uc3m

Universidad
Carlos III
de Madrid

First Assignment: Wikipedia Norms

Puerta de Toledo

Master in Big Data Analytics

Network analysis and data visualization

Ion Bueno Ulacia NIA 100364530
Daniel Martín Cruz NIA 100384121

Ion Bueno Ulacia NIA 100364530
Daniel Martín Cruz NIA 100384121

1. Dataset

A society's shared ideas about how one "ought" to behave govern essential features of economic and political life. And, while the rational structure of rules and laws is an important part of coordinating actions and desires, people determine the legitimacy of these solutions based on beliefs about fairness and authority.

Online communities, such as Wikipedia, provide new opportunities to study the development of norms over time. Along with information and code repositories at the center of the modern global economy, such as GNU/Linux, Wikipedia is a canonical example of a knowledge commons.

In this study we will be focused and the part of this online encyclopedia devoted to information and discussion about the norms of Wikipedia itself. There exists plenty of norm sets depending on the language shared in a given community. In this study we will put the focus only on the norms related to the English-speaking community of Wikipedia.

In this sense, each of the pages will be one node of the studied network and the links between nodes will correspond to the existing links between these norm pages. The resulting network counts with 1976 nodes and 17235 edges connecting them.

This dataset has been found in the **Index of Complex Networks** provided by the University of Colorado and can be found after the name of *Wikipedia norms*.

For this analysis, we will make use of the tools learnt during the course by means of R for the first part of the study and Gephi for visualization and community detection.

2. Metrics

In this section we will go through the main statistics of the studied network. In order to obtain this information we will need the functions contained in the R library `igraph`.

2.1. Degree Distribution and Friendship Paradox

First of all, we will start with one of the most important metrics to know the importance of nodes: the **degree**.

By means of the `degree` function we obtain that the mean degree of the nodes of this network is **17.44**, a relatively low number having into account that there are more than 1000 nodes in this network.

The mean degree of the nodes does not give much information of the behaviour of this metric along the network. For this reason, we will extract some visual information with the goal of knowing how is the distribution of the degree.

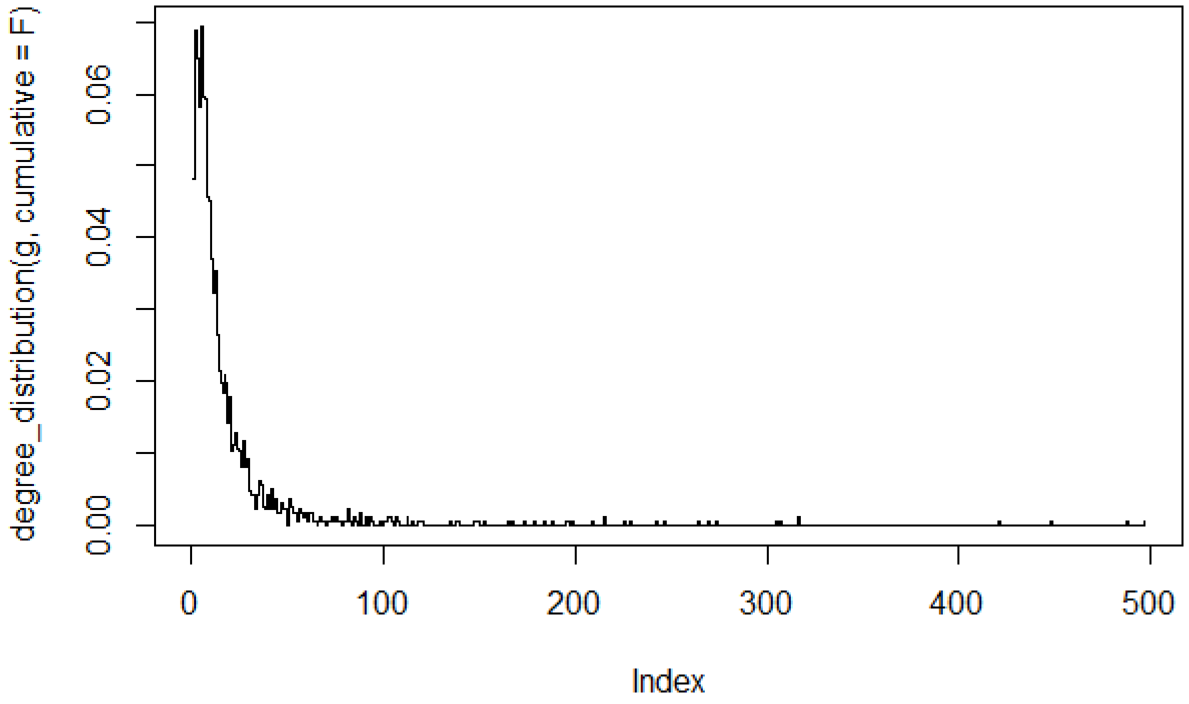


Figure 1: Degree distribution of the network

In this plot we can see that the degree distribution of the Wikipedia norms network follows a clear *Power-law* distribution, where most of the nodes have a small value when it comes to the degree metric while there are a few nodes with a considerably high degree.

