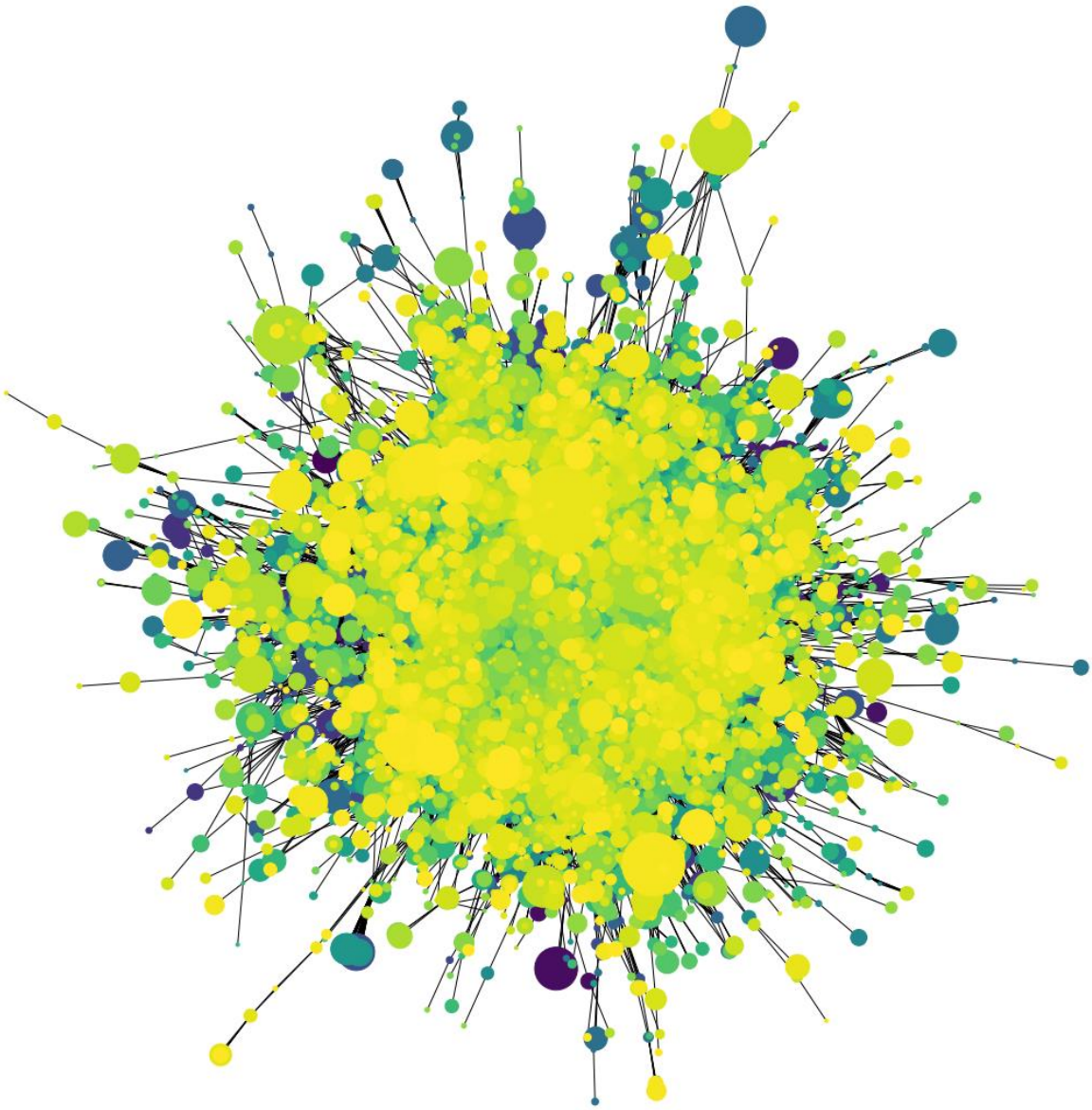
Social and Graph Data

NETWORK ANALYSIS PROJECT



Student: LE Trung Viet

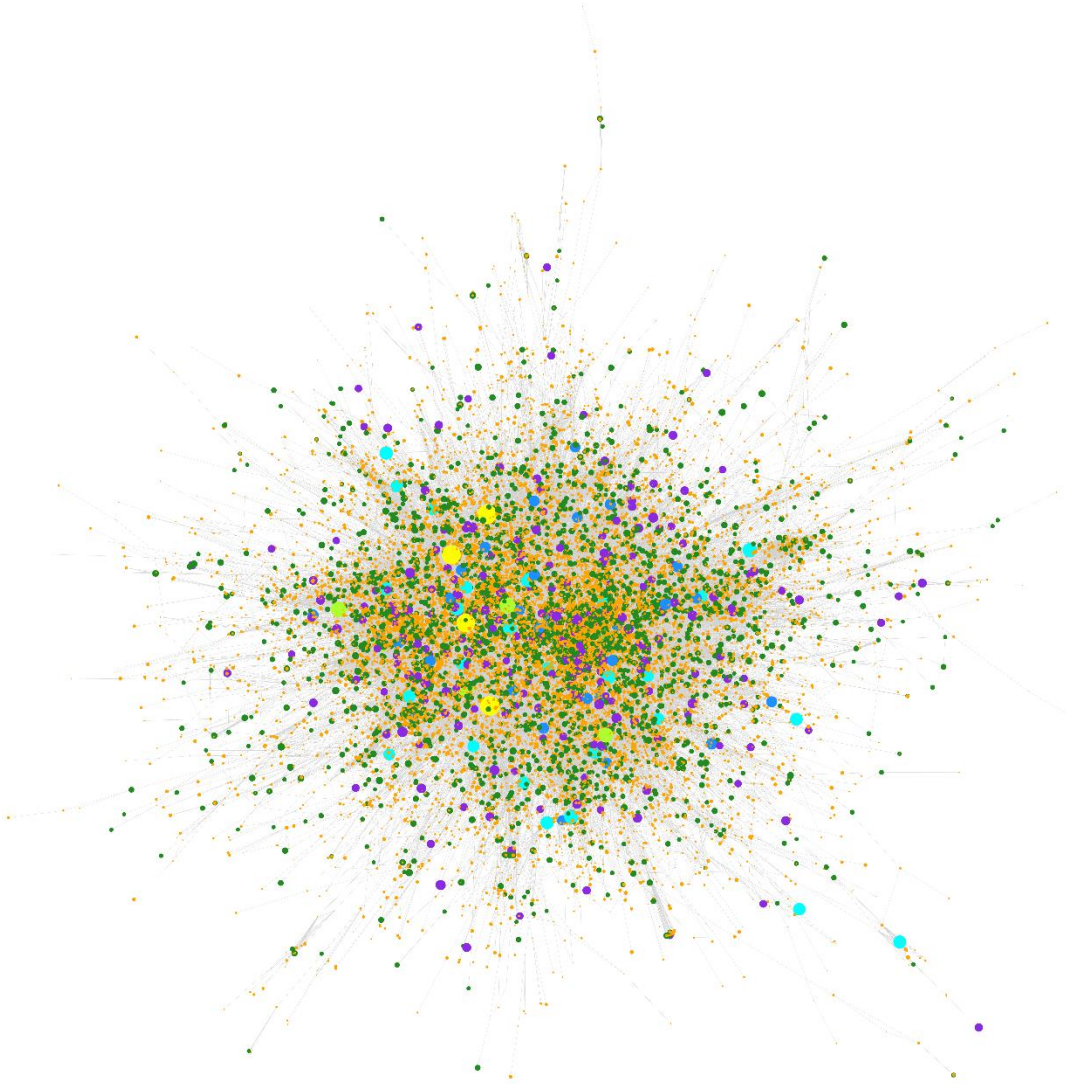
Teacher: MANIU Silviu

Contents

I. Dataset

In this project, I used the social network from the *SNAP – Stanford Network Analysis Project*, *Stanford Large Network Dataset Collection*. The dataset chosen is *Facebook Large Page-Page Network* from the *MUSAE Project*. It is a page-page graph of verified Facebook sites. Nodes represent official pages while the links are mutual links between sites. This dataset is not anonymized, so I find it interesting to have look at what the actual pages are, after the analysis.

