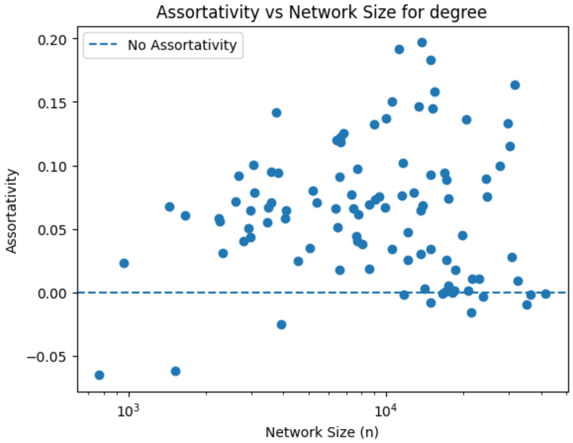

Figure 11: Degree assortativity versus network size

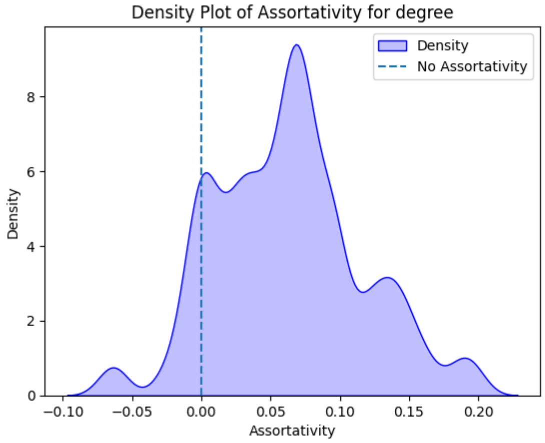

Figure 12: Degree distribution of assortativity

Figure 13 and 14 illustrate the assortativity on the dorm attribute. It can be seen easily that the distribution mostly ranges from 0.1 to 0.25 with the mean value around 0.2. This moderately positive assortativity suggests that people living in the same dorm are more likely to be connected, which is logical given the increased opportunities for social interaction among dorm residents.

Figure 15 and 16 depict the assortativity of the gender attribute. The pattern of this assortativity is explained thoroughly in the question section.

3 Question 4: Link prediction

To evaluate the performances of different link prediction metrics including Common Neighbors, Jaccard, and Adamic/Adar, the experiment is conducted on two different data sets: Princeton with 6596 nodes (a quite large network) and Caltech with 762 nodes (a small network). Tables 2, 3, and 4 show the precision and recall of the Common Neighbors, Jaccard, and Adamic/Adar applying on the Princeton network respectively. Similarly, tables 5, 6, and 7 represent the link prediction metrics results of the Caltech network.

In general, the Jaccard predictor performs the worst compared to other predictors on both datasets. Its precision ranges from 0 to 0.2 in the Princeton data set and from 0.1 to 0.3 in the Caltech one, with different

