MASTER 2 DATA SCIENCE

Cours : Social and Graph Data Management.

Établissement : UNIVERSITE PARIS SACLAY

Enseignant : SILVIU MANIU

Sujet : Network Analysis Project - 27/11/2020

Etudiant : Panongbene Jean Mouhamed Sawadogo.

SOMMAIRE :

I. INTRODUCTION

In this work, we analyzed the properties of a social network dataset. For this, we use the python language with the Jupyter tool. In this folder you will find the Jupyter file which has the name network_analysis.ipynb (this file contains the python code to analyze the network), you will find the README.rd file which explains how the code works and other explanations for the specifics of our code. The Project.pdf pdf file which explains the work we need to do. In the data directory, you will find the wiki-Vote.txt file which contains the data representing our graph. You will also find this report which explains in detail our work as a whole.

II. DATASET

The graph considered here is a directed graph represented by the wiki-Vote.txt text file contained in the data directory. This file contains multiple rows and two columns. In each line, we find two integers which represent two nodes of our graph : indeed the nodes of our graph are represented by integers and the two integers on the same line represent a link between the two nodes represented by the two integers. Our graph is unweighted (all links have the same weight equal to 1). Our graph contains 7115 nodes and 103,689 oriented links.

The data used to construct this graph was downloaded from the website website. A citation to write the data this graph represents was given on the website as follows :

Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. Using the latest complete dump of Wikipedia page edit history (from January 3 2008) we extracted all administrator elections and vote history data. This gave us 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on). Out of these 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users.

The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node i to node j represents that user i voted on user j.

III. GRAPH ANALYSIS

1. Show the number of nodes and edges in the graph.

   Answer : Our graph contains 7115 nodes and 100762 edges.

2. Draw the graph if small enough ; for large graphs this may be unfeasible.

   Answer :

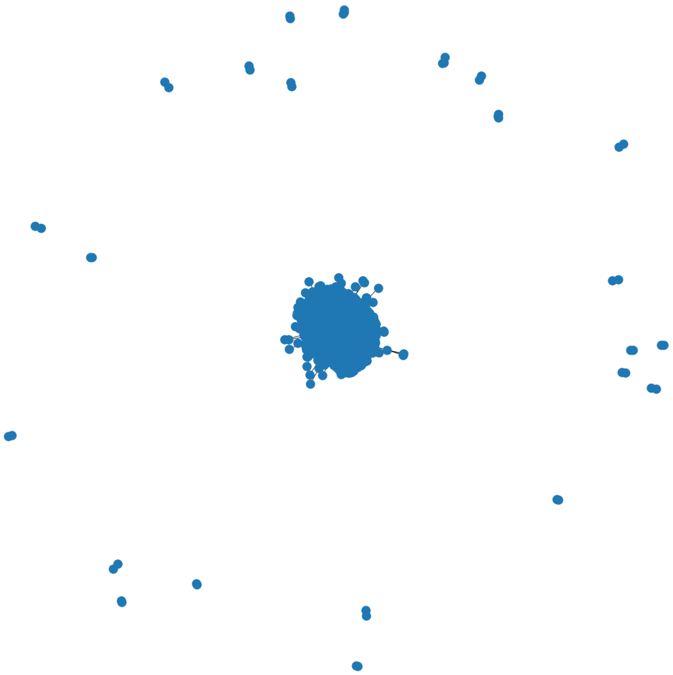


FIGURE 1 – graph wiki-vote

   The very large number of nodes that we have prevents us from having a good visual representation of our graph. However, the representation below allows us to see that our graph is not connected and we can distinguish by looking at this figure that it is composed of a large subgraph and several small graphs.

3. Draw the histogram of degrees. Compare the distribution with the distribution for a random graph having the same average degree. Discuss the results.

Answer : Average degree of wiki-vote graph = 28.323822909346458

The average degree <k> of a random graph G where the node probability exists is p, which have N node is given by

$$< k >= p(N-1) \implies p = \frac{< k >}{N-1}$$

So, we have : p = 0.0039808605635061785.