Below is an attempt to visualize this dataset. The nodes' colors are chosen randomly, their sizes are directly proportional to their degrees. The links are almost transparent since it is hard to effectively visualize these links.



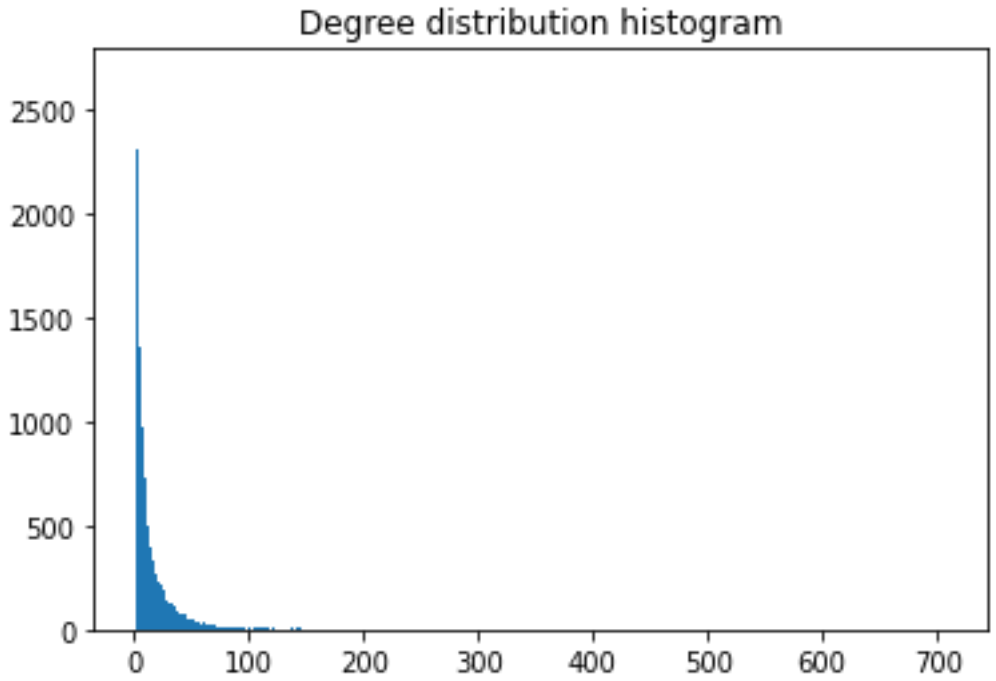
II. Overview

In my analysis, I will use 2 csv files: *musae_facebook_edges* and *musae_facebook_target*. The edges are contained in the former and the node information are contained in the latter. I also included a non-social network for comparison at the last section.