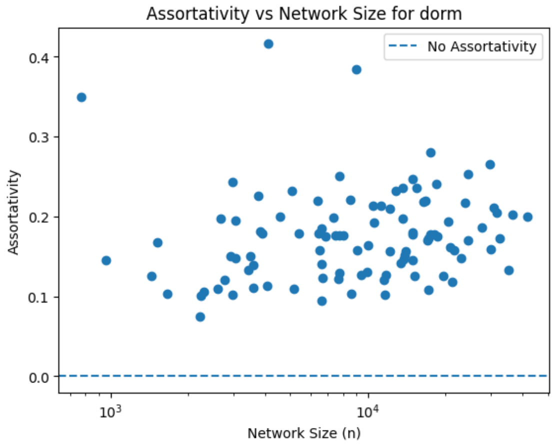

Figure 13: Dorm assortativity versus network size

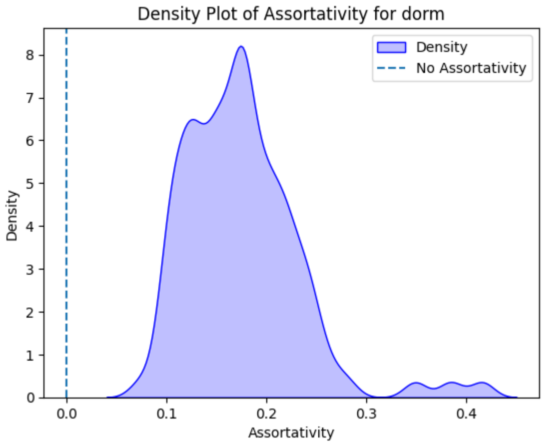

Figure 14: Dorm distribution of assortativity

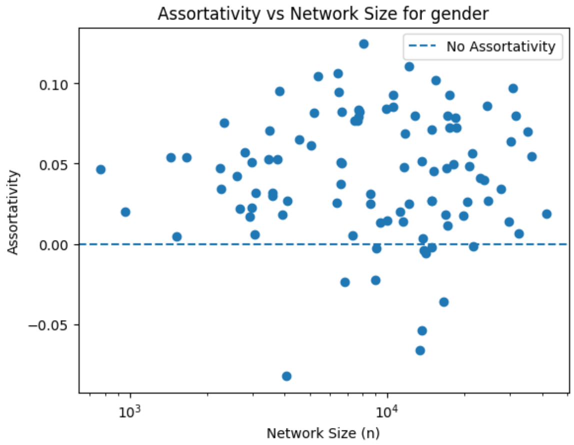

Figure 15: Gender assortativity versus network size

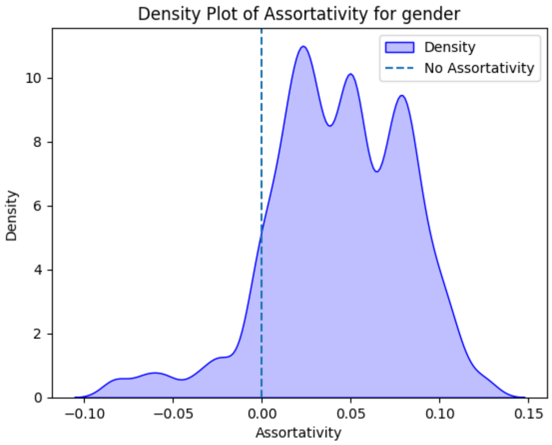

Figure 16: Gender distribution of assortativity

values of the removal fraction and k prediction pairs. The recall is also very low, with only 0 to 0.004 in the Princeton data set and 0.002 to 0.05 in the Caltech data set. It can be understood because Jaccard normally performs poorly on graphs that have many hubs (high-degree nodes), which is one of the characteristics of those social networks, or having an uneven degree distribution. It can be proved by the formula of the Jaccard algorithm that having more hubs leads to a decrease in scores.

On the other hand, Common Neighbors and Adamic/Adar perform better than Jaccard overall, whereas Adamic/Adar is slightly better. In the Princeton data set, Common Neighbors has a precision ranging from 0.23 to 0.44 and a recall ranging from 0.0003 to 0.0063. Besides, Adamic/Adar scores fall between 0.24 to 0.46 for the precision and 0.0004 to 0.0065 for the recall. The differences between those performances stay the same in the small Caltech data set when Common Neighbors can only reach the precision from 0.1 to 0.34 for the precision and 0.0045 to 0.05 for the recall. while in contrast, the values of Adamic/Adar are 0.13 to 0.38 and 0.006 to 0.65 respectively. The reasons behind their better performances are that Common Neighbors normally works well in low clustering graphs (which are shown true in Question 2) and Adamic/Adar discounts high-degree shared neighbors making it usually perform well in most real-world graphs.

