Network Analysis: Assignment 1

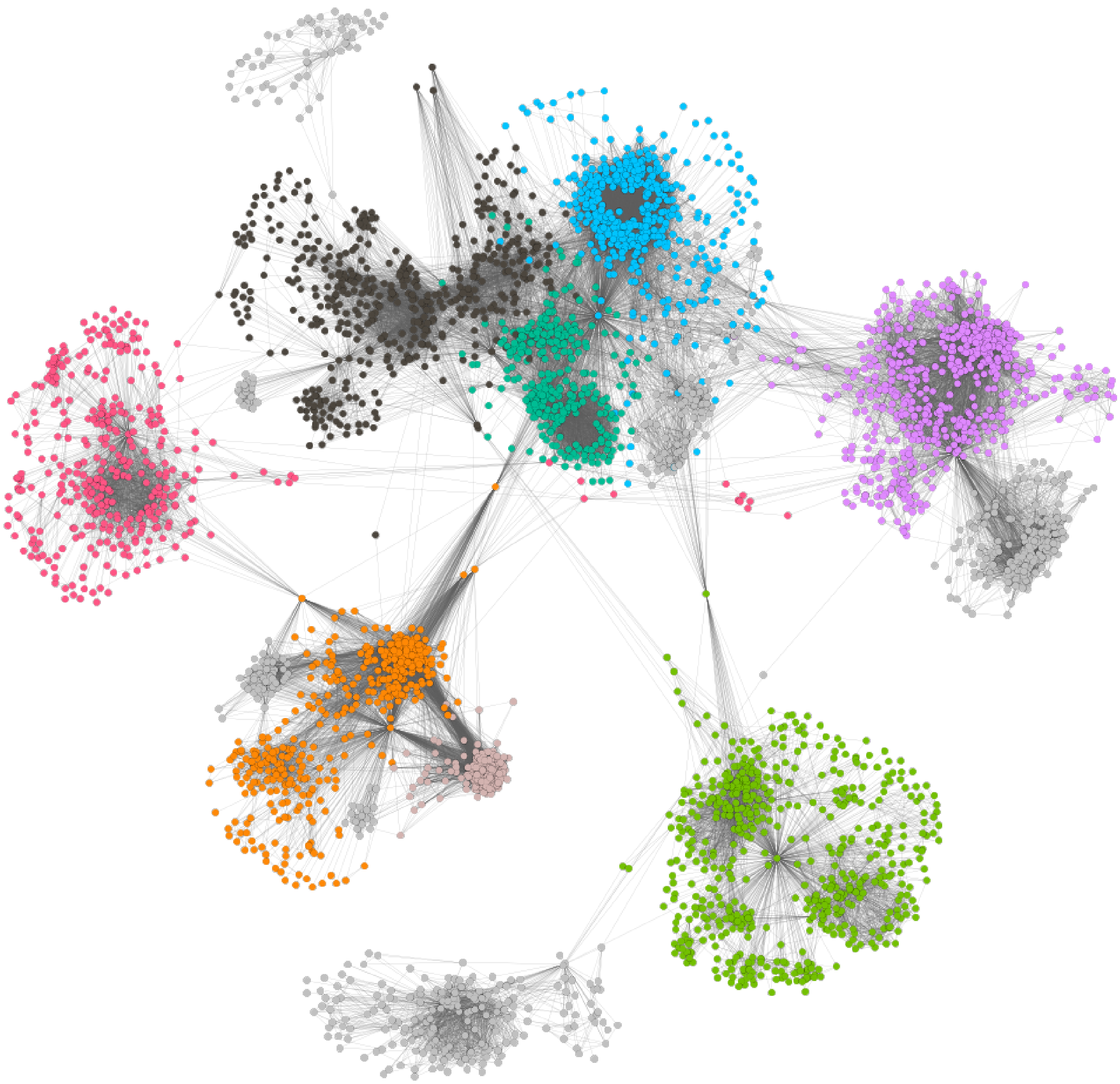
Simone Campisi s4341240

Jacopo Dapueto s4379956

1. Information about the dataset

This dataset consists of 'circles' (or 'friends lists') from Facebook collected from survey participants.
The dataset includes node features (profiles), circles, and ego networks.
It's been downloaded from here.
The linked page provides two different dataset:
An *Ego-centric network* (or *"ego" networks*), is particular type of network in which specifically maps the connections of and from the perspective of a single person (an ego). In fact this networks represent circles of friends of a certains person (ego).