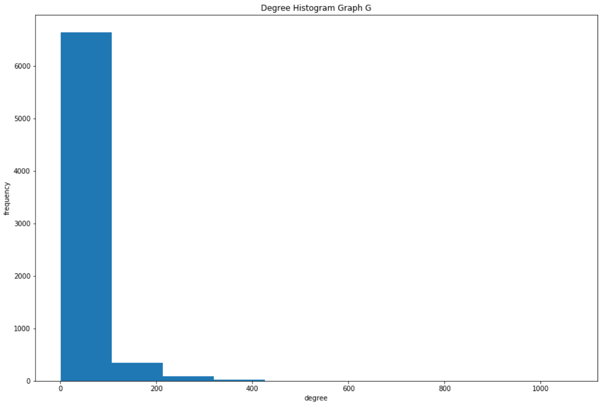


FIGURE 2 – histogram of degrees wiki-vote

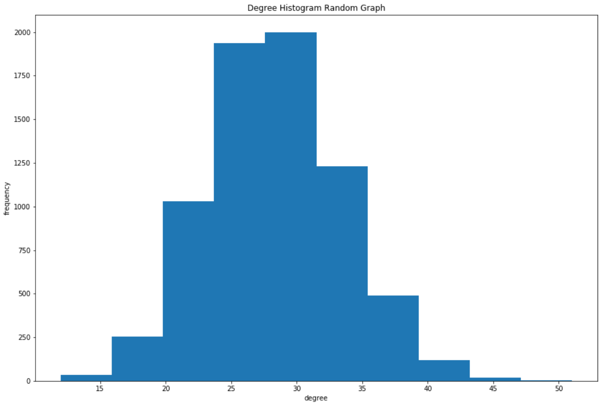


FIGURE 3 – histogram of degrees random graph

We see that for a randomly generated graph with the same number of nodes and the same average degree, the randomly generated graph has a distribution resembling the distribution of a Gaussian law, but the distribution followed by the wiki-vote graph is decreasing as a function of the degree of nodes.
We can say that the distribution of the wiki-vote graph does not respect the central limit theorem and therefore, the votes are not independent.

4. Draw the histogram of clustering coefficient, and the average clustering coefficient. Compare it with the one of a random graph and discuss the results.

   Answer : The average clustering coefficient of wiki-vote graph = 0.14089784589308738