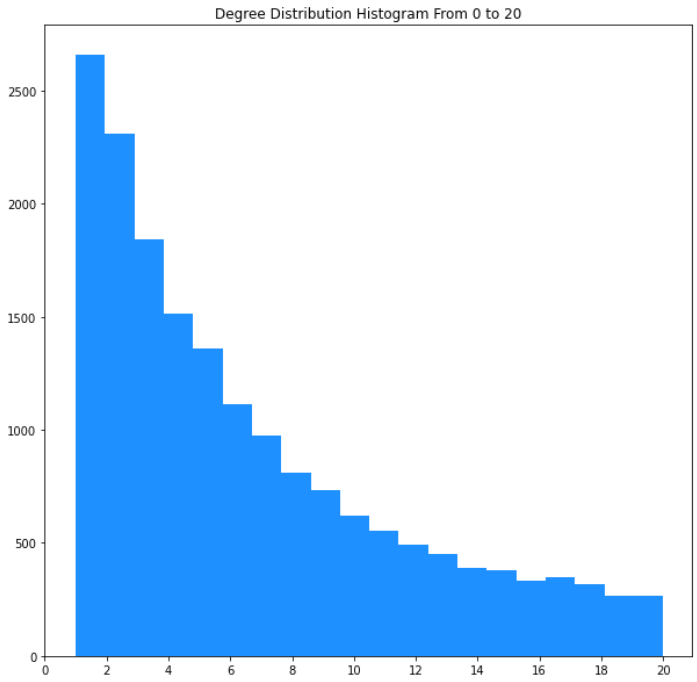
This dataset contains: 171002 edges and 22470 nodes. This dataset is also undirected, which makes the real number of edges become two times bigger – 342004 edges.

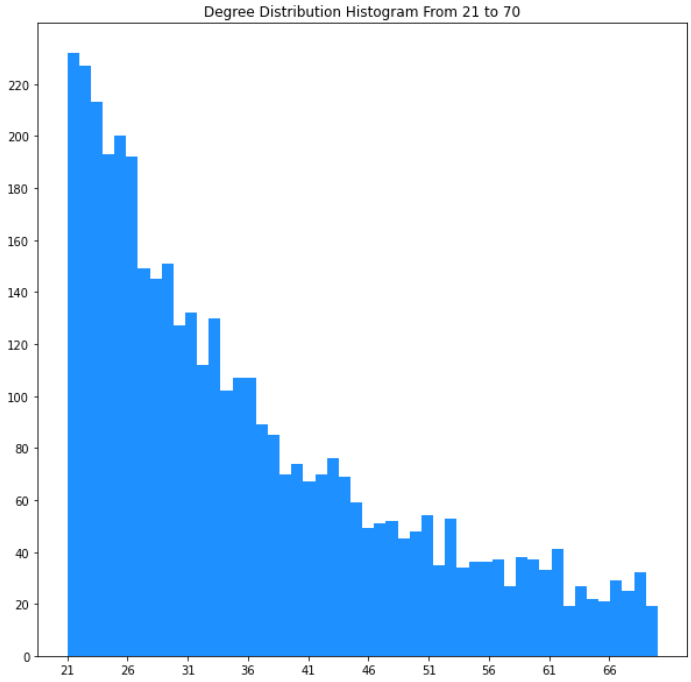
1. Degree

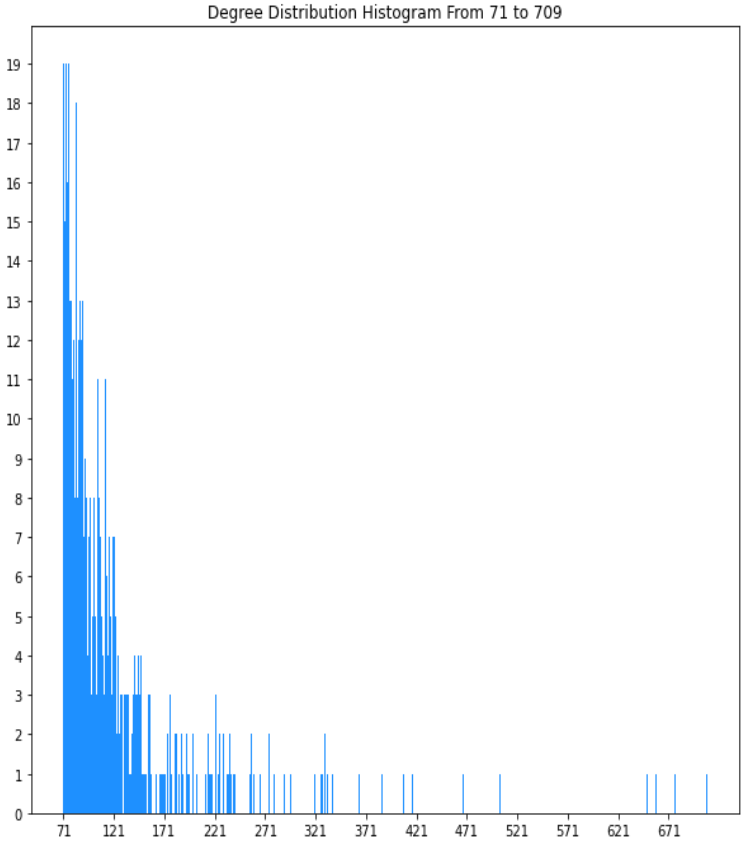
The degrees of this graph range from 1 (2658 nodes) to 709 (1 node). The average degree of each node is 15.22. This shows that this social network is sparse, considering that each node can connect 22469 other nodes in 2 directions: in and out, at maximum.

