uc3m | Universidad Carlos III de Madrid

Master's degree in Statistics for Data Science
Academic Year (2019-2020)

Master Thesis

"Communication vs. Polarization: A network analysis of science communication in Twitter"

Roberto Rey Sieiro

Iñaki Úcar Marqués

Lucas Sánchez Sampedro

11/09/2020

AVOID PLAGIARISM

The University uses the Turnitin Feedback Studio program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

ABSTRACT

The aim of this project is to study the connections between the Science and Pseudo-science of the Spanish speaking community on Twitter and find out if there are any echo chambers. Misinformation is a dangerous tool commonly used in social networks, especially when this misinformation is related to health issues. The project will be carried out with a data set that has 55 million Twitter user profiles. To approach this issue, data filtering techniques were applied to obtain the user profiles of interest in order to build the network. The network was analysed through different network analysis techniques, such as centrality measures, which will help us to detect the main user profiles of the network and clustering that enables us to discover hidden communities belonging to the network. Some 'hinge' profiles were found connecting both communities; however, these profiles are not related with Science or Pseudo-science directly. The Science community seems more aware of the Pseudo-science community than the other way around. There were no echo chambers found according to the Science and Pseudoscience communities, although several sociolinguistic and political communities were found.

Keywords Social Networks; Network Analysis; Echo Chamber; Graph; Data Filtering; Pseudo-science

ACKNOWLEDGEMENTS

I would like to thank my project tutors, Iñaki Úcar and Lucas Sánchez for their guidance during the project. I would also like to express my gratitude towards all my friends who were there for me during this year. Finally, I would like to acknowledge my family that got me this far.

Roberto

INDEX

1. INTRODUCTION

1.1 Motivation

Twitter has proven in recent years to be a key tool for sharing both rigorous information and fake news. In terms of users, it is one of the top 15 most used social networks, with 326 million of active users, where nearly half of the population of Spain uses Twitter and a 12% of Latin American population, which is approximately 102 million users.

Twitter is a social network where the users send messages with a maximum of 280 characters called tweets and they are shown in the profile of the user. Other users can subscribe to the tweets of another user, this is called 'to follow' and the users that are subscribed to your tweets are called 'followers'. In addition, you can share tweets from another user to your followers, this is called a 'retweet', the tweets that you retweet are going to appear in your profile with your conventional tweets.

There are many kinds of 'information' in Twitter, like memes or rigorous scientific information. In the case of science, there are several ways of bringing up this information, from evidence-based information, which helps to bring science closer to our society, to pseudo-scientific theories, which feed the social network with misinformation and alternative therapies. There are many user profiles like these with a large number of followers, although we do not know if there is an open gap between each type of creator of content and their audiences. So, there are two possible options, people only get information from the 'closed' network or chamber that they are following, or Twitter is a perfect tool to reach non-captive audiences.

Subsequently, we can analyse further our case by obtaining the amount of impact, positive or negative, that a user profile has. In other words, if a user profile is following pseudo-scientific profiles just to criticise them or to refute their theories, or on the other hand, they could be sharing their content and spreading the information. This is going to be a harder task to analyse because following someone does not mean that you are sharing their content even if you are an active person on Twitter.

Moreover, misinformation is a very dangerous issue, especially when it is related to health issues. There were some cases where text messages filled with misinformation led to serious problems or even deaths [14]. For example, this kind of misinformation was present in several tweets related to the Ebola outbreak, and furthermore, these tweets containing misinformation had a larger reach than the correct information [14]. Misinformation is a much more common issue than it seems, and unfortunately goes very unnoticed. One of the most recent examples is the information of Coronavirus on Twitter, where approximately 25% of the tweets were classified as misinformation [11].

The use of alternative therapies or pseudo-science have the potential to harm people [15]. These kinds of therapies are widely spread across the internet. As this misinformation may have such a great impact, in this project we are going to see if this platform, Twitter, could be used as a tool to spread this misinformation in an effective way. Referring to an effective way by discovering closed networks to which no true information can reach. We first locate closed networks to which no true information can be reached, followed by identifying if these could enhance the credibility of their tweets and spread their misinformation to other networks.

1.2 Objective

We have a dataset with 55 million user profiles that follow scientific and pseudo-scientific content. The main objective is to study how connected science and pseudo-science users are on Twitter.

More specific objectives would be:

- Analysing how science and pseudoscience are distributed in different communities.
- Analysing the connectivity between the communities.
- Analysing the main user profiles of the network and the possible existence of 'hinge' user profiles, in other words, user profiles who are followed by both networks.

In the first section of the project the main variables of the data sets used will be described. It will also explain how the data filtering was done to achieve the final data sets in order to build the network. Furthermore, it will also talk about the network analysis techniques used during the project. In the next section, the network is going to be visualized and analysed. The analysis results are going to be discussed and visualized in graphs. In addition, the cluster analysis is going to be included in this section. Finally, in the last section, the conclusion of the analysis will be given. Likewise, we will include further work that can be derived from this project. In the appendix you can find a more detailed analysis of each cluster.

The code can be downloaded from: https://github.com/Rober157/TFM

2. METHODS

2.1 Dataset description

This work analyses a pre-existing data set that was downloaded between June and July of 2019. There were obtained two files, one of them shows every connection between the user profiles (who is following who). It has two variables, 'from' and 'to', where the 'from' column says who follows the 'to' column and composed by 165240230 links. The other file contains the information about these user profiles. There are 55377341 user profiles that are divided into two communities.

The user profiles used to create the communities are 'level 0' user profiles. Once we have selected these user profiles, the followers of these user profiles (level 1 user profiles) and the followers of the level 1 user profiles (level 2 user profiles) were downloaded.

The first community was created from 2 twitter users from the scientific communication community, which we will refer to as 'Naukas' and comprises of 54958196 user profiles belonging to this community. The second community was created from 6 twitter user profiles from the pseudo-science community, which we will identify as 'Pseudoscience', with 419145 user profiles. Despite more user profiles were used to create the 'Pseudoscience' community, the 'Naukas' community is a much more popular and therefore, there is an imbalance between the communities.

The data frame of user profiles contains the following variables:

- id_str: (int64) unique user identifier.
- location: (char) if any location is declared by the user.
- description: (char) blank space where you are free to write anything.
- protected: (bool) whether the user profile is protected.
- followers_count: (int) number of followers.
- statuses_count: (int) number of published tweets.
- created_at: (date) when the user profile was created.
- verified: (bool) if the user profile is verified.
- lang: (char) if any language declared by the user.
- community_id: (char) community "Naukas" or "Pseudoscience".
- community_level: (int) 0 main users, 1 their followers, 2 followers of the followers.

The **id_str**, it is an int64 and is the unique user identifier, also used in the links dataset.

The **location**, blank space where anything can be written.

The **description**, another blank space where you are free to write anything.

Protected, a Boolean that indicates if the user profiles are protected, these user profiles cannot be retweeted. In addition, you cannot see the tweets or the followers of these user profiles unless you are a follower.

The **followers count**, an integer that indicates the number of followers of that user profile.

The **statuses count**, an integer that indicates the tweets by the user including retweets.

Created_at, a date referring to when the user profile was created.

Verified, a Boolean that indicates if the user profiles are verified, being verified usually means being famous.

The **lang**, it is a character and refers to the language declared by the user, this can be a field with no value because the user did not declare any language. If any language is declared, the language 'xx' code is displayed.

The **community id**, it is a character, to which of the initial communities (Naukas or Pseudoscience) the user profile belongs to.

The **community level**, it is an integer, referring to which of the 3 levels the user profile belongs to.

2.2 Filtering

The first step is the filtering of the user profiles. The file that is going to be filtered here, is the one which contains all the information of every user profile. This filter is needed because we want to study the Spanish speaking community on Twitter. In addition, we want to create a network that can be managed by a computer.

So, during our filtering we must focus on:

- We want Spanish speakers.
- We need active user profiles.
- Creating manageable network

First, an 'objective' filtering is done, which is the language filter and removing every private profile. Later on, we filter the active user profiles, which on the contrary, is not objective given that you must establish a subjective threshold, even if we had inactivity data (you could establish: one month, one week, etc. of inactivity). The filtering and the rest of the code was done on Rstudio. To filter every private profile, we took the user profiles with variable private == FALSE.

Every time a filter is applied, the level 0 user profiles and the verified user profiles are not going to be filtered out.

2.2.1 Hidden private user profiles

There were several level 1 user profiles that changed their user profile from non-private to private while the download was being done. Because of this, some of the level 1 user profiles did not have any link to level 2 user profiles.

2.2.2 Language

There were two main cases when the language filtering was done, on the one hand, when the user declared his main language, on the other hand, when the user did not.

The first case was easier to approach, as analysing the column ‘lang’ was the only thing needed. This column contains different codes for each language, so ‘Spanish’, ‘Mexican-Spanish’, ‘Spain-Spanish’, ‘Galician’, ‘Catalan’ and ‘Basque’ were filtered as Spanish speakers.

The second case was a bit harder as we cannot analyse the column ‘lang’ as it does not have a value. The columns ‘description’ and ‘location’ were the ones used for this part of the filter. Every description was taken and analysed with a language detector, one of the language detectors called ‘Compact Language Detector 2’ uses Naïve Bayes to classify. The other language detector is called ‘Compact Language Detector 3’ and uses a Neural Network model for language identification. If they cannot detect the language they will return ‘NA’, so every time that one of them detected one of the languages above, we will classify the person as Spanish speaking. Now we have a Spanish speaking group and another group which their description is not in Spanish. The next step is analysing the location column. The location column is written by the user, this means that it can be written in many ways: uppercase, lowercase, city and country, only city, etc. Subsequently, the column location must be processed in order to do the filtering. The text is encoded as ‘UTF-8’, translated to ‘latin-ascii’ in order to remove every accent mark. Every comma was removed and transformed into a space and every letter was transformed to a lowercase letter. In the end, every row of the column location was composed by groups of characters separated by spaces.

To check if these locations belong to a Spanish speaking country or city, we need a list with cities and countries. To compare the location value of each user profile, we used an R database called world.cities from the library maps. It contains world cities of population greater than about 40,000. The names of the cities and the countries were processed in the same way as the location column.

Figure 2.1 Spanish speaking countries (Keith Admin, 2020).

We chose only cities from countries where Spanish is the official or national language, and cities with more than 100k population. The limit of population was established because many small cities in America have similar names to European cities, for example: Roma. Furthermore, a manual approach was done, some cities or countries have some prefixes such as San Francisco or New Zealand, that had to be manually removed from the list. Every time we found a country or city in the location column in the list, the user profile was classified as a Spanish speaker.

2.2.3 Activity

We needed a medium size network that can be managed by a computer (around 300k nodes maximum) but still having enough nodes of both communities. As both communities are unbalanced, we must do a different filtering for each community.

In our case, we do not have data about inactivity, so the frequency of the tweets was taken as an indicator of inactivity. If a user tweets with a certain frequency, we can determine that it is an active user. To obtain this variable we just took the number of tweets divided by the number of days in twitter. Both communities had different tweeting frequencies, so this was another incentive to do a different filtering per community.

TABLE 1. TWEETS PER DAY NAUKAS COMMUNITY

Quantile	**0%**	**25%**	**50%**	**75%**	**100%**
Tweet freq	0	0.64	2.04	5.75	1647.66

TABLE 2. TWEETS PER DAY PSEUDOSCIENCE COMMUNITY

Quantile	**0%**	**25%**	**50%**	**75%**	**100%**
Tweet freq	0	0.02	0.14	0.94	1647.66

The first community filtered was the Pseudoscience one, to find active user profiles a lower threshold was stablished. This threshold was tweeting every two days, which

translates into 0.5 tweets per day (top 33% of the community). In addition, we are interested in users with a great number of followers, or in other words, those who can spread information easier, so a minimum of 100 followers per user profile was applied.