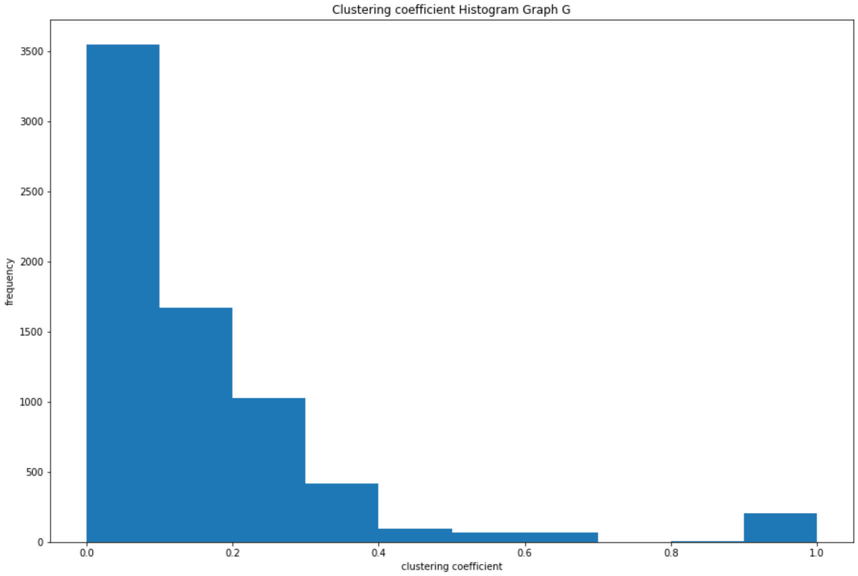


FIGURE 4 – histogram of clustering coefficient wiki-vote

The average clustering coefficient of random graph G = 0.003965959389834505

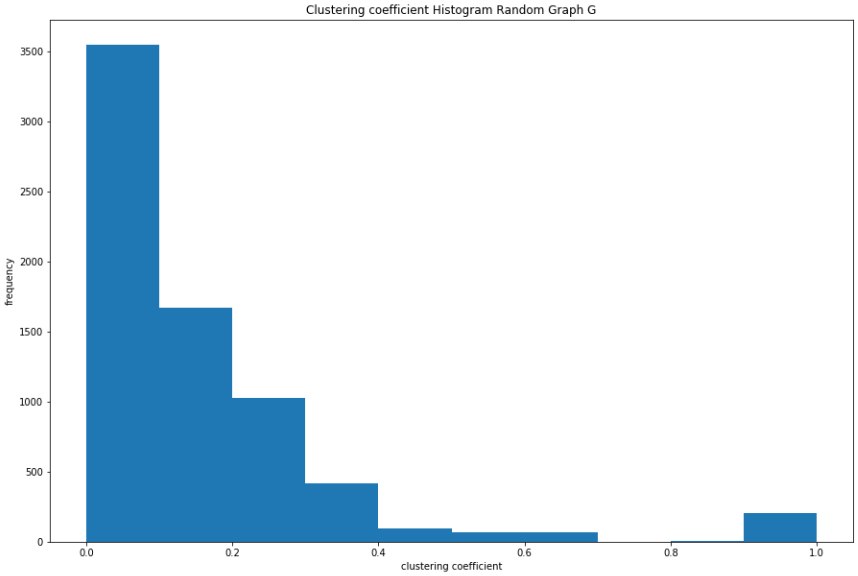


FIGURE 5 – histogram of clustering coefficient random graph

Using the same random graph generated previously which has the same number of nodes and the same average degree, we find that the distribution of the clustering coefficients are the same for the two graphs but the average clustering coefficients are totally different.

We can empirically deduce that the distribution of the clustering coefficient depends on the average degrees

5. Draw the histogram of distances in the graphs, the diameter and the average distance. Compare with random graphs and discuss the results.

   Answer : Our graph is not connected but, we can see that the subgraph connect composing our graph consists of a large connected graph G_sub_big of 7066 nodes 23 subgraph connects of less than 3 nodes.

So we consider in this part that the diameter of the graph G is equal to the Diameter of the connected subgraph G_sub_big.

All the analysis that will be done in this part will be done by considering an approximation of the graph G as equal to the graph G_sub_big