- The first contains 10 ego networks, each one has the edge list, a list of clicles (each one constisting of list of nodes), and other features of the network.
- The second, that is the one used in the assigmentent is a network obtained combing all the ego-networks, including the ego nodes themselves, along with an edge to each of their friends.

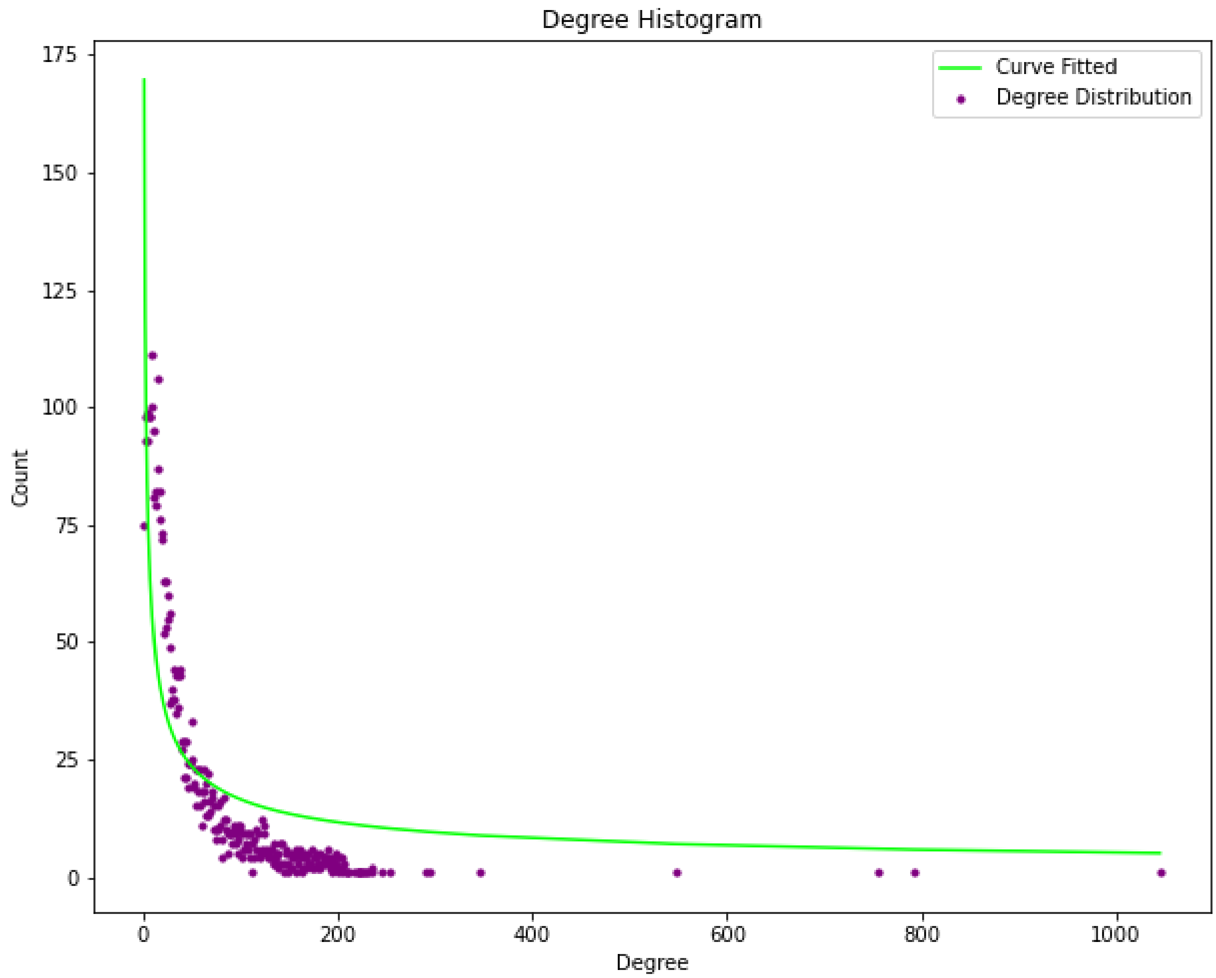


2. Analysis

Dataset statistics	Values
Nodes	4039
Edges	88234
Average Degree	43.691

Dataset statistics	Values
Average clustering	0.605
Global clustering	0.51
Nodes giant component	4039
Diameter	8
Average shortest path	3.69
Density	0.01
Assortativity	

2.1. Does the graph have the same characteristics of a random or a power-law network?



Observing the chart of the degree distribution is possible to notice that the degree distribution follows the trend of a *power law* degree distribution, described by the following equation:

$p_k \sim k^{-\gamma}$

Since the degree distribution follows a power law, the network is called *scale free* network, in which fact there are a fraction of nodes with very high degree, that are the hubs.

2.2. Which are the most important nodes, with respect to a given centrality measure?

We decide to measure the "importance" of the nodes considering the betweenness and the closeness centrality.

Betweenness		**Closeness**	
Node	**Value**	**Node**	**Value**
107	0.480	**107**	0.459
1684	0.337	**58**	0.397
3437	0.236	**428**	0.394
1912	0.229	563	0.393
1085	0.149	**1684**	0.393
0	0.146	171	0.370
698	0.115	348	0.369
567	0.096	483	0.369
58	0.084	414	0.369
428	0.064	376	0.366

The table above shows the top 10 nodes with maximum betweeness and the top 10 nodes with maximum closeness and the nodes present in both ranking are highlighted.

The betweeness measures how many short paths pass to, instead the closeness measures the mean distance from a vertex to the other vertices and therefore the nodes with high closeness can have an easy access to information of influence on other nodes.