The other community, Naukas, is a larger community, so we need to do a more restrictive filter in order to balance the communities. So instead of filtering the top 33%, we are going to take the top 10%, with this condition we can meet the requirements for a manageable network and somewhat address the balance issues. In the Naukas community the top 10% of tweeters tweeted 13 tweets per day or more. In addition, the follower criteria was also established, although it only removed a minimal amount of user profiles.

2.2.4 Components

Creating the network acts as an 'indirect filter'. We must create the network according to the user profiles that we have, but there are some level 2 user profiles that passed the filter, but their respective level 1 user profile did not. These user profiles should be removed, too. If the level 1 user profile is not active, it is not spreading information, this means that the level 2 user profiles are not going to get the information, people do not check if every profile they are following is active. There is a possibility that someone from level 2 is following an inactive profile from the level 1.

2.3 Creating and analysing the network

To represent the network a graph was created. To create the graph, we used the file which contains the connection between the user profiles, thus, we now use the edges database. To create this graph, 'igraph' was used. Igraph is a library and R package that provides tools for building and analysing these kinds of graphs. The graph created is a directed graph, each arrow represents the 'follow' action. For example, if an arrow goes from the vertex V1 to the vertex V2, the vertex V1 follows the vertex V2. However, the vertex V2 is providing information to the vertex V1, hence, the graph was inverted in order to represent the flow of information.

The order of a graph is the number of vertices that a network has, the size of a graph is the number of edges. These are good ways of determining the size of the network.

Centrality in a network can be understood as a measure or a value that has a vertex within a network. This, when evaluated on a scale, determines its relevance within the network and allows this vertex to be compared or contrasted with others. Knowing the centrality helps to determine the impact it causes within the network it is a part of.

The simplest one is the **vertex degree**. It measures the number of links or connections that a node has with the other nodes belonging to a network. When such an analysis is applied, different conclusions can be obtained. For example, in social networks we can measure the degree of entry of a node as its popularity, while the out degree can be defined as an indicator of sociability.

The **closeness** in a network is one of the most applied and developed concepts since with this measure we can determine the shortest or most efficient routes to get from one node

to another. This calculation is widely used because it can be interpreted as the speed at which information can be propagated from one node to all the others. This somehow allows us to understand the accessibility of a given node within a network.

The **betweenness** is a measure of centrality that quantifies the frequency or number of times a node acts or bridges a short path between two given nodes. When high betweenness nodes exist in a network, they usually play an important role in the structure to which they belong. These nodes also have the capacity to be controllers or regulators of the information flows within the overall network structure.

These measures can be studied for the directed graph and the undirected graph, although we are only interested in the directed one because it is going to show which are the main information spreaders. For instance, if a user profile follows all the network but no one follows him back, according to the directed graph he would not be important but according to the undirected measures he would probably be the most important one.

There are other kinds of important measures for the analysis, connectivity measures. The edge density is a connectivity measure that is calculated by dividing the total number of edges by the possible number of edges which is computed as $n \cdot (n-1)$, being 'n' the number of nodes for a directed graph. These connectivity measures are going to be important in order to analyse the cohesion of the different clusters in the future.

2.4 Visualization and clustering

Having finished the analysis, the next step is to visualize the network and obtain the different components or clusters of the network. To perform this part, 'Gephi' was be used as the representation tool. Gephi is open-source network analysis and visualisation software.

Clustering divides a database into different groups. The main goal of the clustering process is to group nodes by their similarities. Although we have two predefined communities (Naukas and Pseudoscience), we want to find 'hidden' communities. There are several cluster algorithms, but in this case, we are going to use the conventional algorithm, the Louvain's method. Louvain's method of community detection allows communities to be extracted from large networks. This community detection method seeks to optimize modularity, a number between -1 and 1 that compares the density of edges inside and outside a community.

These clusters can be analysed individually the same way as the main network. Obtaining the main vertices of each cluster (with the centrality measures) we can guess what that community represents. For instance, if Tolkien were a main vertex, that cluster could be representing Lord of the Rings fanbase.

After analysing the clusters individually, analysing the connection between the clusters is the next step. Observing this connection can lead to interesting conclusions, once we can identify the background of the individual clusters, the connection between clusters can be related to these backgrounds. For example, if there are two groups: football and

basketball and the connection between them is high, this could be because people in that network like sports.

There are different ways of analysing the connection between clusters, one of the measures is the coreness. Although, it is a measure associated to each vertex and we are looking for a more global measure as we are looking to compare the clusters. Modularity is a good measure, however, as we are dealing with big networks, the Louvain's algorithm might find smaller communities inside these two clusters. This means that the modularity measure would be according to these small communities and not according to the two clusters we want to compare. To deal with this problem, a 'connection density' was computed. This connection density is based in the total number of possible connections. First, there are two clusters which are analysed individually.

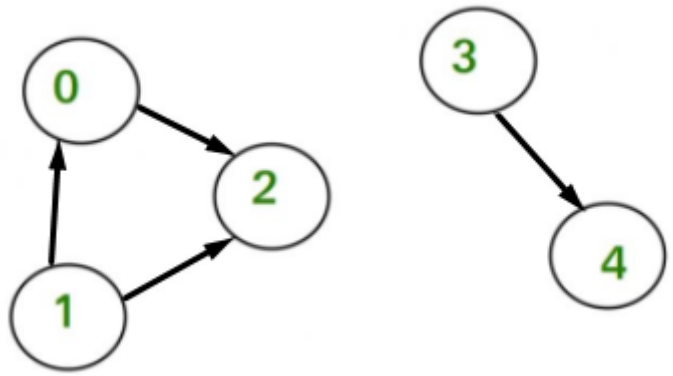

Figure 2.2 Calculating connection density

The first cluster contains 3 edges and 3 vertices and the second one 1 edge and 2 vertices. The next step is, taking the whole network and extracting both clusters, but now, with the connections between them.

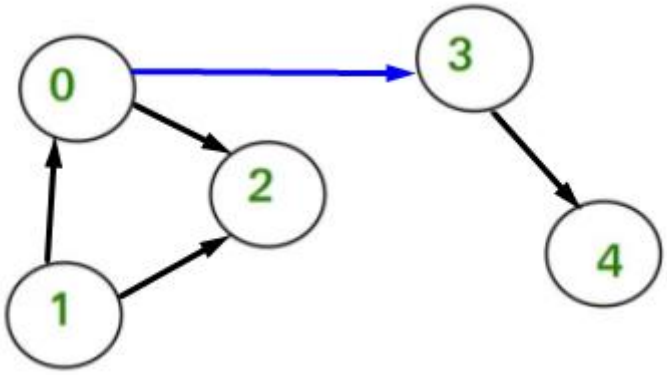

Figure 2.3 Calculating connection density

By taking the whole network, we can observe that there is only one edge that connects the clusters. Each vertex of the first cluster could be connected with the 2 vertices of the second cluster. These are $3 \cdot 2 = 6$ possible connections. Each vertex from the second cluster could be connected with the three vertices of the first cluster, these are $2 \cdot 3 = 6$

possible connections. There is 1 possible connection out of the 6+6 = 12 possible connections, so the connection density is 1/12. To obtain the general case for two clusters:

- Obtain the connection edges as: edges of the total network – (edges 1st cluster + edges 2nd cluster)
- Obtain the order of the first cluster, n.
- Obtain the order of the second cluster, m.

$$Connection\ ratio = \frac{Connection\ edges}{n \cdot m \cdot 2}$$

3. RESULTS AND DISCUSSION

3.1 Filtering

At the beginning we had 55377341 user profiles, once the filtering was done, we ended up with 185919 user profiles.

When the first filter was applied, which was removing the 'hidden private' user profiles, this only removed 9200 from the raw data set. The user profiles eliminated by this filter was a residual percentage of the total amount (0.017%). This was an expected result given the time between downloading the first data set and the second one, some profiles changed to private.

The second filter applied was selecting the non-private user profiles. The community Naukas went from 54958196 user profiles to 50551449, roughly 9% of the user profiles were removed. The community Pseudoscience went from 419145 user profiles to 382334, about 10% of the user profiles were excluded. Both percentages are quite similar, so it seems that they had a similar proportion of private user profiles. It was important to remove these profiles since we did not have their followers.

The third filter applied was detecting Spanish speakers, we got 4912353 user profiles in Naukas and 37229 in the community Pseudoscience. Spanish speakers represent approximately 10% of each community. Despite our level 0 user profiles were Spanish speakers, many user profiles were removed, possibly many of them were followers from English-speaking verified profiles.

The next filter was the activity filter, first we established the minimum number of followers we wanted per profile, 100 in this case. Then we filtered according to the activity thresholds we established in the methods section. Once we applied this, only 211126 Naukas user profiles and 11132 Pseudoscience user profiles were left. Now the community Naukas is 20 times bigger than the community Pseudoscience, not like in the beginning which was 100 times larger.

The last filter was the creation of the network, it acted as an indirect filter, this filter removed the level 2 user profiles which were linked to an inactive level 1 user profile. When this filter was applied, nearly half of the user profiles of Pseudo-science were removed to a total of 5245. The Naukas community was less affected by this filter, it kept 88% of the user profiles, which translates into 185919 user profiles. As the Pseudoscience network is smaller, removing a user affects a lot more than removing user profiles in a bigger network, this is the reason why this filter affected more the Pseudoscience community than the Naukas one.

We have 5245 Pseudoscience user profiles and 180674 Naukas user profiles, which sums up to 185919 user profiles. They are still quite unbalanced but more balanced than previously, Naukas is 35 times bigger than Pseudoscience community. Pseudoscience represents roughly 3% of the total network.

3.2 The whole network

The whole network has 185919 user profiles (nodes) and 1711625 edges.

This is the visualization of the whole network, every cluster detected by Louvain's algorithm is in a different colour. The modularity obtained was 0.424, where having a modularity bigger than 0.3 indicates that the subgraphs of the corresponding partition are modules [8]. This means that the breakdown of the network is not random and the nodes belonging to each community have characteristics in common.