	k pairs									
	50		100		200		300		400	
Frac.	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.38	0.0013	0.37	0.0025	0.28	0.0038	0.2467	0.0051	0.2325	0.0063
0.10	0.44	0.0008	0.35	0.0012	0.29	0.0020	0.2967	0.0030	0.305	0.0042
0.15	0.36	0.0004	0.34	0.0008	0.39	0.0018	0.4033	0.0028	0.3825	0.0035
0.20	0.30	0.0003	0.36	0.0006	0.375	0.0013	0.3767	0.0019	0.39	0.0027

Table 2: Common Neighbors results of Princeton

	k pairs									
	50		100		200		300		400	
Frac.	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.10	0.0003	0.14	0.0009	0.105	0.0014	0.12	0.0025	0.145	0.0039
0.10	0.10	0.0002	0.12	0.0004	0.135	0.0009	0.12	0.0012	0.1625	0.0022
0.15	0.04	0.0001	0.13	0.0003	0.165	0.0008	0.2133	0.0015	0.245	0.0022
0.20	0.00	0.0000	0.17	0.0003	0.13	0.0004	0.1933	0.0010	0.215	0.0015

Table 3: Jaccard results of Princeton

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

Table 8 displays the result from the label propagation algorithm applied on the Columbia network. Throughout the experiment, we can see that gender has the highest accuracy and lowest mean absolute error among the three labels suggesting that gender information is relatively easy to propagate through the graph. In addition, the dorm attribute has a moderate accuracy and mean absolute error with a slight fluctuation with different removal percentages. This indicates that dorm assignments are somewhat well-structured but not as strongly connected as gender. Last but not least, the major has the lowest accuracy with a sharp decline as more labels are removed. The accuracy drops from 0.2795 at 10% removal to just 0.1719 at 30% removal, and the mean absolute error increases significantly. Therefore, we can conclude that major indices are harder to infer using label propagation.

There are many reasons for those dissimilarities. Firstly, graph structure and homophily: Gender likely exhibits strong homophily, which is shown in the assortativity, whereas dorm and major have less. Secondly, gender is typically binary making it easier to infer while major likely has many distinct values, making it much harder to predict with label propagation. Thirdly, students of the same dorm or gender seem more densely connected making label propagation work well. Conversely, propagation becomes unreliable when the major connections are sparse.

5 Question 6: Communities detection with the FB100 datasets

a) The research question can be the impact of labels in forming community groups. The hypothesis believes that in real networks, each node usually contains multiple attributes representing the node's characteristics. Therefore, there might exist dominant attributes having effects on community formation. The purpose of the experiment is to find the dominant attribute of those networks with different community detection algorithms.

b) The experiment is run with three different data sets: Princeton, Harvard, and Indiana. Two community detection algorithms chosen to run are Louvain Community Detection and Label propagation community detection.

c) Tables 9, 10, and 11 show the result after running the experiment with the Louvain Community Detection algorithm on three networks Princeton, Harvard, and Indiana respectively. Note that those tables only represent the 7 largest communities in each network. It is shown that the student/faculty label seems

	k pairs									
	50		100		200		300		400	
Frac.	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.26	0.0009	0.28	0.0019	0.25	0.0034	0.24	0.0049	0.24	0.0065
0.10	0.34	0.0006	0.35	0.0012	0.34	0.0023	0.33	0.0034	0.3225	0.0044
0.15	0.34	0.0004	0.39	0.0009	0.375	0.0017	0.37	0.0025	0.38	0.0035
0.20	0.46	0.0004	0.46	0.0008	0.425	0.0014	0.4333	0.0022	0.4125	0.0028

Table 4: Adamic/Adar results of Princeton

	k pairs									
	50		100		200		300		400	
Fraction	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.2	0.0120	0.2	0.0240	0.115	0.0276	0.1067	0.0385	0.0975	0.0469
0.10	0.3	0.0090	0.26	0.0156	0.255	0.0306	0.2133	0.0384	0.2025	0.0486
0.15	0.34	0.0068	0.27	0.0108	0.245	0.0196	0.2533	0.0304	0.2375	0.0380
0.20	0.3	0.0045	0.29	0.0087	0.3	0.0180	0.2967	0.0267	0.2925	0.0351