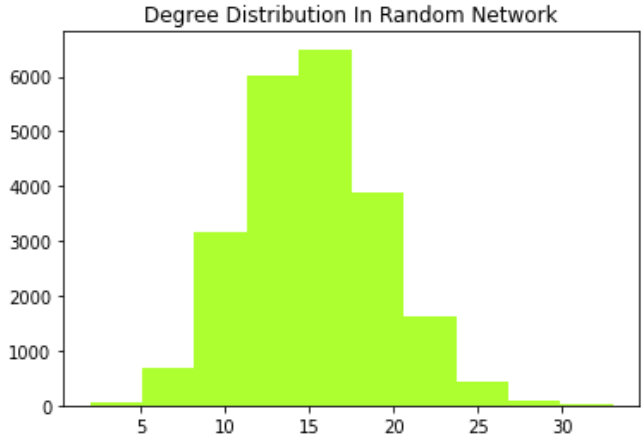


The degree distribution for this graph is almost linear – as the degree increases, the number of nodes decreases. However, this graph provides us with too little information after the 100-degree mark. As a result, I made 3 separate histograms for different ranges of degrees as well as a histogram for random network of the same average degree.









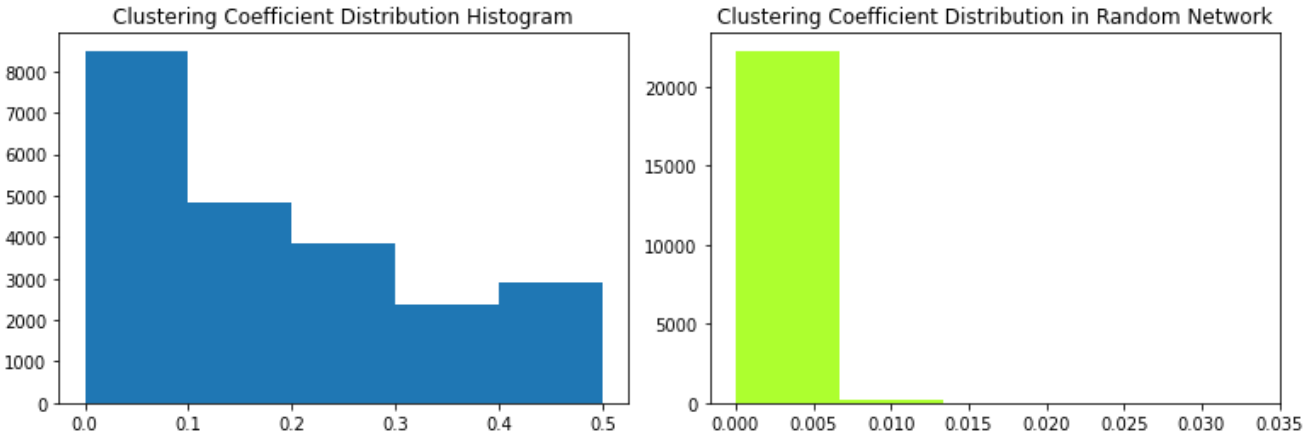
We can easily see that this Facebook pages network has many hubs whose degrees exceed the average by a large margin.

This behavior is completely different from a random network (quite balanced distribution, and less significant outliers) and is more similar to that of a scale-free network.

2. Clustering Coefficient

The clustering coefficients of this Facebook page graph range from 0 to 0.5. Since we have many nodes with a single link in this graph, the number of nodes with clustering coefficient of 0 is expected to be high.

This social network has an average clustering coefficient of 0.362. This shows us a high community structure. The neighbors of each node tend to connect to each other.