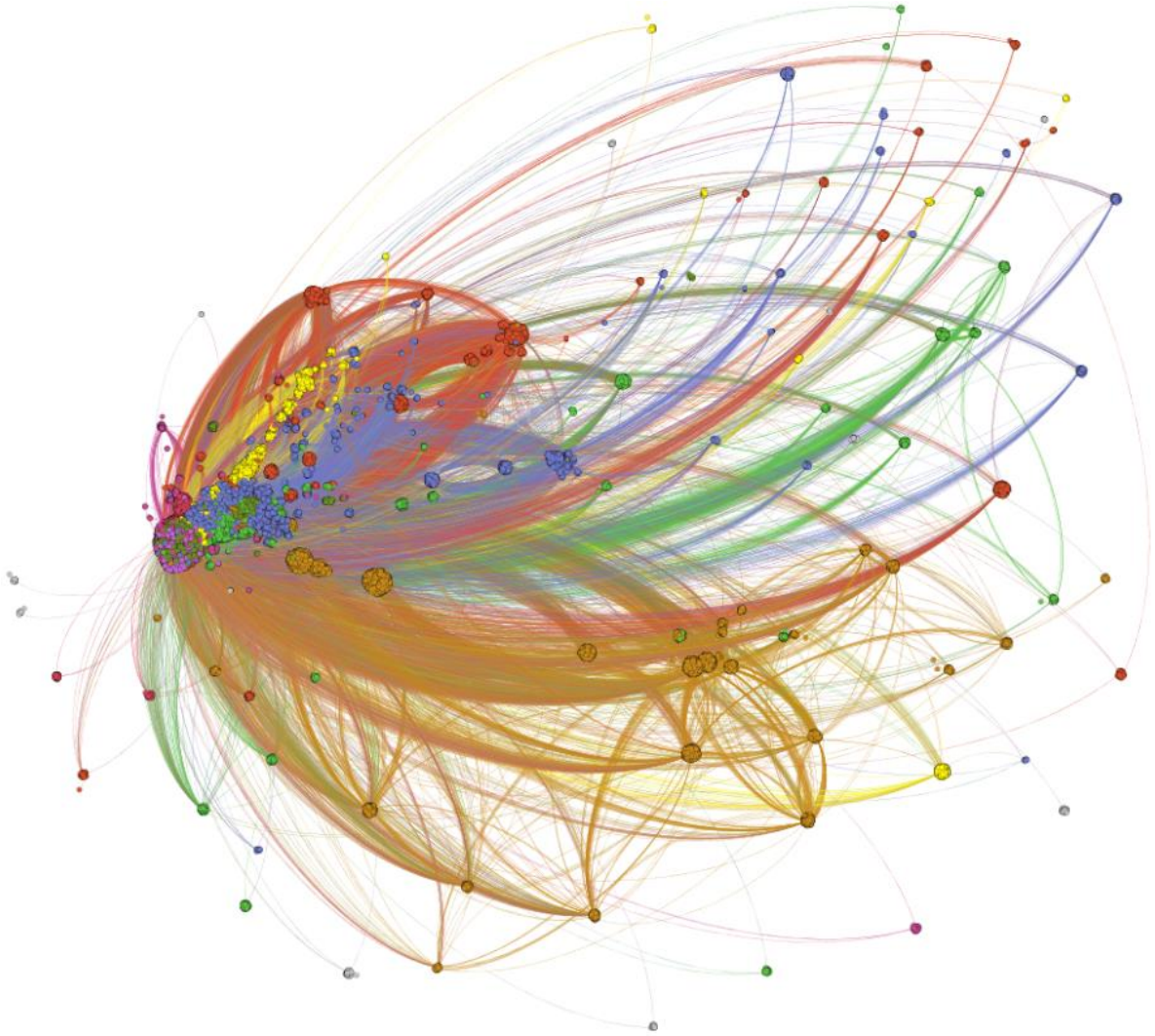

Figure 3.1 Visualization of the network and clusters

20 communities were detected:

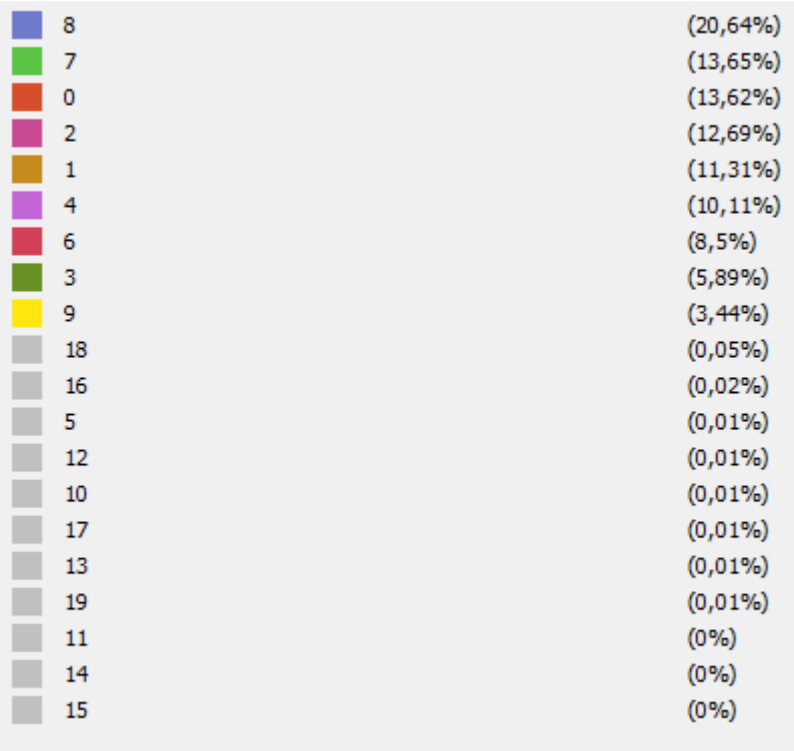

Figure 3.2 Communities detected and percentage of the total network.

We can observe that the first 9 communities represent 99.85% of the total network. The analysis is going to be based on these 9 communities as most of the smaller communities are composed in total by less than 280 nodes.

These 9 communities were analysed with the different centrality measures and by obtaining the most important user profiles of each community, we could guess the 'hidden identity' of these communities. All the tables and the respective analysis for each community will be in the appendix of the project.

According to the most important user profiles of each cluster, we are going to assign them a different name in order to identify them more easily.

- Cluster 0 → Chilean and Scientific Newspapers
- Cluster 1 → EEUU
- Cluster 2 → Comedians, Influencers and Scientific disseminators
- Cluster 3 → Major Brands and Multinationals
- Cluster 4 → Spanish Left-Wing
- Cluster 6 → Spanish Right-Wing
- Cluster 7 → Argentina
- Cluster 8 → Venezuela
- Cluster 9 → Colombia

The connections between the communities seem to be caused by political and cultural trends. Although a language filter was established, we can see a community that contains English speaking user profiles. This is one of the effects of not filtering the verified user profiles regardless of the language. This does not mean that all the user profiles are English speaking, but the most important user profiles are (which are verified user profiles).

Once we have identified each of these communities, we can plot how much percentage of the total network they represent.

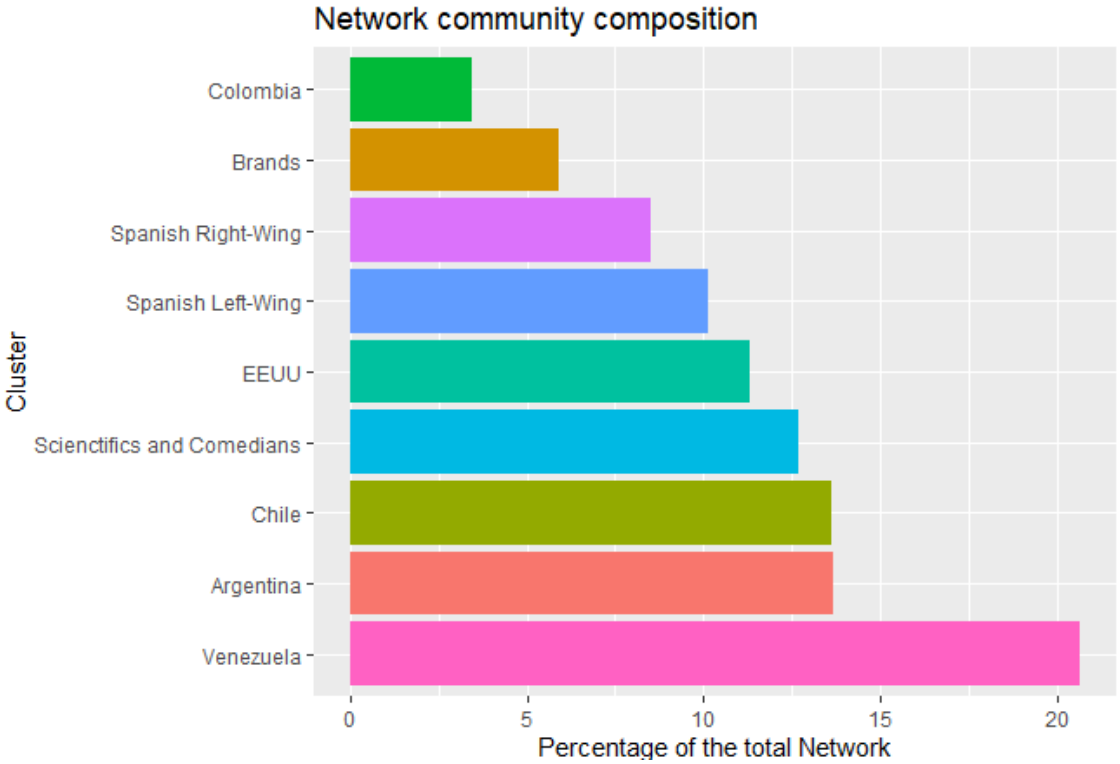

Figure 3.3 Percentages of the total network.

Even though it looks like most of the network is formed by non-Spanish (in terms of the country) user profiles, the Spanish community represents 37.19% of the total network, the highest percentage among all the countries. As the Spanish community is so big (and more dense edge wise), it was broken down into different communities.

3.3 Pseudoscience community

The Pseudoscience community has 5439 edges and 5183 nodes, which represents 3% of the total network, even smaller than the Colombian community which was the smallest one with 3.44%. We are going to represent the pseudoscience community (red) in the whole network graph.

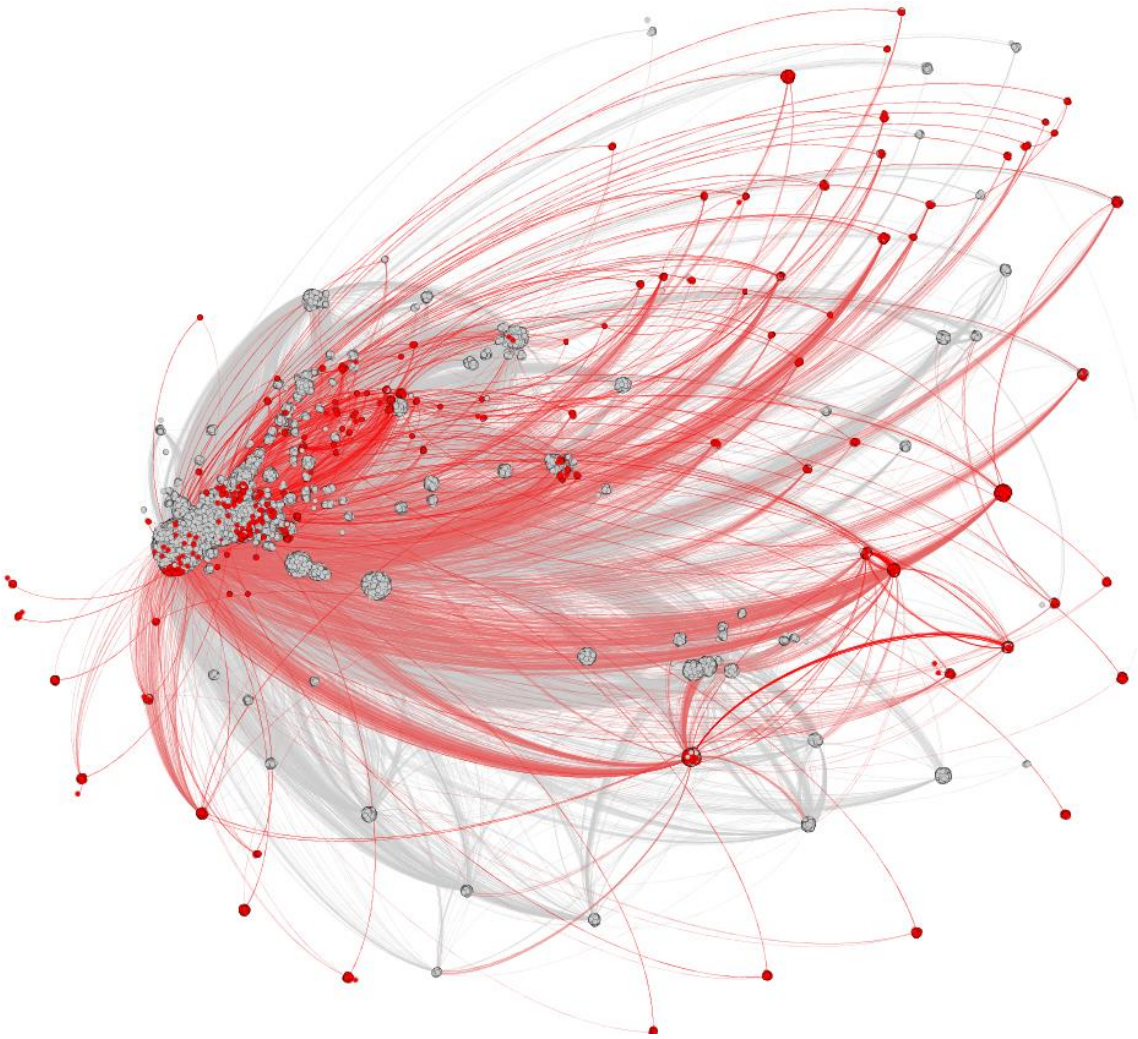

Figure 3.4 Pseudoscience in the whole network.

