Network Analysis: Assignment 1

Simone Campisi s4341240

Jacopo Dapueto s4345255

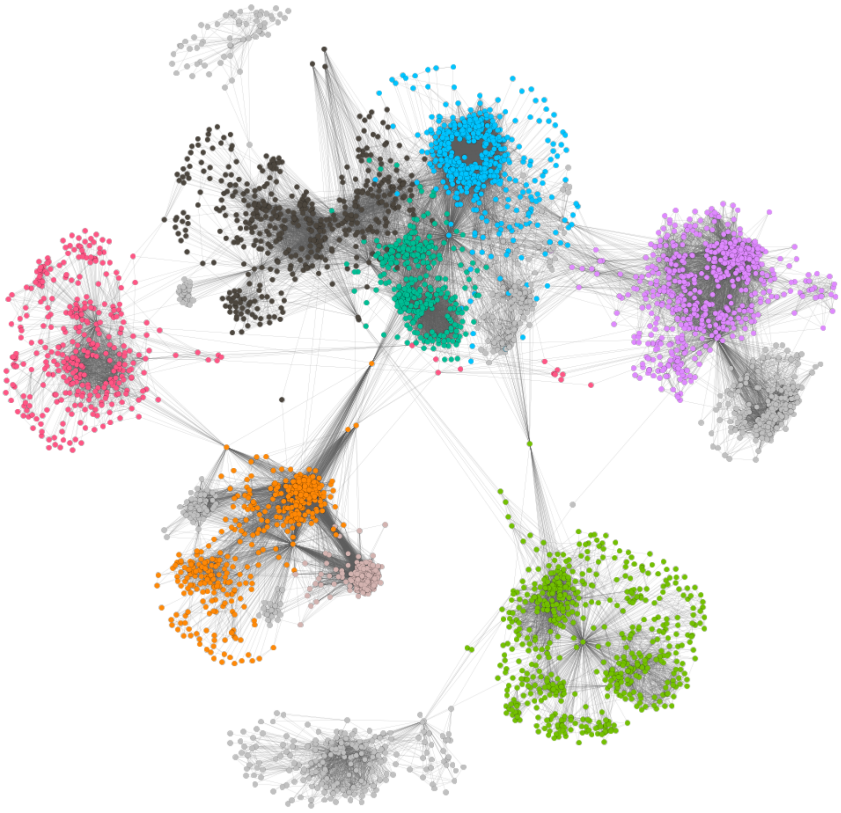
1. Information about the dataset

The dataset we choose 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 Stanford Network Analysis Project. The linked page provides two different dataset:

- The first contains 10 ego networks, each one has the edge list, a list of circles (each one consisting of list of nodes), and other features of the network. An *Ego-centric network* (or "*ego*" *networks*)consist of a focal node ("ego") and the nodes to whom ego is directly connected to (these are called "alters"). In fact this networks represent circles of friends of a certain person (ego).
- The second is the one used in the assignment and it is obtained combining all the ego-networks, including the ego nodes themselves along with an edge to each of their friends.

To visualize the network we used **Gephi** which provides functionalities to process networks, especially for very large networks.

One of them identifies the *communities* through the *modularity* measure and to each of them is assigned a different color. In the image below there are both densely connected communities and sparsely connected ones. So this is a good way to represent the ego network, because is possible to see better the circles of friends.



2. Analysis

The following table shows the global statistics we used to analyze the network in the following chapters, it includes the number of nodes and edges, the average clustering and the global clustering etc. The entire network is connected so exist a path for each pair of nodes.

Dataset statistics	Values
Nodes	4039
Edges	88234
Average Degree	43.691
Average clustering	0.605
Global clustering	0.51
Nodes giant component	4039
Diameter	8
Average shortest path	3.69
Density	0.01
Assortativity	0.06

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

To answer to the question we first visualize the degree distribution together with the fitted curve. Each dot represent a degree and the frequency it appears in the dataset.