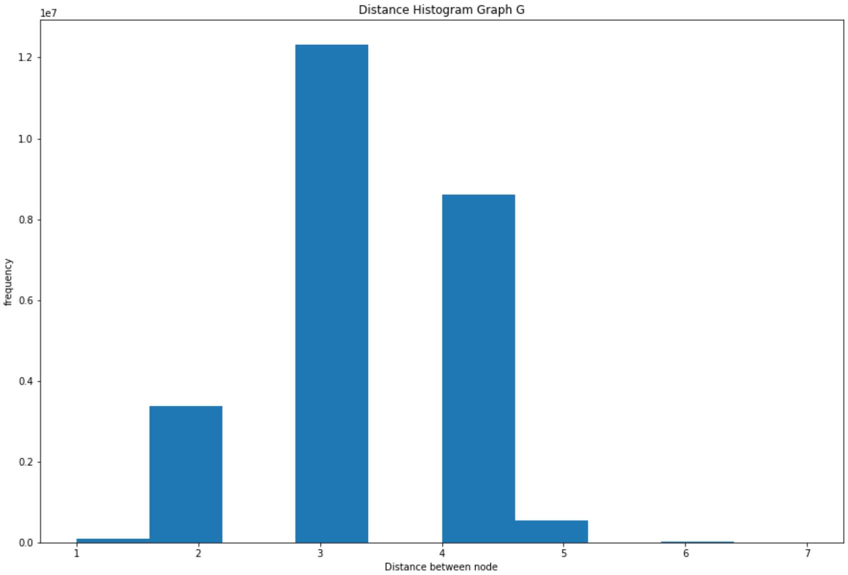


FIGURE 6 – histogram of distance wiki-vote graph

The Diametre of wiki-vote graph G = 7
The Average distance of wiki-vote graph G = 1.6014653042282387

We have generated a random graph that has the same number of nodes and the same average degree as the wiki vote graph. However, the random graph is totally connected unlike the random graph.

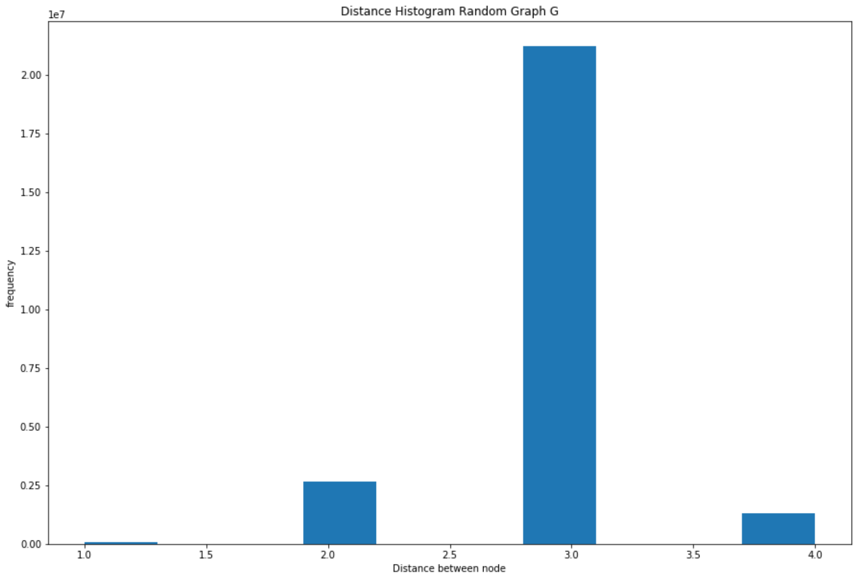


FIGURE 7 – histogram of distance random graph

The Diametre of random graph G = 4
The Average distance of random graph G = 1.4695393423161125

We see that the diameter of the wiki-vote graph is 7 while for the random graph, the diameter is 4 more the average distances are different for our graphs : it is greater than 1 for the wiki-vote graph and less than 1 for the random graph.

However, we see for the two graphs, the number of nodes with a distance equal to 3 is much more numerous. Moreover, the distribution of distances looks like a curve of the law of gaussiance centered in 3.We can conclude that the distance distributions of our two graphs are the same.

6. Analyze the degree correlations of the graph.

   Answer : We calculated the degree of correlation of the wiki-vote graph and we found r = -0.08305248270016433 ; So we have r<0, then we can deduce that wiki-vote graph is an assortative networks

IV. SECOND PART

The extra requirements are to go beyond basic analysis, and discuss other relevant measures. Below are some suggestions, but you can add your own :