It seems that Pseudoscience is not clustered and it is evenly distributed, although we must define the proportions of Pseudoscience in each of the communities. There are a lot of external nodes represented in red, some of them are the small communities that were removed from the analysis and the rest may be close friends' networks.

By determining the percentage of Pseudoscience in the communities, we can check if this community is clustered within one of the communities obtained previously.

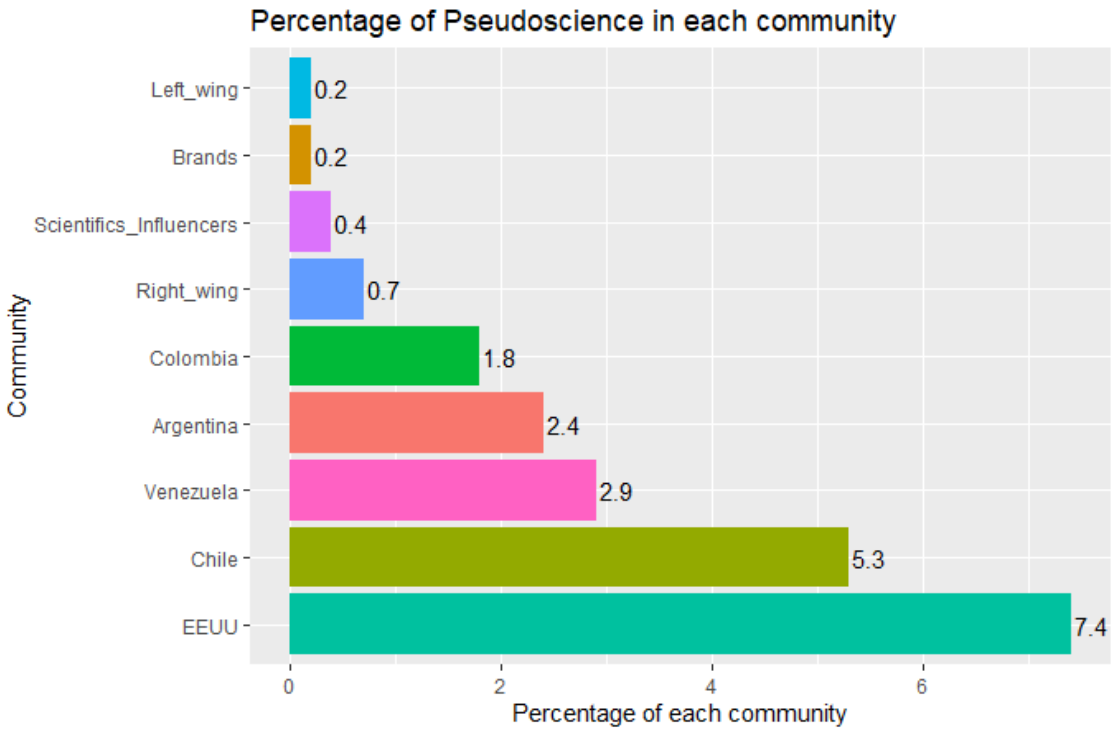

Figure 3.5 Percentages of pseudoscience in each community.

As we can observe, the percentage of pseudo-science in each community is low, the highest one is in the EEUU community. This value is probably due to the user @BoironUSA, a pharmaceutical company famous for the production and distribution of homeopathic medicines. The Spanish communities have the lowest value according to the pseudo-science proportion in their community. Chile has the second highest value; this is interesting because some science magazines were the most important user profiles in the Chile community.

Although it seems that pseudoscience is evenly distributed along the communities, it does not seem that pseudoscience is a closed community.

3.4 Degree Distribution

Comparing the degree distribution of the Pseudoscience network and the whole network can give us a general view of each of the communities. We are going to use the log degree to reduce the skewsness of both graphs.

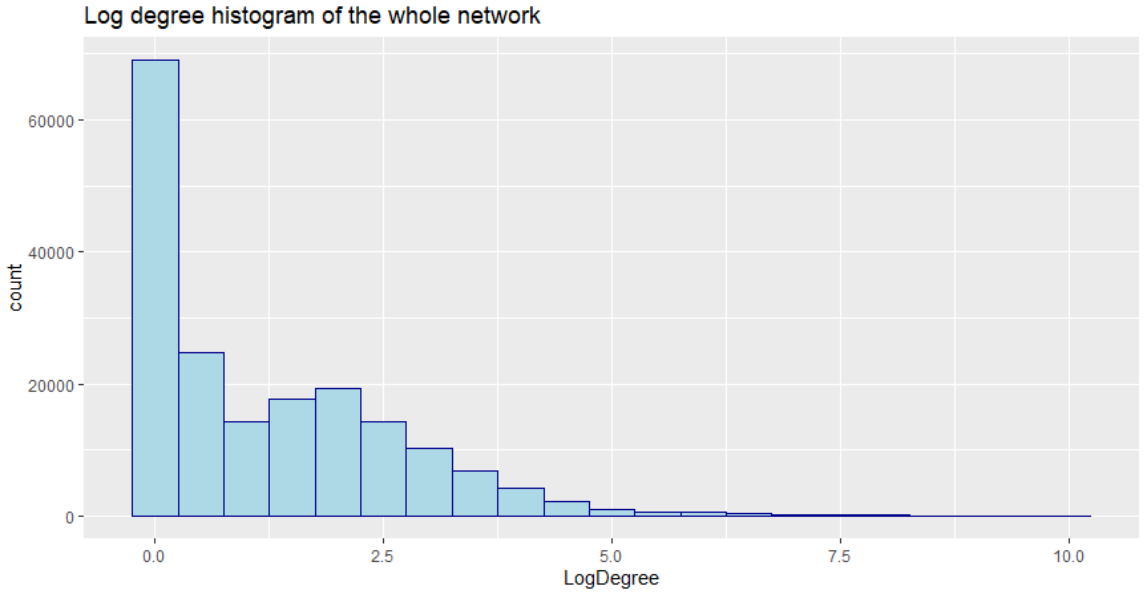

Figure 3.6 Log-Degree histogram of the whole network

This is the log degree histogram of the whole network, it is descendant as expected, as we can observe most of the user profiles have degree equal to 1 (exp (0) = 1). These user profiles are probably level 2 profiles, as they are the followers of the followers, one link means that they are only following one profile within the network. Another case could be level 1 profiles with no active nor Spanish speaking followers, because level 1 profiles are always linked to a level 0 profile, which means that their degree is minimum 1. There are several profiles with a huge number of followers.

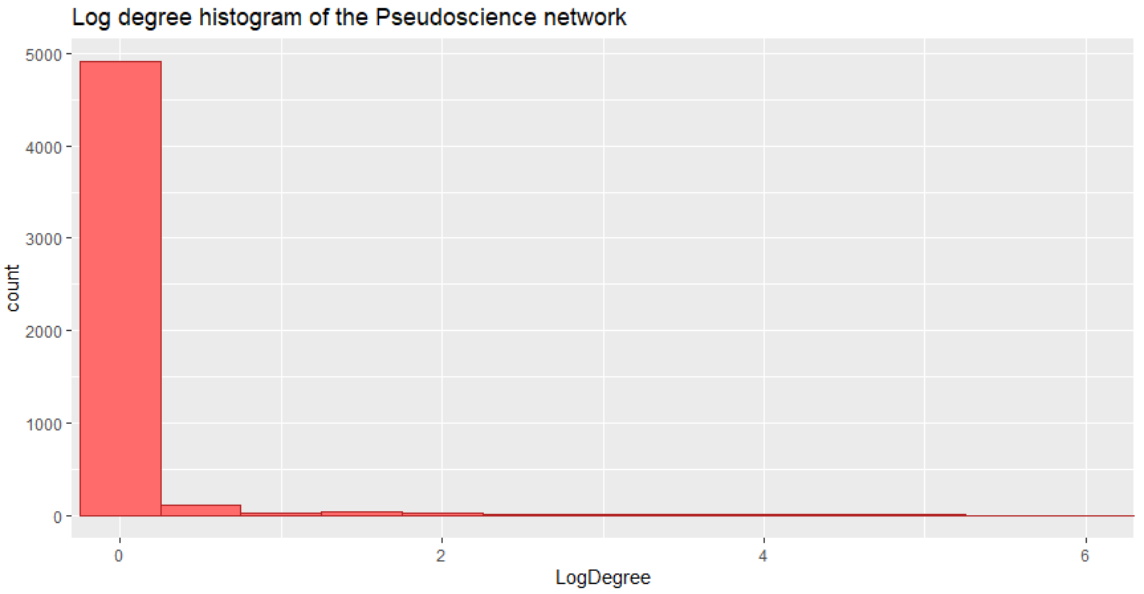

Figure 3.7 Log-Degree histogram of the pseudoscience network

This is the log degree histogram of the Pseudoscience network. We can see a huge jump from the degree 1 to the next degrees. Most of the network is composed by level 2 user profiles, or level 1 user profiles with no followers. If we check our data, approximately 98% of the profiles are level 2 profiles. The network density seems to be low as we have

nearly one edge per vertex. Therefore, this was one of the reasons why most of the smaller communities were created when the network visualization took place.

3.5 Edge density and connection density

In this section, the different density measures of each community will be analysed. First, we are going to start analysing the edge density per community.