After the analysis of the degree distribution, we will study if the Friendship Paradox is fulfilled in this network. This paradox states that "our friends have more friends than we do", or what is the same, that our adjacent nodes generally have a higher degree than ours.

In order to find out if this paradox is true in the case of our network, we have decided to generate some visualization of the distribution of the number of adjacent nodes for every node and the distribution of number of *friends* that each of those adjacent nodes has. If the Friendship paradox is true in this network, we should see that this second distribution is generally more at the right than the first one. Let us take a look at the histogram:

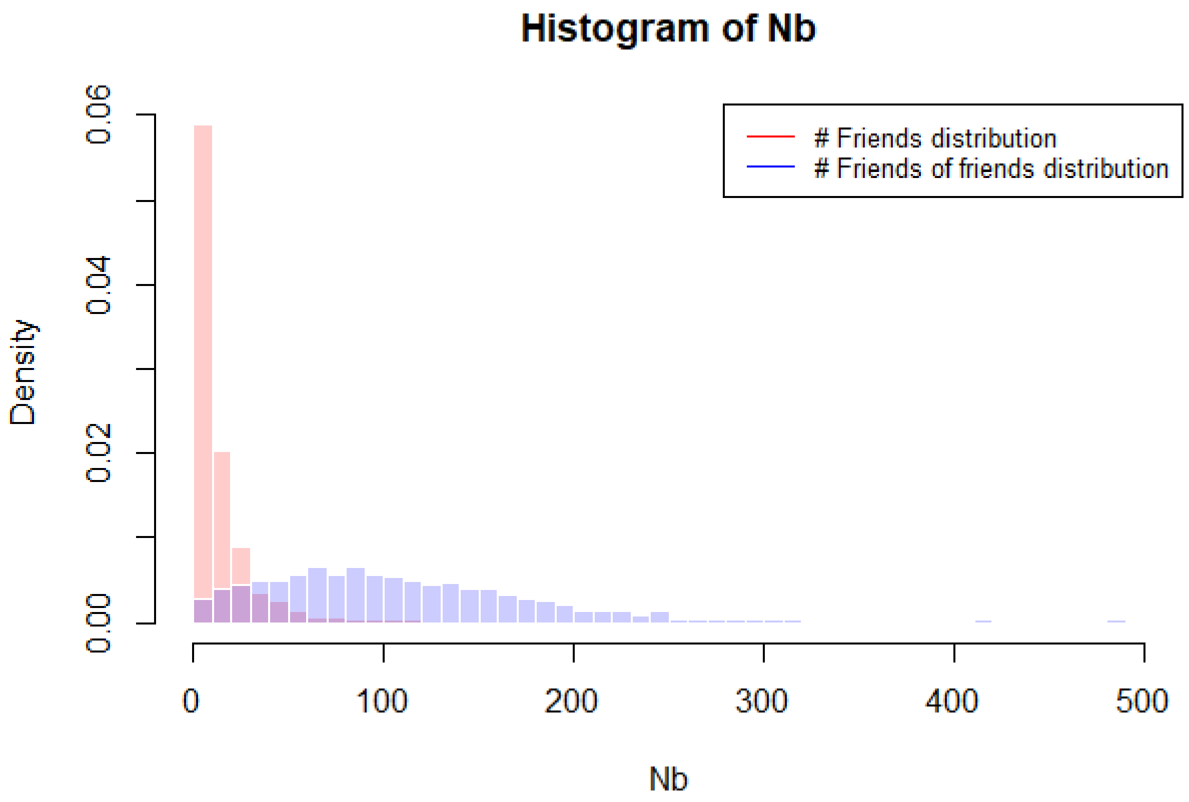


Figure 2: Number of friends and friends of friends distribution

As we can see, in general, the number of friends of a friend node (blue) is higher than the number of friends (red). If we compute the mean value of both variables, we obtain that each node has an average of **14.44 "friends"** while the average number a friends for each friend of a node is of **113 adjacent nodes**. It is important to mention that, in order to compute this second variable (`Nb2` in the code), we have not taken into account the `NA` values, which may have contributed in such a high value. In any case, we can conclude that **the Friendship Paradox is true** in the context of the *Wikipedia norms* dataset.