1. Detect the communities in the graph, and discuss the results.

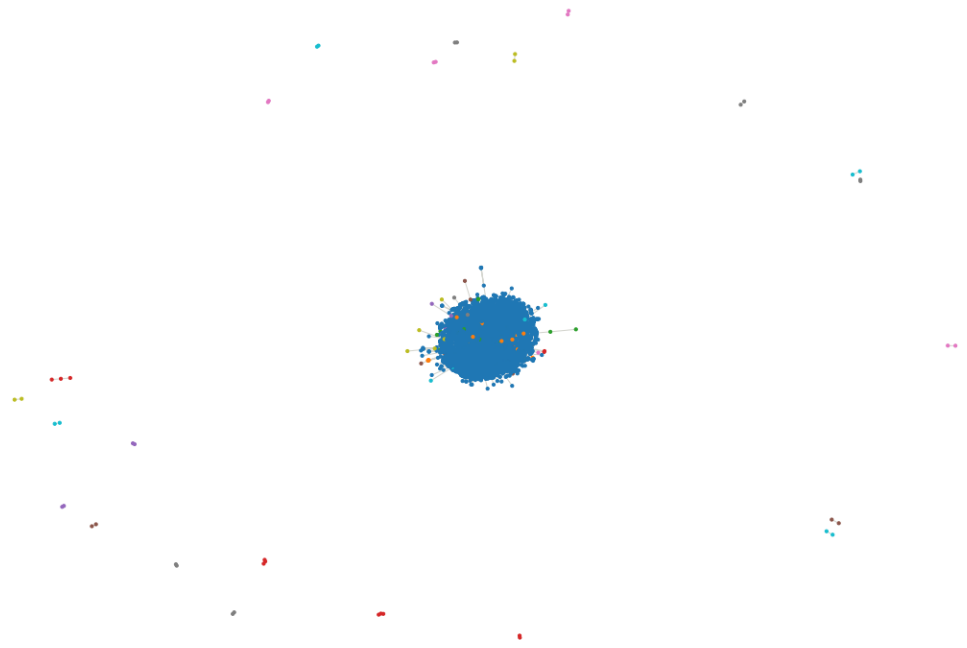


FIGURE 8 – Communities obtained using greedy algorithms

Using greedy optimization to detect communities, we get 55 communities with total modularity equal to 0.34020342296217737.