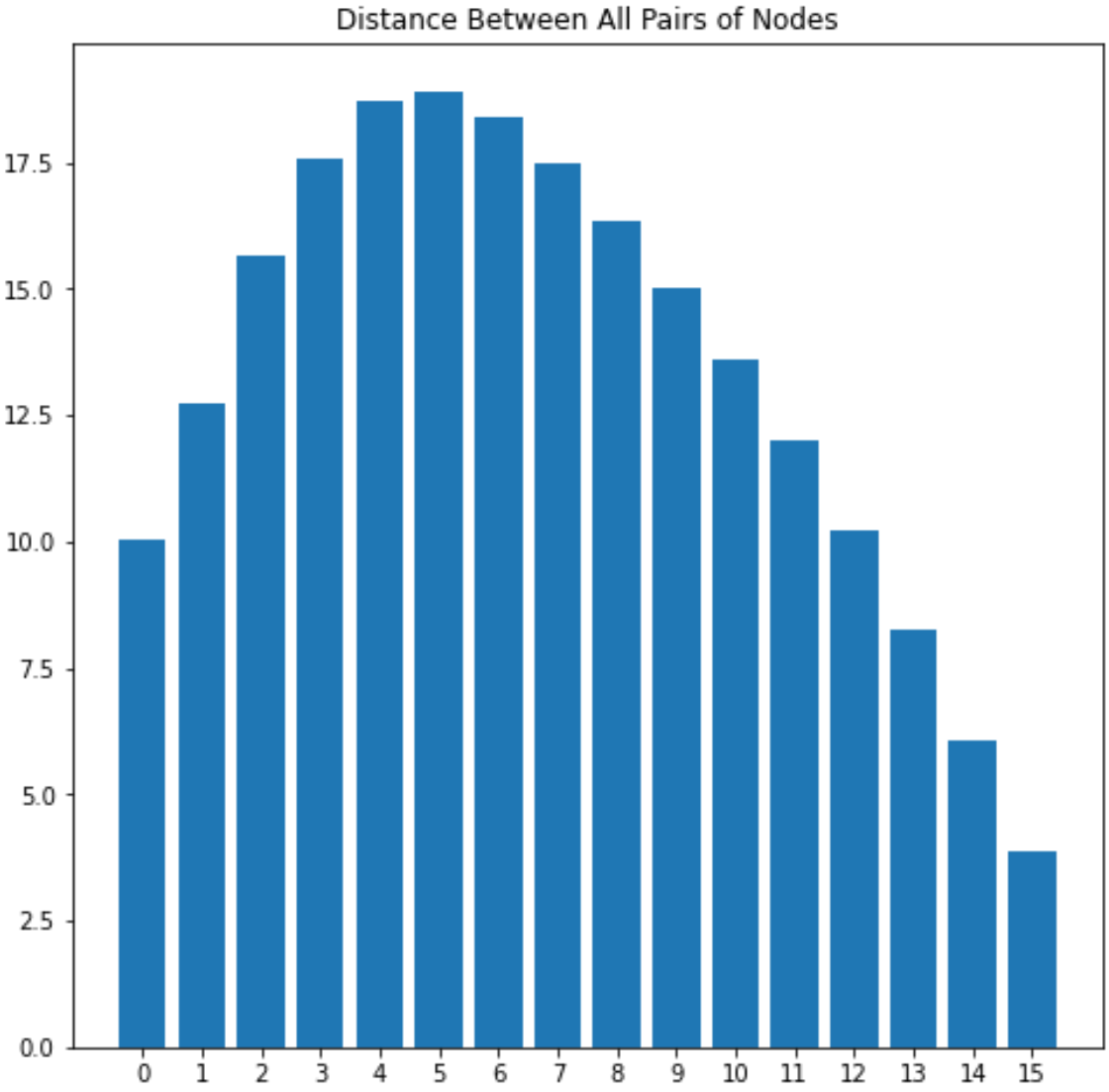


Unlike this network, an instance of a random network of the same characteristics, which I generated, only has an average value of 0.00062. This is also its probability of having a link between 2 nodes.

This difference can be explained by the fact that the nodes in a social network do not randomly have a link with the others. So, there is high chance that their neighbors are directly connected. Moreover, a random network with 22470 nodes and 342004 links is a highly sparse network, hence the low average clustering coefficient.

3. Distance

The average distance between every pair of nodes is 4.97. Although we have large numbers of vertices and edges, the average distance is low. The diameter of this graph is 15. Below is the distance distribution of all pairs of nodes in natural logarithm scale.

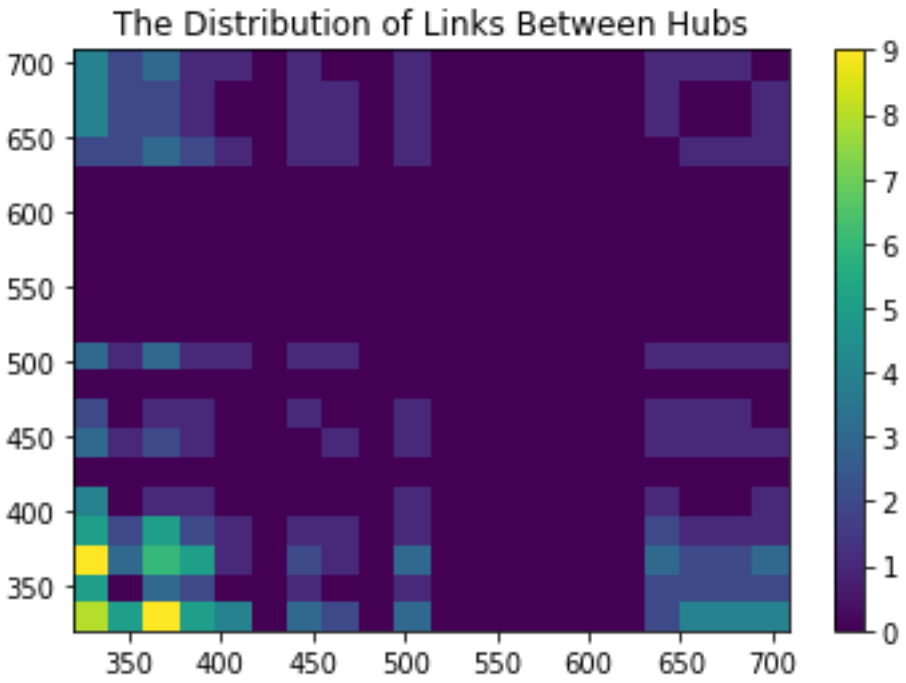


The distance values gather around the average distance (with the occurrences of distance of 4, 5 and 6 are the highest). Meanwhile the occurrences of bigger or smaller distances decreases sharply compare to those. Since the Facebook pages social network has a structure with hubs and slightly high clustering coefficient, it contains distance values far away from the average (both higher and lower values): Hubs can effectively reduce the distance between nodes and the high distance values belong to the pairs of nodes of different clusters.

With the same random network generated in the first sections, I obtain a diameter of 6 and an average distance of 3.95. It suggests that this random network is quite balanced in the distance between all pairs of nodes.

4. Assortativity

The Facebook pages network has an assortativity coefficient of 0.0805, which is close to that of a random network. However, it also suggests slight assortativity of this network. Indeed, the histogram below shows the connection between hubs in this network. The total number obtained is exceedingly high. I have 23 nodes that I consider hubs (Nodes having a degree bigger than 300) and 242 links (121 undirected links).



The random network that I have generated in the previous sections has an assortativity coefficient of -0.000054, which put it in the neutral category.