2.2. Assortativity Coefficient

The next metric to measure if we want to analyze the general behaviour of the network is the **assortativity coefficient** or Pearson coefficient. We can do this by means of the `assortativity_degree()` function included in the `igraph` library we are using for this study.

A high value of this metric implies that the nodes that are highly connected are connected between themselves, meaning that we have an assortative network. One of the consequences of this is the *rich-club* effect: popular nodes are not "mixed" with marginal nodes.

In the case of the *Wikipedia norms dataset*, the resulting assortativity coefficient has a value of **-0.0809**. With this value, we can conclude that this network is neither assortative nor disassortative but neutral. This means that popular nodes are connected between while that have also contact with less popular nodes in the network. This value of Pearson coefficient is close to theoretical value for this metric for the case of the Random graph.

2.3. Other metrics

Other metrics that can be valuable for the analysis are the diameter, the clustering coefficient or transitivity and the mean distance between nodes. In the studied network, we have the following values:

- Diameter: 9.
- Clustering coefficient: 0.109.
- Mean distance: 3.77.

Here we see that both the diameter and the mean distance have a relatively small value taking into account that the network has more than 1000 nodes.

In the case of the clustering coefficient, we see that is higher than the theoretical value for the case of a random graph but it is still a quite small value.

3. Model Networks

In this section of the study we will create new networks synthetically with based on different network models. We will create these new networks imposing that they share some attributes with our original network and then we will compare some metrics of these networks with the ones measured in the one formed using the *Wikipedia norms* dataset.

3.1. ER Random Graph

The first network model we will base on in order to compare metrics with our original network will be **Erdös-Rényi Random Graph**. We will start by building our desired network thanks to the help of the `erdos.renyi.game()` function provided by `igraph`. In the call to this method, we will impose that the created network has the same number of nodes and edges than our original network.

We will begin by analyzing the behaviour of the degree in this network. The mean value is of course the same as before, **17.44**, but we will observe a much different shape if we take a look at the degree distribution, that can be found in 3a. In this case, as it could be expected, the degree distribution follows clearly a normal distribution.

When it comes to the other metrics we saw for our original network, these are the values for this case:

- Assortativity coefficient: 0.003.
- Diameter: 5.
- Clustering coefficient: 0.009.
- Mean distance: 2.93.

The most remarkable metrics here are the mean distance and the diameter because this network is mainly characterized by hazing small shortest path length between nodes.

3.2. Configuration Model

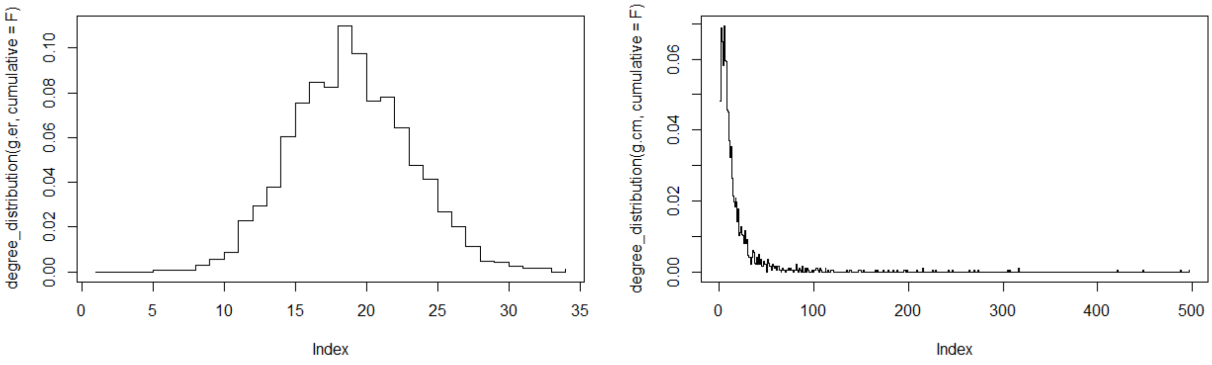
In this second part of the Model Networks section, we will study the behaviour of a network created the same degrees that our original one but following the **Configuration Network Model**.

When it comes to the degree of the network, there is not much to say in this case because the mean degree and the degree distribution of this network will be exactly the same as in our original network. This can be visually checked just by taking a look at 3b.

In order to analyze this network, we will compute the corresponding values for the metrics we used in the analysis of the previous networks:

- Assortativity coefficient: 0.002.
- Diameter: 11.
- Clustering coefficient: 0.075.
- Mean distance: 3.64.

As this series also have a random component as the previous one, we could expect a low clustering coefficient, low assortativity and relatively small shortest path lengths. This is exactly what we see here but, in this, the diameter and the mean distance are higher that in the case of the ER Random Graph. The reason of this is that in the creation of this network we are imposing a degree distribution that corresponds to the one of a real network and is not fully random.