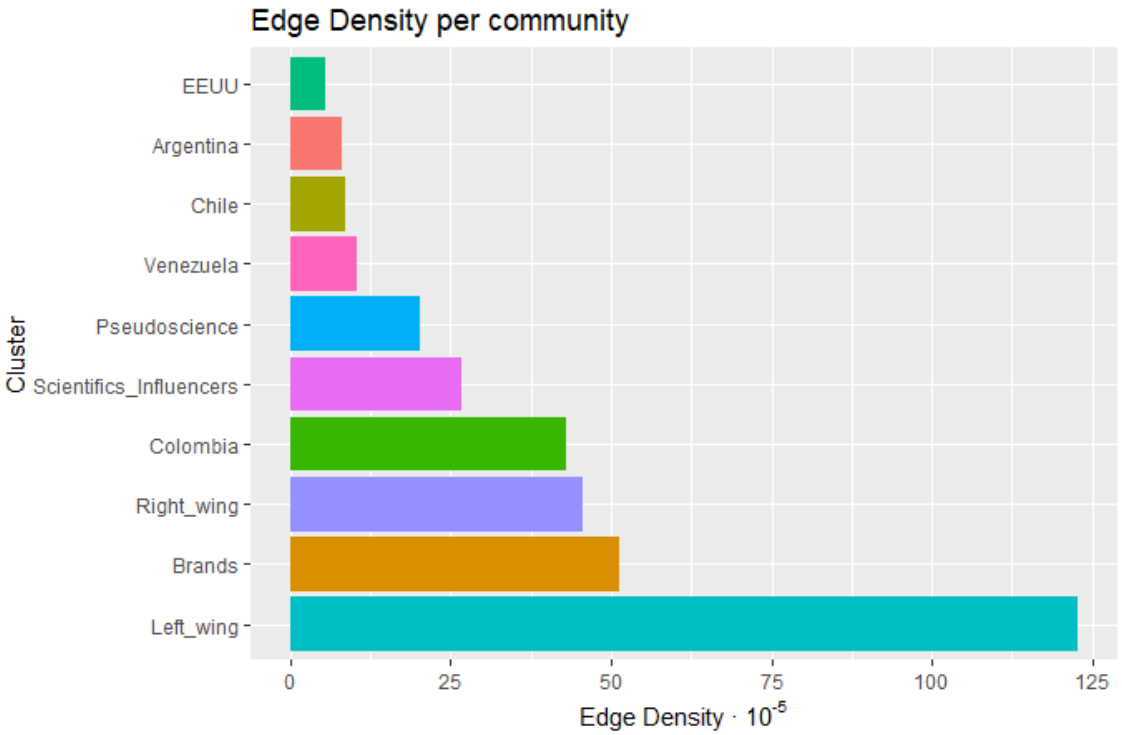

Figure 3.8 Edge density per community

The communities with the highest density are the Spanish communities. The one with the highest density is the Spanish left-wing. As is well known on Twitter, the left-wing is very active in social networks, more so than the right-wing [3]. On the other hand, the Colombian community, has a high density, one of the reasons being that it is the small and therefore it is easier to achieve higher densities. The lowest one is the EEUU community, as most of the non-Spanish speaking followers were eliminated. In the end, many EEUU verified user profiles were not filtered out, which are a large percentage of this community. In addition, this network was relatively big.

The Pseudoscience community seems to have a medium density, although the lowest connection density possible for a connected graph of that size is $19.3 \cdot 10^{-5}$ and its current density is $20.2 \cdot 10^{-5}$, meaning that it has one of the lowest possible densities. These densities might be a bit misleading because we are dealing with networks that must be at least weakly connected.

According to the connection between communities, we are not going to include the Pseudoscience one, this is because it is evenly distributed along the communities. In addition, it is a very small network. This could translate into misleading connection densities.

To represent the connection density between the communities, we are going to use a heat map. The heat map is row scaled, the rows are going to show us which community has the highest and lowest density connection between the rest of the communities. The columns are going to be a non-scaled indicator of the connection density. For example, the row ‘Chile’ and the column ‘Brands’ show the connection density between Chile and Brands but scaled according to the Chile values.

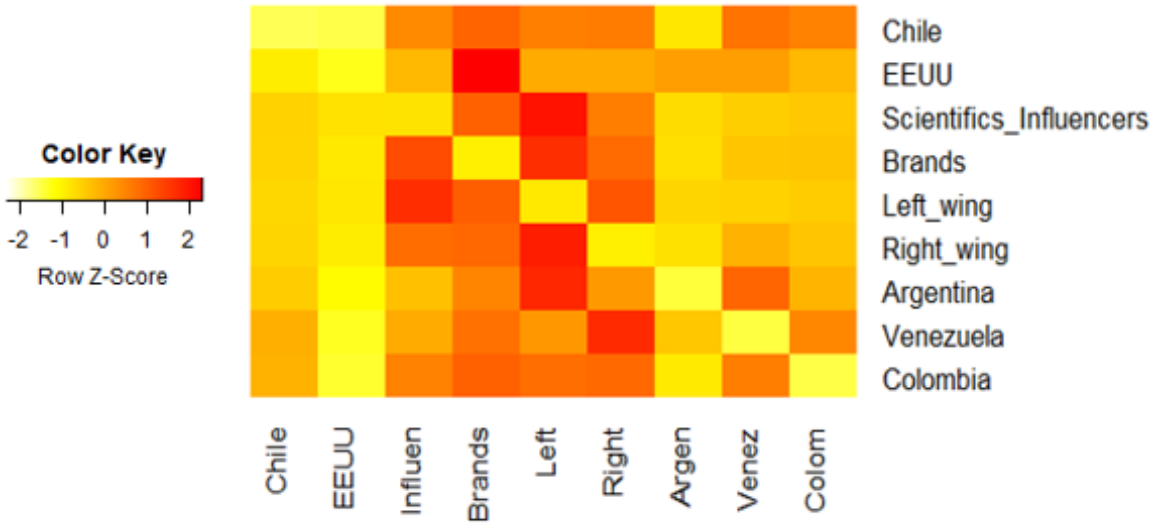

Figure 3.9 Connection Density between communities

Now let us look at the different connection densities. Communities with a high edge density will probably have a higher connection density.

- The Chilean community has the lowest connection density with EEUU, this is because both are very low connection clusters (EEUU the least). The highest connection densities are with Brands, Venezuela and Colombia.
- The EEUU community has the lowest connection densities, the highest is with cluster brands community, while its connections with other countries are very low.
- The community with influencers and Scientifics has the highest connections with the Spanish communities, as expected. The highest with the Spanish left-wing and the lowest with the EEUU community.
- The brand community has a similar connection density between the Spanish communities. The heat map emphasizes that this community has better connections with the other countries; it seems that it is a community that builds more 'bridges' with other countries.
- Left wing cluster is similar to the community of influencers and comedians (with whom it has the greatest connection). It has a low connectivity with other countries.
- The right-wing cluster has the highest connection density with the Spanish left-wing. Another point to note is that it has many connections with the Venezuelan cluster, probably due to political issues.
- The Argentine cluster is very balanced around connections, most of the connections are proportional to the density of each of the clusters.

- The Venezuelan community has many connections with the brands community, but the one with the greatest connection is with the Spanish right-wing.
- The Colombian community connections with the Spanish communities are evenly distributed, without much preference for anyone. On the other hand, it has a high level of connectivity with the Venezuela cluster, which is an adjacent country.

3.6 User profiles analysis

In this subsection, we are going to analyse the user profiles found during the cluster analysis. Next, we will analyse the connections between Naukas and Pseudoscience communities.

The most important user profiles of each cluster are:

- Cluster 0 → EresCurioso, Muy interesante and El Ciudadano.
- Cluster 1 → Lil B, Eric Schiffer and Science Friday.
- Cluster 2 → Alex Riviero, @lavecinarubia, Raquel Sastre and @aberron.
- Cluster 3 → Movistar España, TodoStartUps and Vodafone España.
- Cluster 4 → Ignacio Escolar, El Diario and Público.
- Cluster 6 → ABC, Arturo Perez Reverte and EuropaPress.
- Cluster 7 → C5N, Marcelo Parrilli and Jorge Arreaza.
- Cluster 8 → Miguel Henrique Otero, Emilio Gómez Islas and TalCual.
- Cluster 9 → Vanessa de la Torre, Artur Duanga and Telemedellin.

Most of the profiles we can see are related to the media, both television and newspapers. When we talk about individuals, except for some specific clusters, they are journalists or politicians. As it was expected, some of the more important user profiles are scientific disseminators or related with science. There were no Pseudoscience user profiles at the top 10 of each community. A brief description of these user profiles is found in the appendix of the project where we analyse each of the clusters separately.

3.7 Naukas to Pseudoscience

There are several nodes that connect the Pseudoscience and the Naukas community. These nodes could be responsible for the exchange of information between the communities. We are going to see which are the user profiles with more links that belong to the community Naukas and reach the Pseudoscience community.

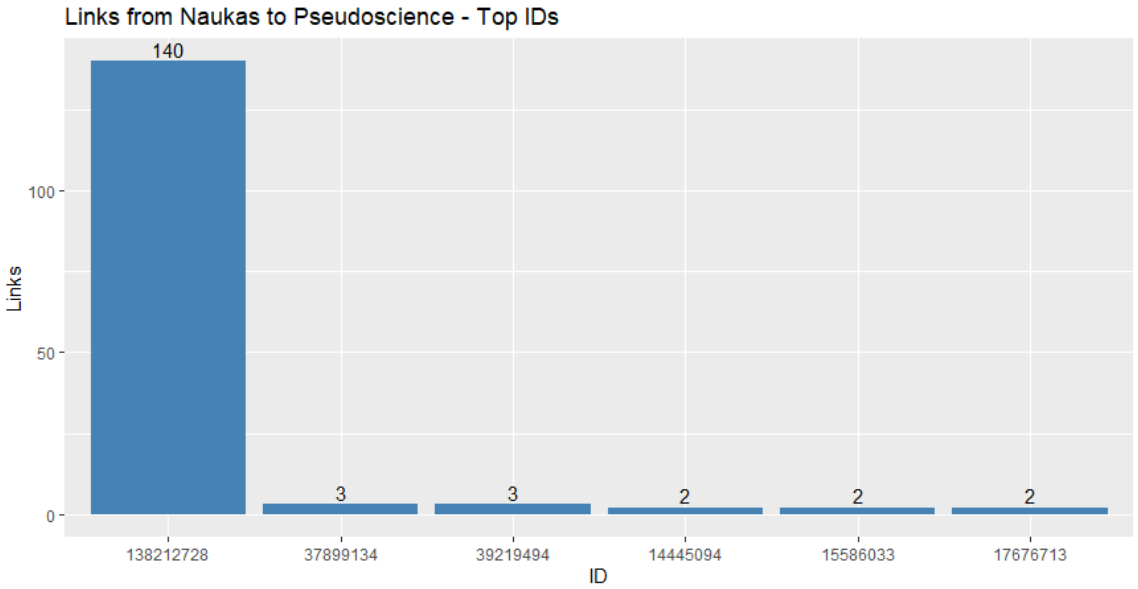

Figure 3.10 Links from Naukas to Pseudoscience

There seems to be only one user profile that connects well with the Naukas community with the Pseudoscience community, the other user profiles only reach 3 user profiles or less. This user profile belongs to Jorge Alberto Arreaza Montserrat, professor and Venezuelan politician, who has held various positions in the cabinet of President Hugo Chávez. Although he is not related with science or pseudoscience directly.

3.8 Pseudoscience to Naukas

Now, we are going to check which user profiles from the Pseudoscience network send information to the Naukas community.