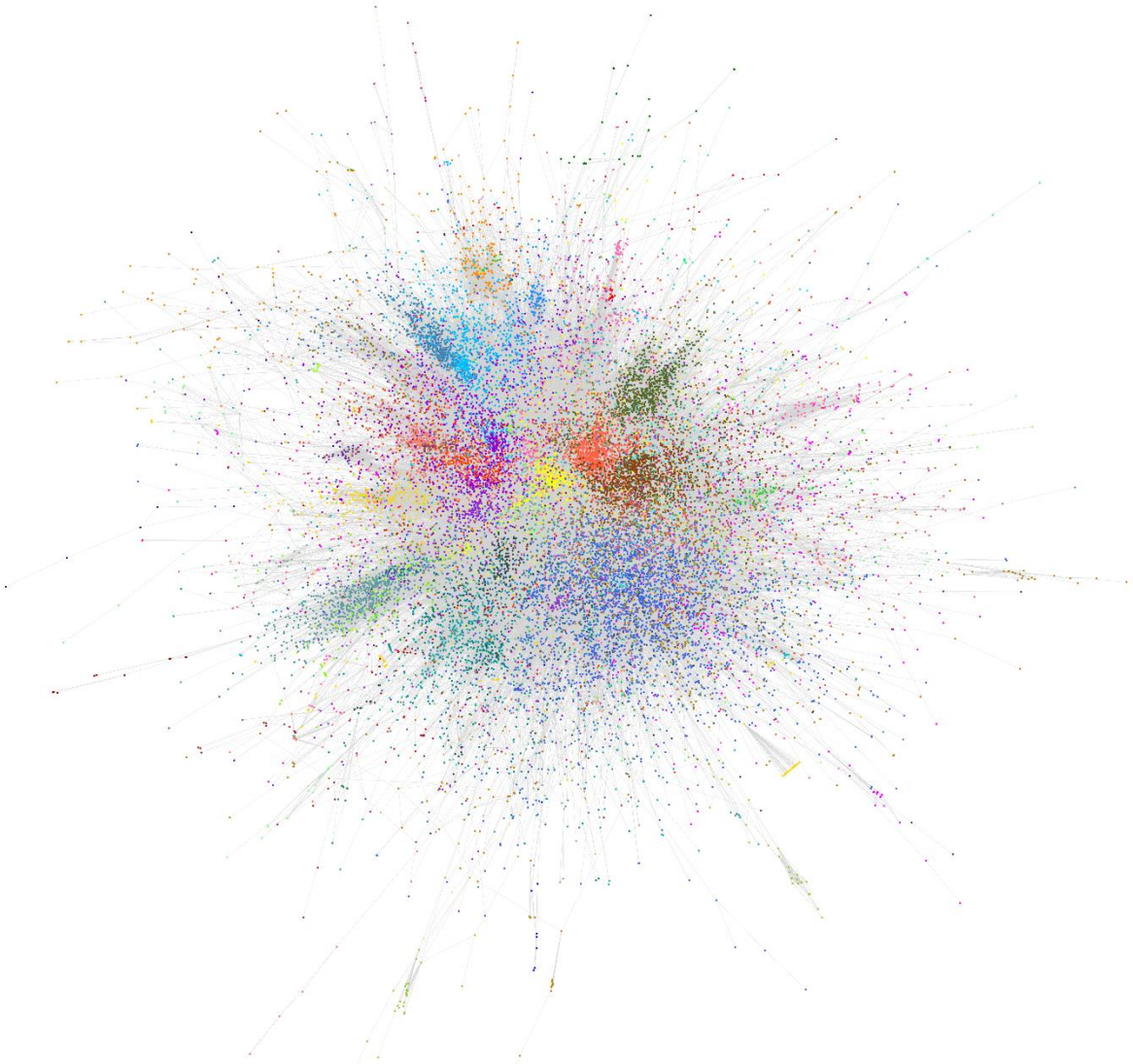
While the difference is not significant, the Facebook pages network suffers from being collected manually, and from being limited in the number of pages. Thus, this network can be heavily biased. With the advantage of having the Facebook pages' name (no data anonymization), I have found several criteria for this low assortativity. For instance, language is a big factor. Indeed, I found in my network small to medium communities formed by pages coming from Brazil or Eastern Europe. Therefore, this assortativity property will be clearer if the dataset grows bigger.

III. In-depth Analysis

1. Community

This Facebook large page network is extremely categorized. Using the Louvain method for communities detection, I found a partition with 58 communities and an extremely high modularity value of 0.814 (Details in the visualization below). With different runs, the network is divided into different partitions, but the modularity is around that point.

Below is a visualization of the communities in the network for this instance. We can see that large communities are formed in this network.