Table 5: Common Neighbors results of Caltech

to be the dominant label among all other labels with the occupy rate always greater than 50% and around 90% on average. This suggests that the Louvain algorithm effectively captures structural patterns aligned with institutional roles, reinforcing the notion that academic affiliations play a crucial role in community formation within these university networks. However, there is a bias for the student/faculty attribute since in most universities, the number of students usually outweighs the number of other roles. This might lead to the unreliability of the experiment when the value ”student” of the student/faculty is way higher than other values. Therefore, it is not enough to prove that the student/faculty label is the dominant label that is in charge of forming communities. The same result pattern happens with the Label propagation community detection algorithm shown in tables 12, 13, and 14 with 3 largest communities.

	k pairs									
	50		100		200		300		400	
Fraction	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.16	0.0096	0.15	0.0180	0.105	0.0252	0.1033	0.0373	0.1025	0.0493
0.10	0.18	0.0054	0.23	0.0138	0.17	0.0204	0.17	0.0306	0.1675	0.0402
0.15	0.12	0.0024	0.22	0.0088	0.195	0.0156	0.2033	0.0244	0.21	0.0336
0.20	0.24	0.0036	0.32	0.0096	0.27	0.0162	0.2667	0.0240	0.2575	0.0309

Table 6: Jaccard results of Caltech

	k pairs									
	50		100		200		300		400	
Fraction	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
0.05	0.28	0.0168	0.26	0.0313	0.19	0.0457	0.1533	0.0553	0.135	0.0649
0.10	0.38	0.0114	0.33	0.0198	0.26	0.0312	0.2567	0.0462	0.2425	0.0583
0.15	0.32	0.0064	0.31	0.0124	0.285	0.0228	0.2733	0.0328	0.265	0.0424
0.20	0.38	0.0057	0.35	0.0105	0.35	0.0210	0.33	0.0297	0.295	0.0354

Table 7: Adamic/Adar results of Caltech

Attribute	**% Removed**	**Accuracy**	**Mean Absolute Error (MAE)**
Dorm	10%	0.5259	11.4359
	20%	0.4269	14.8849
	30%	0.5296	11.2855
Major Index	10%	0.2795	46.4911
	20%	0.1971	50.1138
	30%	0.1719	54.1192
Gender	10%	0.5641	0.4894
	20%	0.4758	0.5743
	30%	0.4619	0.5738

Table 8: Label Propagation Accuracy and MAE of Columbia

Size	**Label**	**Occupy**
1835	student_fac	55.42%
1121	student_fac	97.23%
1066	student_fac	98.87%
1046	student_fac	51.91%
960	student_fac	97.71%
286	gender	52.80%
131	student_fac	82.44%

Table 9: Dominant labels of Princeton using Louvain

Size	**Label**	**Occupy**
4800	student_fac	87.19%
2722	student_fac	55.95%
1858	student_fac	92.95%
1761	student_fac	92.28%
1669	student_fac	97.72%
1258	gender	53.50%
978	student_fac	55.83%

Table 10: Dominant labels of Harvard using Louvain

Size	**Label**	**Occupy**
8404	student_fac	79.49%
4928	student_fac	99.61%
4278	student_fac	64.94%
4224	student_fac	89.58%
1748	student_fac	82.84%
1471	student_fac	88.58%
1221	student_fac	83.87%

Table 11: Dominant labels of Indiana using Louvain

Size	**Label**	**Occupy**
4435	student_fac	55.83%
1053	student_fac	98.29%
1044	student_fac	99.33%

Table 12: Dominant labels of Princeton using Label propagation

Size	**Label**	**Occupy**
11693	student_fac	50.45%
1620	student_fac	96.42%
1484	student_fac	98.72%

Table 13: Dominant labels of Harvard using Label propagation

Size	**Label**	**Occupy**
25396	student_fac	79.56%
4317	student_fac	99.72%
3	student_fac	33.33%

Table 14: Dominant labels of Indiana using Label propagation