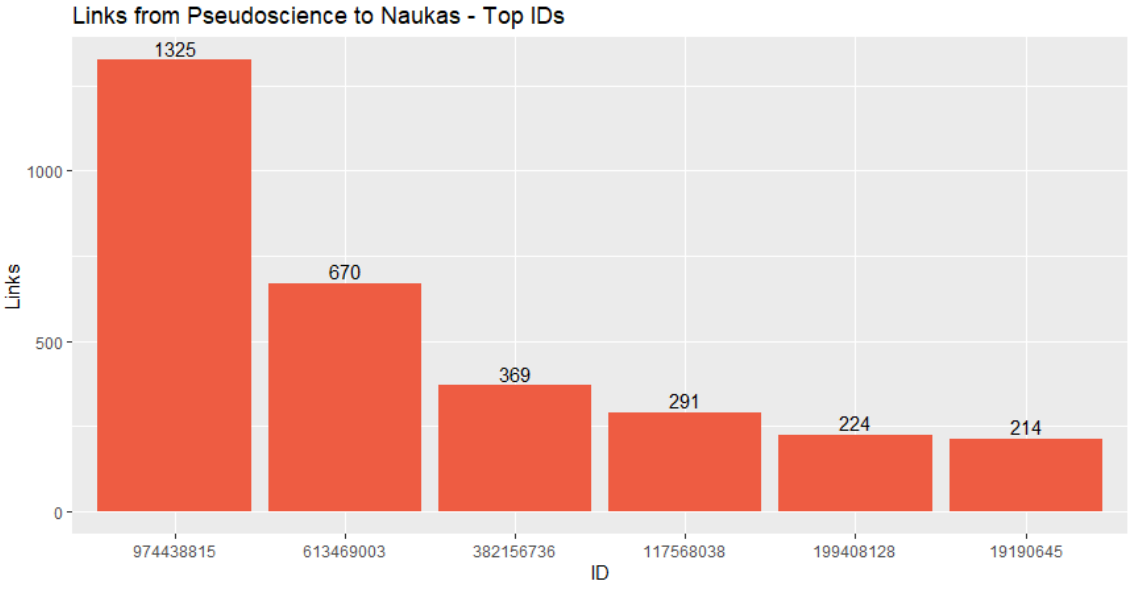

Figure 3.11 Links from Naukas to Pseudoscience

The first user profile and the second user profile seem to be the central nodes, especially the first user profile. The first user profile is Teresa Forcades, the second user profile is Caroline Criado. As we can observe, these 3 ‘hinge’ user profiles (1 from Naukas and 2

from Pseudoscience) are related with politics or social movements. The Naukas user profile is a Venezuelan politician. Teresa Forcades is related with the 'Proceso Constituyente' which is a Spanish social movement of Catalan scope, created to promote a change in the political, economic and social model of Catalonia. Caroline Criado campaigned for women experts to be better represented in the media.

The numbers differ a lot given that one of the communities is much bigger than the other one, so it is easier to reach more user profiles. Although this explanation is not enough, as there is practically one profile connecting Naukas to Pseudoscience. It seems that the profiles of the Naukas community are more aware of the Pseudoscience community than the other way around.

3.9 Do they share followers?

Although when we talk about the Naukas and the Pseudoscience communities, according to the results, it seems that we are talking about the same network. However, they do not appear to be separate networks, so, do they share audience? If so, how much? In order to do a comparison, we took one of the most important profiles of Pseudoscience, Teresa Forcades and the most important level 0 user profile of the Naukas network @aberron and we checked how many followers they share.

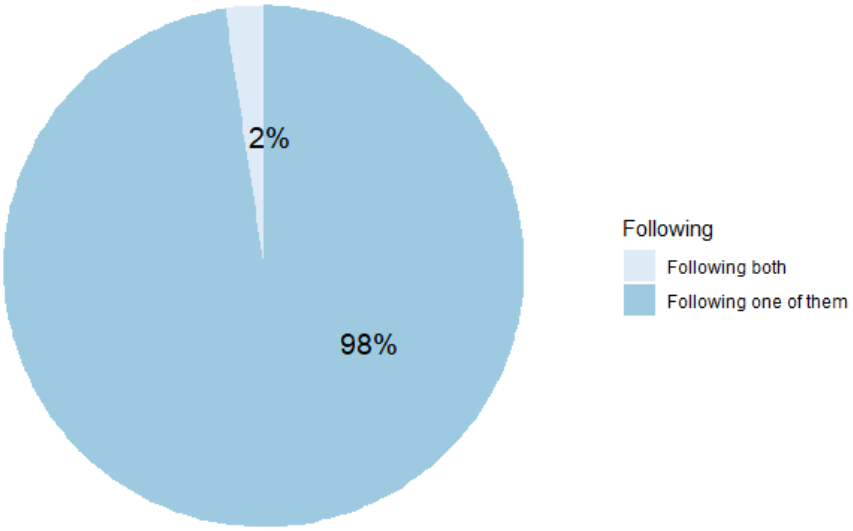

Figure 3.12 Followers shared between Teresa Forcades and @aberron

Only 80 out of the 3334 possible user profiles (2.4%), where 2066 followed @aberron and 1348 Teresa Forcades. These 80 user profiles were mainly: journalists, Pablo Echenique a Spanish politician and physicist, the newspaper 'Público' and average twitter user profiles.

Therefore, they do not seem to share audiences.

4. CONCLUSIONS AND FUTURE WORK

4.1 Conclusions

The intention was to find out if there was a connection between the scientific network and the 'pseudo-scientific' network, and the possibility of 'hinge' user profiles that act as a bridge between the two networks. In the analysis we could observe that both networks become blurred among other even larger networks to which they belong. Likewise, the 'hinge' user profiles did not appear to be related with science or pseudo-science.

The first step of this project was the filtering that was quite effective, nearly 90% of the users belonged to Spanish-speaking countries. Nevertheless, one of the communities was English-Speaking, this means that it was not a perfect filtering. Maintaining the verified profiles led to the formation of the English-Speaking community. This does not necessarily imply that every member of this community was English-speaking, rather that they followed many English-speaking profiles. This reflects the great impact that EEUU has globally.

The next step was the visualization process and the clustering process. The Pseudoscience community represented only 3% of the total network, demonstrating people's preference of science over pseudo-science. With the visualization, we were able to appreciate the small communities that were created by the level 2 user profiles of the Pseudoscience community. Some of these communities were removed given their minor weight in the total network, leaving us with nine main communities, being 99.85% of the total network.

In the analysis of the network, we obtained the main user profiles of the networks. In the scientific community, the level 0 user profiles, in this case two, only one of them was considered an important node of the network (according to some measures of centrality). The most important user profiles were journalists, media and popular science magazines. On the other hand, when we took the pseudoscience network, the most important user profiles were most of the level 0 user profiles, the user profiles with which the network was created. However, the pseudo-science profiles did not play an important role in the whole network, only in their own network, as being a level 0 user profile does not imply being a main node of the network.

Furthermore, the analysis of these main user profiles works as a tool to identify the background of each of the communities. The communities detected were mainly due to political and sociolinguistic issues. Most of the main profiles of each community were important journals or politicians of different countries. Twitter users are divided by social or political tendencies rather than science or pseudo-science beliefs. There were found several echo chambers according to these political trends, for instance, the Catalan case [6]. Likewise, more echo chambers were found according to sociolinguistic trends [7].

These nine communities are connected to each other, although the connections between two communities may vary depending on the relationships between them. Neighbouring countries had more connections than distant countries. Venezuela, for example, was very much related to Spanish politics, where we found that there was a great connection

between the Spanish right-wing network and the Venezuelan community. Therefore, the connections are also due to cultural or political tendencies.

These connections result when a user profile is followed by both communities. These user profiles act as a ‘bridge’ of information between both communities (‘hinge’ profiles). All 'hinge' user profiles do not seem to be user profiles that polarise information, we refer to polarisation as the act of criticising this information or sharing it in a negative way. Most user profiles are politicians or influencers who are followed by both communities. However, the content that they share or are known for is not focused on science or pseudo-science.

In conclusion, the communities and the links are not related due to scientific beliefs, but rather due to the culture and political tendency. Moreover, the Pseudoscience network does not seem to be a closed network, although it appears to have a different audience than the science network, where the scientific audience is more aware of the Pseudoscience audience than the other way around.

4.2 Further work

Taking even more pseudoscience user profiles or bigger pseudoscience user profiles to achieve a better balance between the communities would be a good addition to the possible improvements.

Looking at the results, we see that there is a connection between both communities by user profiles not related to science or pseudoscience. With this result it would be interesting to make a similar project with the political tendencies in Spain, since pseudoscience, not being such a popular subject, diffuses among other much larger communities.

Another interesting way of approaching the work, would be studying specific tweets. In our project, we are only studying connections between the profiles but not their content distinctively. Studying the content could determine if the polarizing content reaches both communities and could detect this ‘hinge’ profiles we are looking for.

5. REFERENCES

[1] BORGATTI, Stephen P. Centrality and network flow. *Social networks*, 2005, vol. 27, no 1, p. 55-71.

[2] BOSTRÖM, Harry; RÖSSNER, Stephen. Quality of alternative medicine—complications and avoidable deaths. *International Journal for Quality in Health Care*, 1990, vol. 2, no 2, p. 111-117.

[3] CARRINGTON, Peter J.; SCOTT, John; WASSERMAN, Stanley (ed.). *Models and methods in social network analysis*. Cambridge university press, 2005.

[4] CINELLI, Matteo, et al. Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603*, 2020.

[5] COLLEONI, Elanor; ROZZA, Alessandro; ARVIDSSON, Adam. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, 2014, vol. 64, no 2, p. 317-332.

[6] DEL VALLE, Marc Esteve; BRAVO, Rosa Borge. Echo chambers in parliamentary twitter networks: The catalan case. *International journal of communication*, 2018, vol. 12, p. 21.

[7] DUSEJA, Nikita; JHAMTANI, Harsh. A sociolinguistic study of online echo chambers on twitter. En *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. 2019. p. 78-83.

[8] FERNÁNDEZ GÓMEZ, Jorge David; HERNÁNDEZ-SANTAOLALLA, Víctor; SANZ-MARCOS, Paloma. Influencers, marca personal e ideología política en Twitter. *Cuadernos. info*, 2018, no 42, p. 19-37.

[9] FORTUNATO, Santo; BARTHELEMY, Marc. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 2007, vol. 104, no 1, p. 36-41.

[10] JIN, Fang, et al. Misinformation propagation in the age of twitter. *Computer*, 2014, no 12, p. 90-94.

[11] KOUZY, Ramez, et al. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 2020, vol. 12, no 3.

[12] KUŠEN, Ema; STREMBECK, Mark. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, 2018, vol. 5, p. 37-50.

[13] LANDHERR, Andrea; FRIEDL, Bettina; HEIDEMANN, Julia. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2010, vol. 2, no 6, p. 371-385.

[14] OYEYEMI, Sunday Oluwafemi; GABARRON, Elia; WYNN, Rolf. Ebola, Twitter, and misinformation: a dangerous combination?. *Bmj*, 2014, vol. 349, p. g6178.

[15] POSADZKI, Paul; ALOTAIBI, Abdulellah; ERNST, Edzard. Adverse effects of homeopathy: a systematic review of published case reports and case series. *International journal of clinical practice*, 2012, vol. 66, no 12, p. 1178-1188.

[16] SCOTT, John. Social network analysis. *Sociology*, 1988, vol. 22, no 1, p. 109-127.

[17] USHER, Nikki; HOLCOMB, Jesse; LITTMAN, Justin. Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The international journal of press/politics*, 2018, vol. 23, no 3, p. 324-344.

[18] WHITE, Douglas R.; BORGATTI, Stephen P. Betweenness centrality measures for directed graphs. *Social networks*, 1994, vol. 16, no 4, p. 335-346.