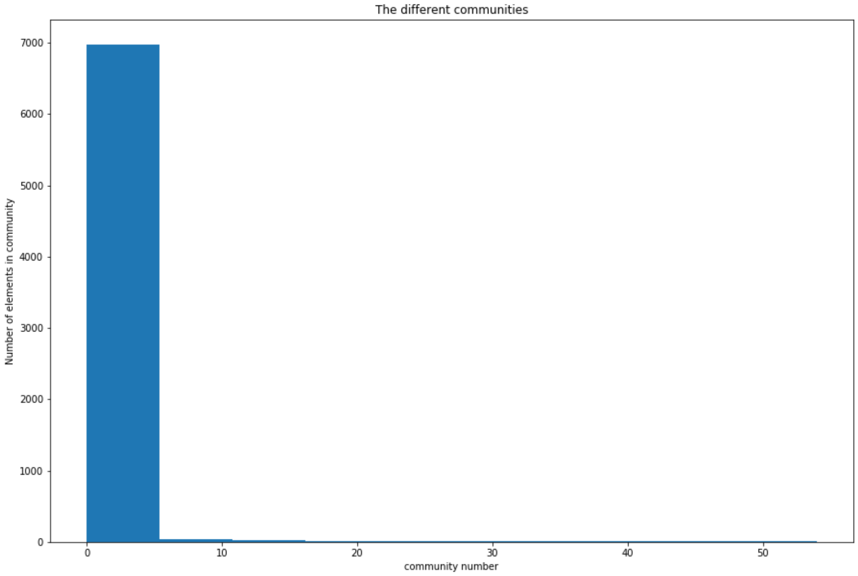


FIGURE 9 – The numbers of nodes per community

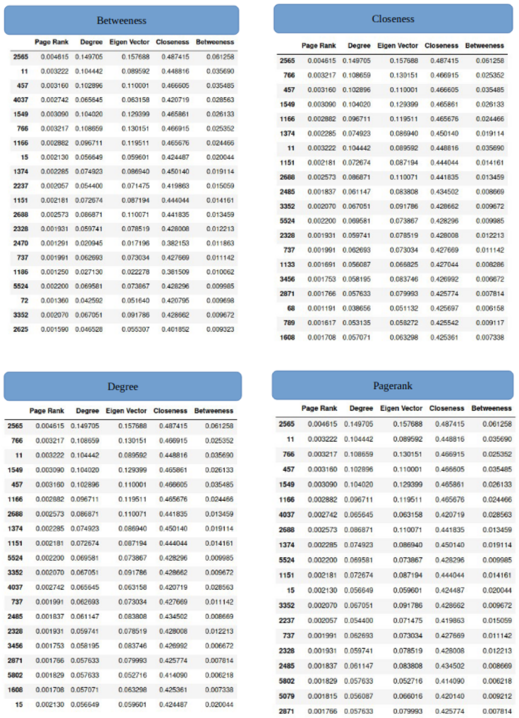
We can also see in the figure above that we have a large community with almost 7,000 nodes and 54 small communities with at least 10 nodes.We can deduce from this that the results of votes on wiki are biased by a homogeneous community that votes among themselves.

2. Count the number the triangles in the graph, and compare to a random graph.

   Answer : We found in wiki-vote graph 608389 triangles while in the random graph with the same number of nodes we found 3884 triangle. We can explain this by the fact that in our graph the links within a community are strong.

3. Compute and discuss other centrality measures : betweenness, PageRank, etc..

By calculating the centrality of "Page Rank", "Degree", "Eigenvector", "Closeness" and "Betweenness", we get the arrays below.



Betweeness

	Page Rank	Degree	Eigen Vector	Closeness	Betweeness
2565	0.004615	0.149705	0.157688	0.487415	0.061258
11	0.003222	0.104442	0.089592	0.448816	0.035690
457	0.003160	0.102896	0.110001	0.466605	0.035485
4037	0.002742	0.065645	0.063158	0.420719	0.028563
1549	0.003090	0.104020	0.129399	0.465861	0.026133
766	0.003217	0.108659	0.130151	0.466915	0.025352
1166	0.002882	0.096711	0.119511	0.465676	0.024466
15	0.002130	0.056649	0.059601	0.424487	0.020044
1374	0.002285	0.074923	0.086940	0.450140	0.019114
2237	0.002057	0.054400	0.071475	0.419863	0.015059
1151	0.002181	0.072674	0.087194	0.444044	0.014161
2688	0.002573	0.086871	0.110071	0.441835	0.013459
2328	0.001931	0.059741	0.078519	0.428008	0.012213
2470	0.001291	0.020945	0.017196	0.382153	0.011863
737	0.001991	0.062693	0.073034	0.427669	0.011142
1186	0.001250	0.027130	0.022278	0.381509	0.010062
5524	0.002200	0.069581	0.073867	0.428296	0.009985
72	0.001360	0.042592	0.051640	0.420795	0.009698
3352	0.002070	0.067051	0.091786	0.428662	0.009672
2625	0.001590	0.046528	0.055307	0.401852	0.009323

Closeness

	Page Rank	Degree	Eigen Vector	Closeness	Betweeness
2565	0.004615	0.149705	0.157688	0.487415	0.061258
766	0.003217	0.108659	0.130151	0.466915	0.025352
457	0.003160	0.102896	0.110001	0.466605	0.035485
1549	0.003090	0.104020	0.129399	0.465861	0.026133
1166	0.002882	0.096711	0.119511	0.465676	0.024466
1374	0.002285	0.074923	0.086940	0.450140	0.019114
11	0.003222	0.104442	0.089592	0.448816	0.035690
1151	0.002181	0.072674	0.087194	0.444044	0.014161
2688	0.002573	0.086871	0.110071	0.441835	0.013459
2485	0.001837	0.061147	0.083808	0.434502	0.008669
3352	0.002070	0.067051	0.091786	0.428662	0.009672
5524	0.002200	0.069581	0.073867	0.428296	0.009985
2328	0.001931	0.059741	0.078519	0.428008	0.012213
737	0.001991	0.062693	0.073034	0.427669	0.011142
1133	0.001691	0.056087	0.066825	0.427044	0.008286
3456	0.001753	0.058195	0.083746	0.426992	0.006672
2871	0.001766	0.057633	0.079993	0.425774	0.007814
68	0.001191	0.038656	0.051132	0.425697	0.006158
789	0.001617	0.053135	0.058272	0.425542	0.009117
1608	0.001708	0.057071	0.063298	0.425361	0.007338