(a) ER Random Graph

(b) Configuration Model

Figure 3: Degree distributions of the synthetically created networks.

4. Visualization and Community Detection

In this last section of the analysis of the *Wikipedia norms* network we will proceed with the visualization of the network and the detection of possible communities that may appear in our network. Both these tasks can be performed using R as we have been doing all along the study but, for the sake of simplicity and better visualization, we will use Gephi, an open-source tool that is one of the most used softwares when it comes to network exploration and visualization.

With the help of R, we have been able of exporting our network in `.gml`, which is a valid format for reading it using Gephi. After loading this `.gml` file in Gephi, we apply *ForceAtlas2*, a graph layout algorithm for a better network visualization designed by Gephi. We will also modify the size of the nodes depending on its degree, setting the minimum size to 8.5 and the maximum size to 18.5.

After this first preprocessing, we will use the community detection algorithm implement by Gephi using a resolution value of 1. Colouring each node according to the community it belongs, this is the resulting network:

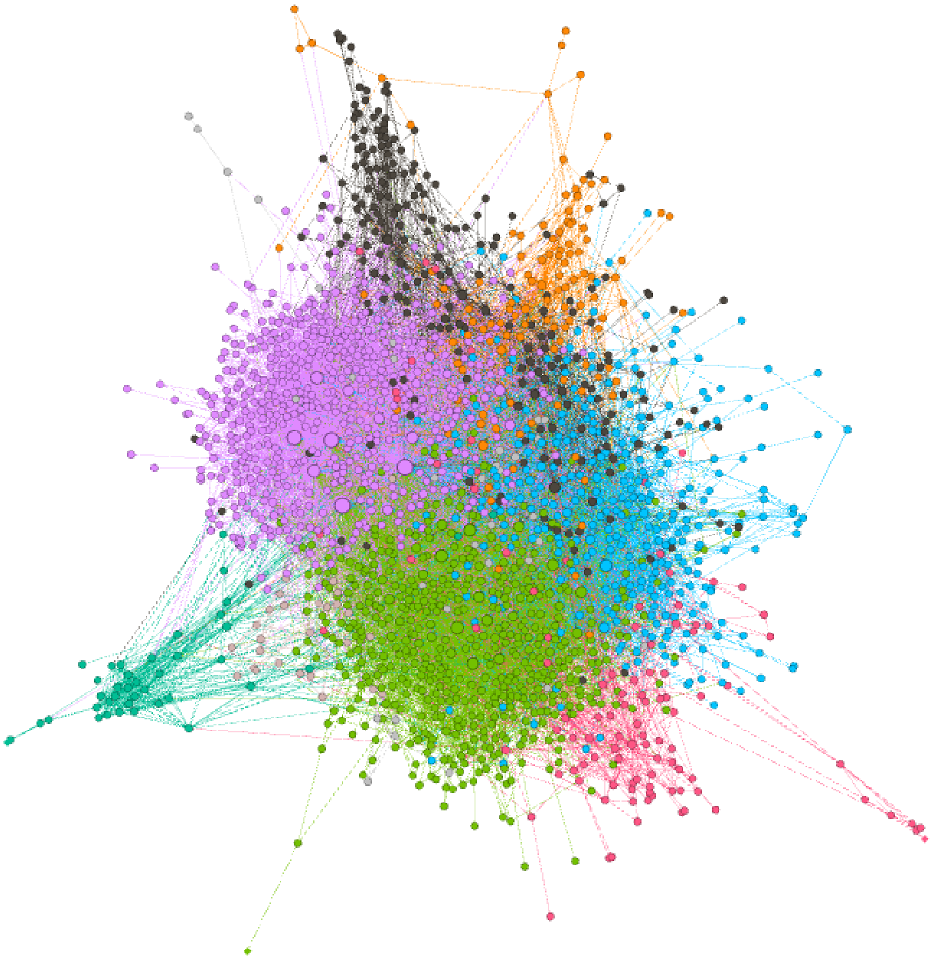


Figure 4: Network.

Here we see that Gephi detects three main classes in the core of the network identified by different colors: purple, green and blue, respectively. We observe that Gephi also considers other smaller communities on the periphery of the network: in the upper part we see the black community and the green one; in the bottom-right part we can find the community identify with the red color and in the bottom-left we see, maybe the clearest one, the community identified with a blue-green color.

As we have a considerably big number of nodes, we have not gone too deep in the interpretation of each of the communities but, after a first inspection, we have concluded that each the communities is related with the types of norms Wikipedia uses for classification. Among these types we can find: essay, convention, guideline, etc.