A. APPENDIX

A.1 Whole network analysis

The first centrality measure that we are going to apply is the vertex degree. The total degree is going to tell us which is the vertex which has more links in the network. Every table is going to have the ids in the top row and the results in the lower row.

TABLE 1. TOTAL DEGREE OF THE NETWORK

IDs	49658852	15648827	19923515	121385551	115516167	111933542
Degree	24123	21358	18252	14607	13456	13352

The first user profile is Miguel Henrique Otero, a Venezuelan journalist and president of the journal 'El Nacional', a Venezuelan newspaper. The second user profile is the newspaper 'Muy Interesante'. Muy Interesante is a monthly magazine of popular science. Its contents range from biomedical sciences, technology and astrophysics. The third user profile is the ABC. The ABC is a Spanish journal. The fourth user profile is Europa Press. Europa Press is a Spanish news agency, it was founded in 1953 by members of Opus Dei and broadcasts mostly in Spanish for 24 hours. The fifth user profile is a user profile called @ifilosfia. His real name is Miguel Olmo, he is a mathematician who one day in November 2010 opened a Twitter user profile where he publishes philosophical quotes, riddles and phrases for reflection. The sixth user profile is Arturo Perez Reverte. Arturo is a Spanish journalist and writer, member of the Real Academia Española since 2003.

Each of these user profiles are level 1 user profiles, this is interesting because according to followers (degree) none of the level 0 user profiles are top 6. Most of the journals or newspapers have a big number of followers but they follow only a few user profiles, these user profiles are mainly other journals or journalists. The first difference that we can see between both degrees is the fifth user profile. The user profile ifilosofía is following 83 thousand of user profiles. The EresCurioso user profile is following 73 user profiles, but if we check them closely most of them are related with science. Although ifilosofia has a higher total degree, that is because he is following a lot of user profiles, so by chance some of them are in the network.

The next centrality measure is the closeness centrality.

TABLE 2. TOTAL CLOSENESS OF THE NETWORK

IDs	49658852	10274252	2260150651	19923515	15648827	127122971
Closeness	0.50	0.48	0.46	0.46	0.46	0.46

The top ten values lie close to 0.46 except the first two ones which are the ones that are going to be analysed.

The first user profile is the Miguel Henrique Otero user profile. The second user profile is @aberron user profile, he is one of the level 0 user profiles. His name is Antonio Martínez, he is a journalist and scientific disseminator.

The next centrality measure is the betweenness. This betweenness is not normalized because it is an estimated betweenness, the conventional betweenness could not be calculated (too computationally expensive).

TABLE 3. ESTIMATED BETWEENNESS OF THE NETWORK

IDs	**49658852**	**115516167**	**31090827**	**37836873**	**117568038**	**33923443**
Betweenness	66702767	27512913	17722371	16850247	15186261	14643607

The first user profile and the second user profile have significantly higher values than the rest of the user profiles. The first user profile is the Miguel Henrique Otero user profile. The second user profile is ifilosofia. These user profiles have a big number of followers.

User profiles with a big number of followers and followed usually are important nodes of the network according to many algorithms. Being a level 0 user profile does not mean to be one of the main nodes of the network according to the centrality measures. Many journals are within the most important nodes of the network.

A.2 Pseudoscience network analysis

The network has 5439 edges and 5183 nodes. We are going to analyse the Pseudoscience network in the same way as the whole network.

TABLE 4. TOTAL DEGREE OF THE PSEUDOSCIENCE NETWORK

IDs	613469003	143114599	399207129	19190645	117568038	48150548
Degree	982	893	340	236	179	155

The first two user profiles have a total and out degree significantly higher than the rest of the network. The first user profile is Caroline Criado, she is a British journalist and feminist activist. She is a level 1 user profile. The second user profile is Mundo Rosa, it is a digital newspaper about fashion, fitness and spirituality. It is a level 1 user profile.

The next centrality measure is the closeness of the network.

TABLE 5. TOTAL CLOSENESS OF THE PSEUDOSCIENCE NETWORK

IDs	117568038	974438815	966632480	143114599	279700787	169845949
Closeness	0.41	0.37	0.33	0.32	0.31	0.30

Values from the top 10 user profiles lie within 0.33 and 0.29, the first two user profiles have the most disparate values. The first user profile is the Planeta Holistico user profile, it is a level 0 user profile. This user profile also appeared during the whole network analysis. The second user profile is Teresa Forcades, she is a Spanish theologian and nun of the Order of St. Benedict, as well as a medical doctor. She is also a level 0 user profile.

This network is smaller, so we can compute the betweenness.

TABLE 6. BETWEENNESS OF THE PSEUDOSCIENCE NETWORK

IDs	117568038	974438815	613469003	143114599	399207129	382156736
Betweenness	0.012	0.007	0.003	0.002	0.001	0.001

The top 3 user profiles are PlanetaHolistico, Teresa Forcades and Caroline Criado.

The most important user profiles according to the centrality measures that appear in the Pseudoscience community are user profiles which are directly related to homeopathy or alternative therapies. Most of these user profiles are the level 0 user profiles. In addition, we can observe that most of the user profiles do not have a high number of followers except Caroline Criado who has 100 thousand (much less than other nodes of the Naukas community).

A.3 Cluster 0

The names of the clusters are the ones given by the Louvain's algorithm. This cluster has 54598 edges and 25331 vertices.

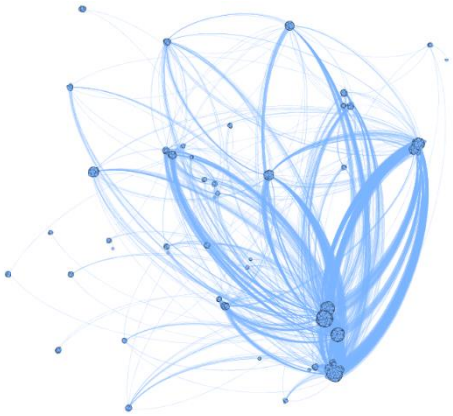

Figure 1. Cluster 0

The first centrality measure is the degree of the nodes.

TABLE 7. TOTAL DEGREE OF THE CLUSTER 0

IDs	**228609209**	**15648827**	**115516167**	**14445094**	**1586287452**	**62186451**
Degree	7812	4987	4227	2295	1407	1380

There seem to be 3 user profiles that have a big impact on this network by looking at the degrees, especially the first one. The first one is the EresCurioso user profile, Muy interesante is the second user profile and ifilosofia is the third user profile. The next four user profiles are Chilean user profiles from newspapers and brands.

The next measure is the closeness.

TABLE 8. CLOSENESS OF THE CLUSTER 0

IDs	**228609209**	**15648827**	**14445094**	**115516167**	**42020940**	**39137321**
Closeness	0.50	0.49	0.48	0.48	0.43	0.43

The top 4 user profiles seem to be the more relevant here. We have EresCurioso as the first one, Muy Interesante as the second one and ifilosofia as the fourth one. The third user profile is the user profile @el_ciudadano, it is a Chilean monthly newspaper.

The next centrality measure is the betweenness.

TABLE 9. BETWEENNESS OF THE CLUSTER 0

IDs	**115516167**	**14445094**	**62186451**	**1586287452**	**34121775**	**64460534**
Betweenness	0.0008	0.0007	0.0005	0.0005	0.0004	0.0002

The top two user profiles are ifilosofia y El Ciudadano. We can think that a great deal of the information handled in this cluster is scientific since the main user profiles are user profiles related to science and technology (EresCurioso and Muy Interesante). In addition, we can see that some of the main user profiles are Chilean user profiles.

A.4 Cluster 1

This cluster has 24061 edges and 21027 vertices.

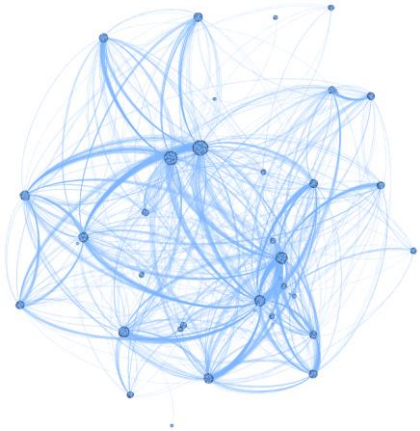

Figure 2. Cluster 1

The first centrality measure is the degree of the nodes.

TABLE 10. VERTEX DEGREE OF THE CLUSTER 1

IDs	**37836873**	**33274835**	**16817883**	**18655567**	**613469003**	**18363508**
Degree	6000	3830	2132	1970	1482	1125

The first user profile belongs to Lil B an American rapper. The second user profile belongs to Eric Schiffer who is a successful American entrepreneur. The third user profile belongs to Science Friday, an American user profile of science. Each of these user profiles are level 1 user profiles.

The next centrality measure is the closeness.

TABLE 11. CLOSENESS OF THE CLUSTER 1

IDs	**37836873**	**33274835**	**297882633**	**162535413**	**938657858**	**16817883**
Closeness	0.54	0.46	0.46	0.43	0.43	0.43

The first user profile seems to be substantially more relevant than the other user profiles. He is Lil B again. The next 9 user profiles have approximately the same value. All these user profiles are American user profiles and most of them are related with Science.

The next centrality measure is the betweenness.

TABLE 12. BETWEENNESS OF THE CLUSTER 1

IDs	**37836873**	**33274835**	**2307675307**	**938657858**	**18655567**	**297882633**
Betweenness	6.15e-04	2.64e-04	1.25e-04	1.23e-04	7.78e-05	7.07e-05

The top 2 user profiles seem to be appreciably more relevant than the other user profiles. The first one is the Lil B user profile and the second one is the Eric Shiffer user profile.

This cluster seems to encompass all the people in the US network. Many of the user profiles in Latin America might follow people from the United States. The most important nodes are verified user profiles, which passed the filter automatically. Some like Lil B are related to music, but most are related to data science or science itself.

A.5 Cluster 2

This cluster has 148203 edges and 23598 vertices.

Figure 3. Cluster 2

The first centrality measure is the degree of the nodes.

TABLE 13. VERTEX DEGREE OF THE CLUSTER 2

IDs	82388403	872086682	14306573	243608028	253499949	14457019
Degree	6000	3830	2132	1970	1482	1125

According to the total degree the first user profile stands out from the rest. It is the Alex Riviero he is a scientific disseminator. Considering the out degree, the top 3 user profiles are the most relevant. The second user profile is @lavecinarubia, she is a Spanish influencer. The third user profile is Berto Romero, he is a Spanish comedian. These user profiles are level 1 user profiles.

The next centrality measure is the closeness.

TABLE 14. CLOSENESS OF THE CLUSTER 2

IDs	10274252	17185251	213264156	243608028	208074928	82388403
Closeness	0.50	0.47	0.46	0.45	0.45	0.45

The first user profile is @aberron user profile, one of the level 0 user profiles from Naukas' community. The second user profile is Pepo Jimenez, he is a Spanish influencer. The third user profile is Raquel Sastre, she is a Spanish comedian and twitter influencer. The fourth user profile belongs to Jot Down, it is a Spanish cultural magazine.

The next centrality measure is the betweenness.

TABLE 15. BETWEENNESS OF THE CLUSTER 2

IDs	82388403	213264156	872086682	108099486	17185251	10274252
Betweenness	0.0028	0.0020	0.0019	0.0016	0.0015	0.0012

The top 3 user profiles according to the betweenness are Alex Riviero, Raquel Sastre and @lavecinarubia.