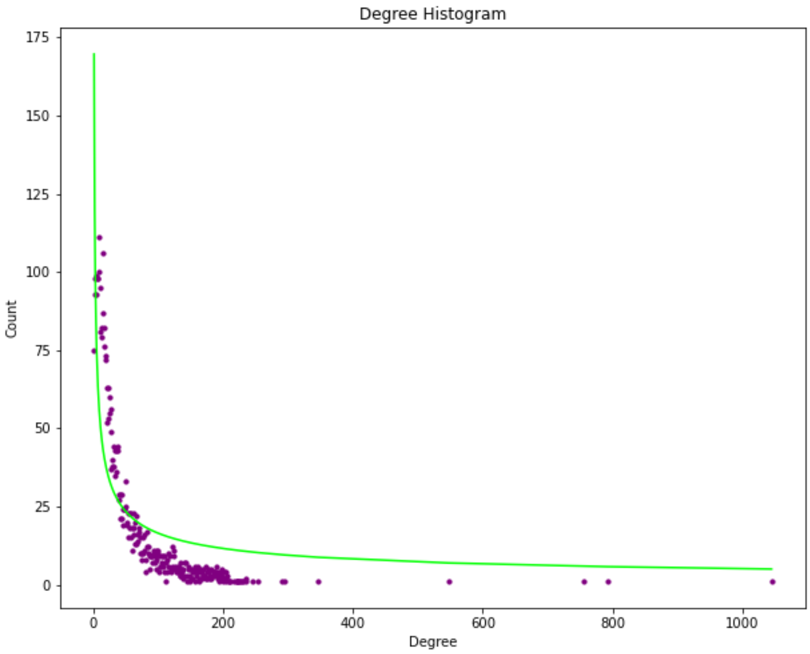


Figure 1 - Degree Distribution and fitting of the curve

At the first sight it seems that the chart in *figure 1* shows the degree distribution following the trend of a *power law* and the fitted curve is described by the equation $p_k \sim C * k^{-\gamma}$.

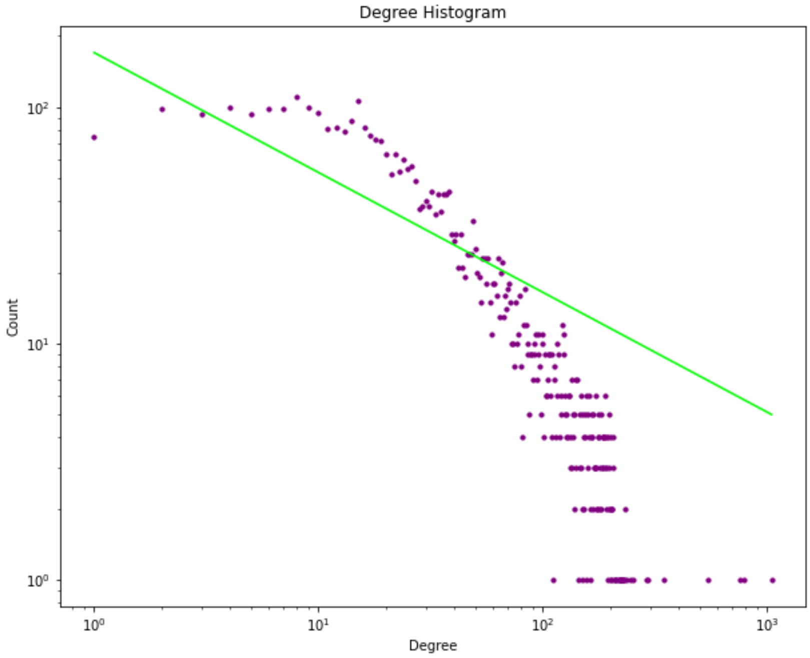


Figure 2 - Curve fitted plotted in logarithmic scale

Hence the *figure 1* represent the degree distribution in a scatterplot along with with the curve that represents the trend of the distribution. The curve has been computed fitting the power law function, estimating the parameter *C* and γ. This latest curve allows to prove that the degree distribution is effectively a power law, in fact showing the curve in logarithmic scale is possible to observe that the green line is a straight one, and this is a characteristic of a degree distribution that follows a power law *(figure 2)* .

What can be further observed in the figure 2, is that our network doesn't have a degree distribution that follows a pure power law. In fact real networks rarely display a degree distribution following a pure power law, instead real systems display a shape similar to what is shown in the figure 2 that share some common features:

- **Low-degree saturation** that is shown in the initial flatten P_k region. This happen when the network have fewer small degree nodes than expected for a pure power law.
- **High-degree cutoff** appears as a rapid drop in P_k : which means that the network has fewer high-degree nodes than expected in a pure power law, and also limiting the size of the hubs. This happens when there is a limitation in the number of links a node can have. Since our system is taken from a social network, is cutoff may be a due to the fact that one person can hardly maintain a deep and meaningful relation with a lot of people.

The degree distribution follows a power law distribution but it's not a *scale free* network, as can be observed in the plot above there are a few hubs (already described as **High-degree cutoff**) and this suggests that the second moment of the degree distribution doesn't diverge. We tried to simulate the behavior of the average path length as the network grows: starting from a random set of nodes, nodes and edges are added step by step so that the network remains connected. The following image shows the tendency of the distance, since the first nodes are randomly selected the distance doesn't follow one of the highlighted curves but then after some steps it converges to the $\frac{\ln N}{\ln \ln N}$ one.