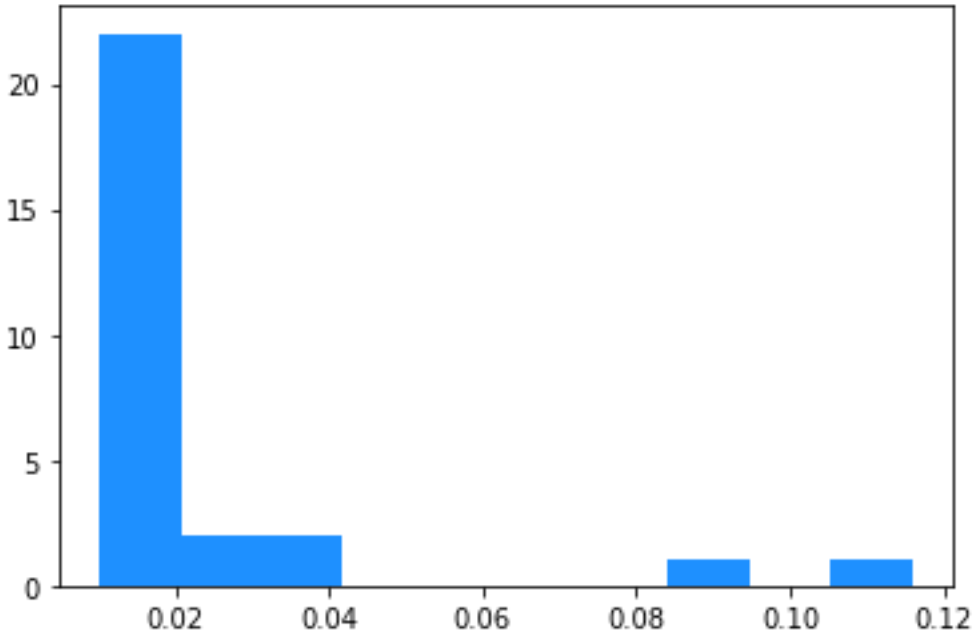


This indicates a high community structure in this kind of network. This also explained the high clustering coefficient in the previous section. And as I stated above, the communities' size varies from several nodes to several thousand nodes because the concerned pages are chosen manually.

For the random network, an instance of this network gives the modularity of 0.193. It does not have such clear community structure like a social network.

2. Betweenness Centrality

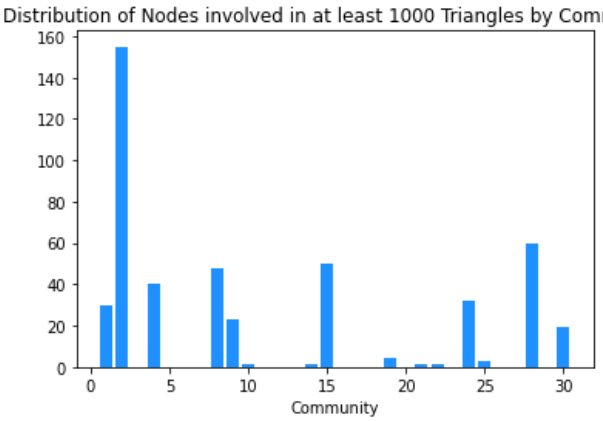
The average betweenness centrality value of the nodes in this network is low. This suggests that the nodes in this Facebook network share close role. We have known in previous section that this network consists of hubs and normal nodes. Therefore, generally, most nodes are not important than the others in the network.

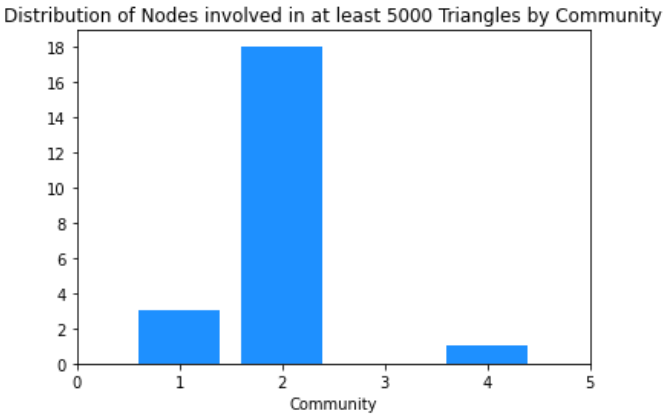


However, I also found several outliers – nodes with extremely high betweenness centrality. The histogram above shows 28 nodes that have betweenness centrality of at least 0.01. For instance, *Facebook* page has a betweenness value of 0.1158 and this value of *Barack Obama* page is 0.0896. In this mutual-like scenario, these high values tell us that these pages are the centers of the Facebook social network. Besides, it may imply that these pages connect pages of different clusters/communities. It also suggests that they might have less influence on tasks such as recommendation of "pages you might like" because most pages tend to have ties with them.

3. Triangles

Another interesting property is the number of triangles that each node is involved in. In average, each node in this network is part of 106 triangles. But there are 468 nodes that are involved in at least 1000 triangles, and the total number triangles (with duplicate) of these 468 nodes takes account for more than 38% of that of the whole network.





Moreover, when I increase the threshold to 5000, I find that most of the nodes with extremely high number of triangles belong to the same community (18 over 22 nodes). This result can be easily explained by the fact that this Facebook network has clear community structure. As a result, the network is dense inside a community – which is an important factor for the number of triangles. Additionally, for a dense community, a greater community also means that there might be more triangles inside.

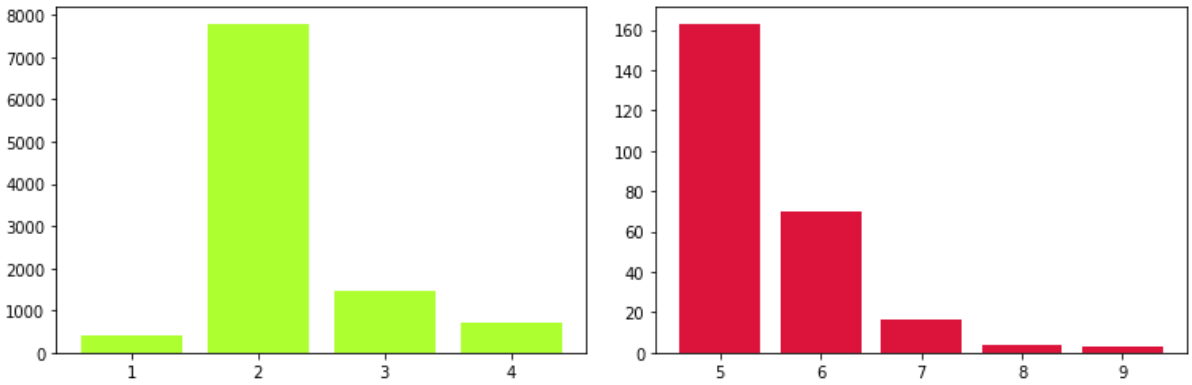
4. Non-social Network

In this section, I will compare the properties of this Facebook network with a non-social network. The dataset chosen is the transportation of Paris. This dataset consists of many smaller datasets. Here, I choose the bus dataset from those. This bus network is extraordinarily sparse.