This community is mainly related to the Spanish twitter community. Some Spanish scientific user profiles appear, such as Alex Riviero or @aberron. On the other hand, there are also several Spanish comedians and influencers.

A.6 Cluster 3

This cluster has 10947 nodes and 61298 edges.

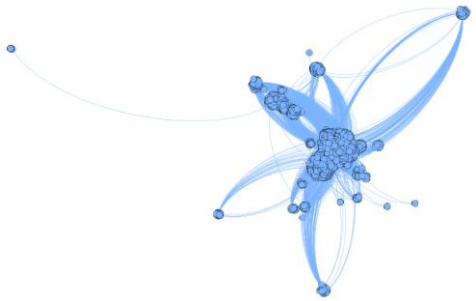

Figure 4. Cluster 3

The first centrality measure is the degree of the nodes.

TABLE 16. TOTAL DEGREE OF THE CLUSTER 3

IDs	44992826	18939115	42046651	145580132	11231412	3345981
Degree	2030	1379	1364	1259	1097	1073

Most of the top user profiles have a similar total degree and out degree except the top 1 user profile that stands out a little. This user profile belongs to Movistar España, it is an operator of the Spanish multinational telecommunications company Telefónica. It is a level 1 user profile.

TABLE 17. TOTAL CLOSENESS OF THE CLUSTER 3

IDs	44992826	145580132	8075962	1638691	731573	11231412
Closeness	0.495	0.492	0.480	0.478	0.474	0.473

According to the total closeness, the top two user profiles are the one that stand out above the rest. The first one is Movistar España again. The second one is TodoStartUps.

The next centrality measure is the betweenness.

TABLE 18. BETWEENNESS OF THE CLUSTER 3

IDs	44992826	145580132	42046651	17019192	1638691	18939115
Betweenness	0.003	0.002	0.00198	0.0019	0.0015	0.0015

The first user profile here is the one that stands out, it is the Movistar España user profile.

As we can observe most of the user profiles that we obtained in this cluster are Spanish user profiles, although all these user profiles with are related to trademarks and companies. It seems that the business cluster has its own community, and all the entrepreneurs and telecommunications companies tend to cluster together.

A.7 Cluster 4

This cluster has 18799 nodes and 433620 edges.

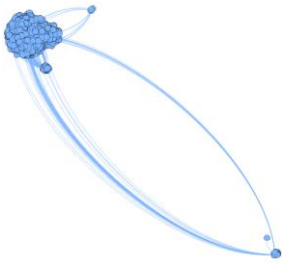

Figure 5. Cluster 4

Degree and closeness results are nearly the same.

TABLE 19. TOTAL DEGREE OF THE CLUSTER 4

IDs	14436317	535707261	17676713	11904592	20909329	76593802
Degree	7296	7052	6818	6730	6246	5083

TABLE 20. CLOSENESS OF THE CLUSTER 4

IDs	14436317	535707261	17676713	11904592	20909329	107153756
Closeness	0.61	0.60	0.60	0.59	0.58	0.56

Every user profile in the top 6 has a similar value, although the 6th one seems to be less relevant. The first user profile belongs to Ignacio Escolar, he is a Spanish journalist. He is the founder and director of the digital newspaper eldiario.es. The second user profile is el Diario, it is a Spanish digital newspaper. The third user profile is @Publico_es, it is a digital newspaper in Spain, published in Spanish and belonging to Display Connectors. The fourth user profile belongs to Alberto Garzón, a Spanish economist and politician.

The next centrality measure is the betweenness.

TABLE 21. BETWEENNESS OF THE CLUSTER 4

IDs	76593802	14824411	425119821	15183391	14436317	17186899
Betweenness	0.0047	0.0024	0.0017	0.0016	0.0014	0.0011

The first user profile is @caval100, he is a blogger, he is left-wing. The second user profile is Izquierda Unidad, it is a Spanish political and social movement formed in 1986. They are left-wing. Both are level 1 user profiles.

Most of the user profiles found in this cluster is a Spanish user profile, they are mostly left-wing. The most important nodes are usually journals like ‘El Diario’ which is known to be a left-wing journal, in addition his owner also appears (Ignacio Escolar). Furthermore, all the journalists and the politicians are related with left-wing political parties.

A.8 Cluster 6

This cluster has 15806 nodes and 113692 edges.

Figure 6. Cluster 6

The first centrality measure is the degree of the nodes.

TABLE 22. TOTAL DEGREE OF THE CLUSTER 6

IDs	19923515	111933542	121385551	343447873	31090827	155629354
Degree	5884	4877	4866	4230	3502	2673

Most of them seem relevant, so we are going to check the top 4 user profiles. The first user profile is the ABC, it is a Spanish newspaper. The second user profile is Arturo Perez Reverte. The third user profile is EuropaPress. The fourth user profile belongs to Mariano Rajoy, he is a Spanish politician of the Partido Popular, sixth president of the Spanish government. They are level 1 user profiles.

The next centrality measure is the closeness.

TABLE 23. TOTAL CLOSENESS OF THE CLUSTER 6

IDs	19923515	111933542	121385551	343447873	31090827	155629354
Closeness	0.593	0.559	0.559	0.522	0.511	0.510

This ranking is the same one as the vertex degree ranking.

The next centrality measure is the betweenness.

TABLE 24. BETWEENNESS OF THE CLUSTER 6

IDs	19923515	31090827	155629354	343447873	146150851	208911269
Betweenness	0.00196	0.00156	0.00151	0.00135	0.00126	0.00092

Every ranking seems similar, the new addition to this ranking is Juanfran Escudero, former councillor of the political party Ciudadanos. Another addition is the ‘20minutos’, it is a Spanish newspaper. Both are level 1 user profiles.

According to this cluster we can observe that most of the politicians or journals that appear here are right-wing. They include the right-wing of Spanish politics. We can observe two networks related with politics in Spain.

A.9 Cluster 7

This cluster has 25374 nodes and 51709 edges.

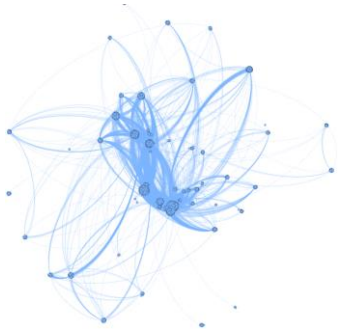

Figure 7. Cluster 7

The first centrality measure is the degree of the nodes.

TABLE 25. TOTAL DEGREE OF THE CLUSTER 7

IDs	152325528	138212728	183997763	33923443	142040148	76133133
Degree	6783	4610	4495	3185	2288	2280

The top 3 user profiles stand out above the rest. The first user profile belongs to C5N, it is an Argentine open and subscription television news channel. The second user profile belongs to Jorge Arreaza, a Venezuelan politician. The third user profile belongs to Marcelo Parrilli, an Argentinean lawyer. They are level 1 user profiles.

TABLE 26. TOTAL CLOSENESS OF THE CLUSTER 7

IDs	183997763	33923443	152325528	2260150651	76133133	30500021
Closeness	0.505	0.497	0.492	0.457	0.451	0.447

According to the total closeness we have a clear top 3. This top 3 corresponds to, Marcelo Parrilli, @6BillionPeople, he is an American influencer and the C5N.

The next centrality measure is the betweenness.

TABLE 27. BETWEENNESS OF THE CLUSTER 7

IDs	183997763	33923443	152325528	76133133	138212728	744150240136044544
Betweenness	0.00077	0.00054	0.00047	0.00028	0.000267	0.000200

This ranking is pretty like the total closeness ranking. The top 3 is the same and they are the most relevant user profiles.

Almost all the user profiles belonging to this cluster are from Argentina, there are many user profiles belonging to Venezuela. Within these user profiles we can see Argentinean actors, Argentinean TV channels and some Venezuelan politicians. This cluster seems to represent the Argentinean community mainly and little of the Venezuelan one.

A.10 Cluster 8

This cluster has 38377 nodes and 152702 edges.

Figure 8. Cluster 8

The first centrality measure is the degree of the nodes.

TABLE 28. TOTAL DEGREE OF THE CLUSTER 8

IDs	49658852	806043	147589125	42832810	53480462	42888442
Degree	16077	7516	7495	6549	5515	5017

There is one user profile that is more remarkable than the others, it is the top one user profile, it was an important node in the whole network. This user profile belongs to Miguel Henrique Otero, he is a Venezuelan journalist.

TABLE 29. TOTAL CLOSENESS OF THE CLUSTER 8

IDs	49658852	806043	127122971	15973392	147589125	42832810
Degree	0.6017594	0.4885862	0.4876239	0.4829661	0.4815600	0.4802222

The ranking is similar to the vertex degree one, this one includes two user profiles. One of the user profiles included is Emilio Gómez Islas, he is the director of ATV Latino (a news channel). The other user profile is TalCual a Venezuelan newspaper.

The next centrality measure is the betweenness.

TABLE 30. BETWEENNESS OF THE CLUSTER 8

IDs	49658852	127122971	117568038	806043	15973392	15973392
Betweenness	0.00354	0.00118	0.00109	0.00048	0.00042	0.00042

The top 3 user profiles have values significantly higher than the others, especially the first one. The first one is Miguel Enrique Otero, the second one is Emilio Gómez and the third one is Planeta Holístico.

This cluster seems to accumulate the other part of the Venezuelan community. It seems to be more related to politics than the previous one, all this is due to Miguel Hotero as he is the maximum influence of the whole network.

A.11 Cluster 9

This cluster has 6392 nodes and 17591 edges.

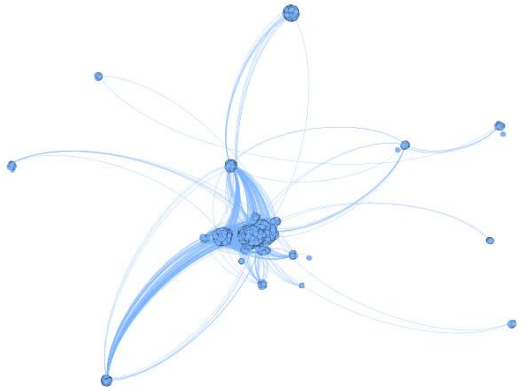

Figure 9. Cluster 9

The first centrality measure is the degree of the nodes.

TABLE 31. TOTAL DEGREE OF THE CLUSTER 9

IDs	47753979	346510738	44186827	97100000	62796429	850339922
Degree	1526	1427	1156	1105	1042	784

We see that all user profiles have a similar degree. The first user profile is Vanessa de la Torre, she is a Colombian journalist. The second user profile Artun Duanga, his user profile is private, although he is related to a radio station in Colombia. The third user profile is Telemedellin, it is a local Colombian open television channel. The fourth user profile is El Nuevo Siglo, a Colombian daily newspaper. Each of these user profiles is a level 1 user profile.

The next centrality measure is the closeness.

TABLE 32. TOTAL CLOSENESS OF THE CLUSTER 9

IDs	346510738	44186827	47753979	18786579	38373082	2157729691
Closeness	0.45	0.45	0.45	0.44	0.44	0.42

The first user profile, it is the user profile of Artur Duanga. The second one is Telemedellin and the third one is Vanessa de la Torre.

TABLE 33. BETWEENNESS OF THE CLUSTER 9

IDs	346510738	18786579	44186827	38373082	62796429	382157490
Betweenness	0.0014	0.0009	0.0007	0.0006	0.0006	0.0005

The first user profile stands out from the others, it is the user profile of Artur Duanga.

As we can see, this community consists mainly of Colombian people. We can find many Colombian newspapers and journalists.