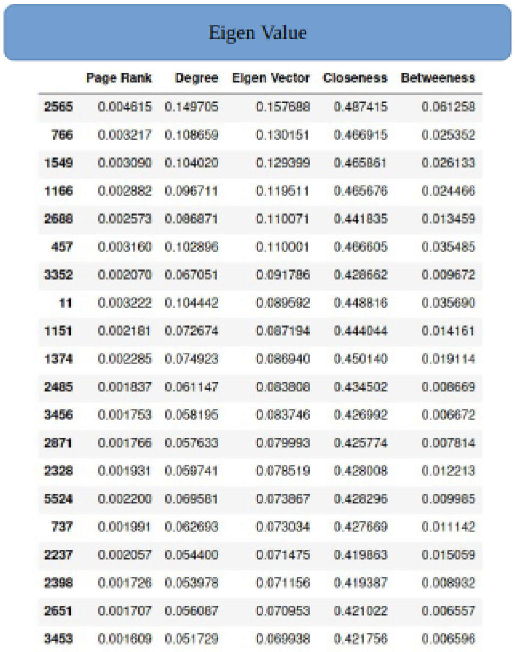
Degree

	Page Rank	Degree	Eigen Vector	Closeness	Betweeness
2565	0.004615	0.149705	0.157688	0.487415	0.061258
766	0.003217	0.108659	0.130151	0.466915	0.025352
11	0.003222	0.104442	0.089592	0.448816	0.035690
1549	0.003090	0.104020	0.129399	0.465861	0.026133
457	0.003160	0.102896	0.110001	0.466605	0.035485
1166	0.002882	0.096711	0.119511	0.465676	0.024466
2688	0.002573	0.086871	0.110071	0.441835	0.013459
1374	0.002285	0.074923	0.086940	0.450140	0.019114
1151	0.002181	0.072674	0.087194	0.444044	0.014161
5524	0.002200	0.069581	0.073867	0.428296	0.009985
3352	0.002070	0.067051	0.091786	0.428662	0.009672
4037	0.002742	0.065645	0.063158	0.420719	0.028563
737	0.001991	0.062693	0.073034	0.427669	0.011142
2485	0.001837	0.061147	0.083808	0.434502	0.008669
2328	0.001931	0.059741	0.078519	0.428008	0.012213
3456	0.001753	0.058195	0.083746	0.426992	0.006672
2871	0.001766	0.057633	0.079993	0.425774	0.007814
5802	0.001829	0.057633	0.052716	0.414090	0.006218
1608	0.001708	0.057071	0.063298	0.425361	0.007338
15	0.002130	0.056649	0.059601	0.424487	0.020044