As can be seen from the table the node *107* it has the highest betweeness and the higher closeness, so we can say that such node is very important in the network and in the main cluster that is in the center of it.

The closeness measures don't change too much from one node to the other, and this can be a
By constrast the betweeness measures in the ranking dedreases very fast, so there are few nodes which lead the communication between the clusters of the network. In fact it can be seen from the graph above that few nodes connect the cluster to the "center" of the network.

2.3. Are the paths short with respect to the size of the network?

The *shortest path (or geodesic path)* between two nodes in a network the number of links. The length of a shortest path is called *shortest distance* . The *average shortest path* , instead, is the average of the shortest paths between all the pairs nodes, and in the network we are considering is about 3.69. So, is possible to say that in this network there is a **small-world effect** because the average shortest path is surprisingly short when compared with the numkber of nodes of the network, that are 4039.

2.4. Is the network dense?

The **density** of a network is the defined by the following formula:

$$\rho = \frac{L}{\frac{1}{2}N(N-1)}$$

, in which L is the total number of links and N is the total number of nodes, represents the total number of edges on the total possible edges, and in this case this values is very small, 0.01, this means that the network is **sparse**.
2.5. Is the network assortative?

2.6. Average clustering

The *average clustering* is the average of all the *clustering coefficients*, defined as

$$C_i = \frac{k_i}{k_i(k_i-1)}$$

, in which
L_i
is the number of links between the
k_i
neighbors. This value captures the degree to which the neighbors of a node link to each other, and it is in the range [0,1]. In this case, the mean value in the network is 0.605 and this means that there is in average a 60.5% of probability that a node randomly selected has two neighbors linked.