As this bus network is disconnected, I only take out the biggest connected component to do my analysis. The original network has 10880 nodes and 12547 edges. The component that I take out will have 10644 nodes and 12309 edges (hence 24618 edges for undirected graph), taking 97.8% and 98.1% of the original network, respectively.

	Paris Bus Network	Facebook Network
Average Degree	2.31	15.22
Clustering Coefficient	0.0054	0.362
Average Shortest Path	47.63	4.97
Diameter	159	15
Degree Assortativity	0.0274	0.0805
Modularity	0.932	0.814

There are many differences in these 2 networks. Firstly, I will look into the degree distribution of the Paris bus network. Below is the degree distribution of this network.

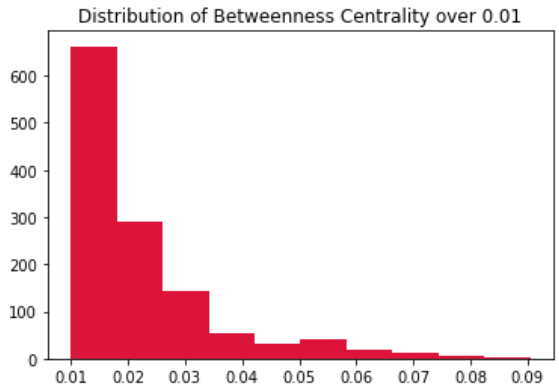


We know that this network is extraordinarily sparse, so the average degree is expected to be low as well. The highest degree in this network is only 9 so we can assume that no hub exists. Nodes with degree of 2 take the majority of this network because in fact, in a line of bus, all the nodes have degree of 2 except the 2 terminuses. The nodes with higher degree are intersection of lines. The degree assortativity is not significant in this network.

Because we do not have many hubs, the distance between nodes is much longer for this network. There are many factors that affect the growth of this network (e.g., geography, population), so we cannot just establish new lines or modify the already existing lines. This explains the exceedingly long average distance as well as the high diameter.

Due to the nature of a line of bus, there will be hardly any station that has their neighbors connected. This is also explained by the degree distribution with high occurrence of 2. Therefore, the clustering coefficient is low.

The best partition for this network using the Louvain method only divides it into 53 communities with the largest takes only 4% of the total number of nodes. As the bus network consists of many smaller lines of bus, it is perfect to form a community structure, a community only connect to another community by very few links (preferably at intersection nodes). This is the reason for the high modularity.



Here is the distribution of 1258 nodes with betweenness centrality over 0.01. Apart from the intersections that have high betweenness centrality, the nodes that start an isolated line are also be included in many shortest paths. Thus, we have many nodes with degree 2, 3 and 4. For instance, the top 3 are LA DÉFENSE-METRO-RER-TRAMWAY, GLACIÈRE – TOLBIAC, VAN GOGH (4, 6, 6).

IV. Conclusion

Facebook pages mutual like network is a clear example of how a social network behaves. Unlike non-social network and random network, social network grows freely with strong community structure. Social networks revolve around ultra large hubs to form community with high density inside that community. Besides, as its name suggests, social network structure is perfect for socialization between the nodes of that network.

Bibliography

Kujala, R., Weckström, C., Darst, R., Mladenović, M., & Saramäki, J. (2018). A collection of public transport network data sets for 25 cities. *Sci Data*.

Leskovec, J., & Krevl, A. (2014, June). *Stanford Large Network Dataset Collection.* Retrieved from http://snap.stanford.edu/data

Sarkar, R., Rozemberczki, B., & Allen, C. (2019). Multi-scale Attributed Node Embedding.

Appendix

1. Environment

All the code and visualization in my project were done on Google Colab.

I used the standard settings of Colab for this project, so the execution time might be long for high complexity algorithms and the RAM is not sufficient in one run for certain ones.

2. Communities by Louvain Method

This is only the complement information for the visualization of the network in the community section, colors were chosen randomly from the list of CSS colors in matplotlib.

ID	Cardinality	Color
0	546	
1	704	
2	1122	
3	414	
4	1465	
5	224	
6	190	
7	153	
8	1216	
9	815	
10	1545	
11	171	
12	234	
13	429	
14	3358	
15	1278	
16	201	
17	349	
18	98	
19	526	
20	949	
21	1001	
22	567	
23	477	
24	952	
25	544	
26	329	
27	194	
28	145	

ID	Cardinality	Color
29	187	
30	261	
31	271	
32	148	
33	49	
34	271	
35	110	
36	103	
37	73	
38	202	
39	88	
40	34	
41	8	
42	8	
43	111	
44	23	
45	14	
46	37	
47	9	
48	108	
49	21	
50	23	
51	7	
52	9	
53	10	
54	55	
55	14	
56	6	
57	14	