Pagerank

	Page Rank	Degree	Eigen Vector	Closeness	Betweeness
2565	0.004615	0.149705	0.157688	0.487415	0.061258
11	0.003222	0.104442	0.089592	0.448816	0.035690
766	0.003217	0.108659	0.130151	0.466915	0.025352
457	0.003160	0.102896	0.110001	0.466605	0.035485
1549	0.003090	0.104020	0.129399	0.465861	0.026133
1166	0.002882	0.096711	0.119511	0.465676	0.024466
4037	0.002742	0.065645	0.063158	0.420719	0.028563
2688	0.002573	0.086871	0.110071	0.441835	0.013459
1374	0.002285	0.074923	0.086940	0.450140	0.019114
5524	0.002200	0.069581	0.073867	0.428296	0.009985
1151	0.002181	0.072674	0.087194	0.444044	0.014161
15	0.002130	0.056649	0.059601	0.424487	0.020044
3352	0.002070	0.067051	0.091786	0.428662	0.009672
2237	0.002057	0.054400	0.071475	0.419863	0.015059
737	0.001991	0.062693	0.073034	0.427669	0.011142
2328	0.001931	0.059741	0.078519	0.428008	0.012213
2485	0.001837	0.061147	0.083808	0.434502	0.008669
5802	0.001829	0.057633	0.052716	0.414090	0.006218
5079	0.001815	0.056087	0.066016	0.420140	0.009212
2871	0.001766	0.057633	0.079993	0.425774	0.007814

FIGURE 10 – Comparison of first 20 centralities by ordering the values in decreasing ways according to the differents algorithms columns



Eigen Value

	Page Rank	Degree	Eigen Vector	Closeness	Betweeneness
2565	0.004615	0.149705	0.157688	0.487415	0.061258
766	0.003217	0.108659	0.130151	0.466915	0.025352
1549	0.003090	0.104020	0.129399	0.465861	0.026133
1166	0.002882	0.096711	0.119511	0.465676	0.024466
2688	0.002573	0.086871	0.110071	0.441835	0.013459
457	0.003160	0.102896	0.110001	0.466605	0.035485
3352	0.002070	0.067051	0.091786	0.428662	0.009672
11	0.003222	0.104442	0.089592	0.448816	0.035690
1151	0.002181	0.072674	0.087194	0.444044	0.014161
1374	0.002285	0.074923	0.086940	0.450140	0.019114
2485	0.001837	0.061147	0.083808	0.434502	0.008669
3456	0.001753	0.058195	0.083746	0.426992	0.006672
2871	0.001766	0.057633	0.079993	0.425774	0.007814
2328	0.001931	0.059741	0.078519	0.428008	0.012213
5524	0.002200	0.069581	0.073867	0.428296	0.009985
737	0.001991	0.062693	0.073034	0.427669	0.011142
2237	0.002057	0.054400	0.071475	0.419863	0.015059
2398	0.001726	0.053978	0.071156	0.419387	0.008932
2651	0.001707	0.056087	0.070953	0.421022	0.006557
3453	0.001609	0.051729	0.069938	0.421756	0.006596

FIGURE 11 – Comparison of first 20 centralities by ordering the values in decreasing ways according to the Eigen values algorithms columns

We note that the values of the centralities of the nodes obtained strongly depend on the algorithm used for the calculations. The calculation of centralities using Degree, PageRank and eigenvector algorithms is quite fast (Degree = 0.01 seconds, PageRank = 4.37 seconds and eigenvector = 0.69 seconds). However the Closeness and Betweeneness algorithms take more time (Closeness = 10 minutes 37.66 seconds and Betweeneness = 11 minutes 2.51 seconds).

We see that for some nodes, the centrality is almost the same for all the algorithms used, for these nodes, we can therefore consider the centrality as equal to the average values of the centralities given by the different algorithms.

4. Do a comparative analysis of your social dataset and a non-social one (e.g., transport, Web).

Answer : Unlike the non-social network, our graph does not present control centers where almost all the nodes will meet. In addition, our graph does not have a tree structure common to non-social graphs.

On the other hand, our graph as for a non-social graph presents a single community which is very large. But unlike a non-social graph, the bonds between individuals in a group are very strong.

On the other hand, our graph as for a non-social graph presents a single community which is very large. But unlike a non-social graph, the bonds between individuals in a group are very strong.

5. Other comparisons or analysis that you may find interesting.

   Answer : From the analysis we performed on the wiki-vote graph, we can say that the votes within the wiki community are quite homogeneous. Indeed for the candidates in the elections, either the people who vote accept in majority the integration of a new member or else they vote in majority the non-adhesion of a new member.