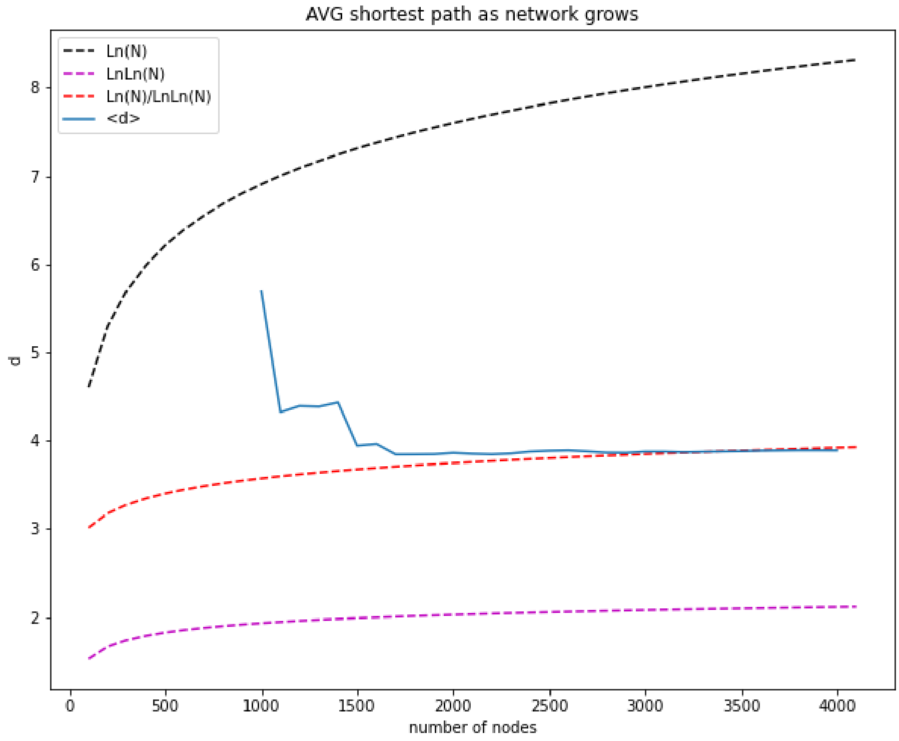


Figure 3 - average shortest path as the number of nodes increases

This result suggests that the γ might be equal to 3, which is the critical point between the *Ultra-small world* and the *Small world* regimes where the hubs are still enough to shrinks the distances compared to a random network of similar size. However we cannot be sure about that because of the limited dimension of the dataset.

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, the closeness centrality and the degree.

Betweenness		**Closeness**		**Degree**	
Node	**Value**	**Node**	**Value**	**Node**	**Value**
107	0.480	**107**	0.459	**107**	1045
1684	0.337	58	0.397	**1684**	782
3437	0.236	428	0.394	1912	792
1912	0.229	563	0.393	3437	547
1085	0.149	**1684**	0.393	0	347
0	0.146	171	0.370	2543	294
698	0.115	348	0.369	2347	291
567	0.096	483	0.369	1888	254
58	0.084	414	0.369	1800	245
428	0.064	376	0.366	1663	235

The table above shows the top 10 nodes with maximum betweenness and the top 10 nodes with maximum closeness together with the nodes and their degree: the nodes present in all the ranking are highlighted.

The betweenness measures how many short paths pass to a certain node, 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 betweenness and the higher closeness, so we can say that such node is very important in the network and it is in the main cluster.

The closeness measures don't vary too much from one node to the other, and this can be due to the logarithmic growth of shortest paths. By contrast the betweenness measures in the ranking decreases 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 pheriperical clusters to the "center" of the network.

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

The *average shortest path* is the average of the shortest paths between all the pairs nodes, the average on this network 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 comparing it with the total number of nodes of the network. Also considering the diameter equal to 8 which means that the longest path is made up of 8 nodes.

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, it is the number of edges in the network over maximum number of possible edges in the network. On our network the value is 0.01, this means that the network is particularly **sparse**.

2.5. Is the network assortative?

In a network can be measured the assortative mixing according to the degree distribution and:

- In an **assortative network** high-degree nodes tend to stick together and the structure of the network is characterized by a *core* of high-degree nodes. Hence hubs tend to link to each other and avoid linking to small-degree nodes meanwhile small-degree nodes tend to connect to other small-degree nodes avoiding hubs.
- In a **disassortative network** hubs avoid linking each other, and tends to link to small-degree nodes. The network result in a hub-and-spoke topology.

The pearson correlation coefficient of our network is 0.06, which means that the network is non-assortative and there is no a particular correlation between the degrees. It can be understood looking at the image of the network: the network is made up of clusters where some of them seems to be a hub-and-spoke topology and the others seems to have a *core* of high-degree nodes surrounded by a sparser periphery. So there isn't a dominant mixing.

2.6. Average clustering

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

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

, in which L_i is the number of links between the k_i neighbors. It captures the density of links in i’s immediate neighborhood, and it is in the range [0,1]. In this case the mean value of the network is 0.605, it means that on average a node has his 60% of neighbors connected to each other so *structural holes* doesn't seem to be a problem in such network.

On the other hand the global clustering is 0.51 and this could be due to the different local topologies in the networks, in particular there are components that seems to be densely connected and others components in a hub-and-spoke topology.