ALMA MATER STUDIORIUM
UNIVERSITÀ DI BOLOGNA
Department of Physics and Astronomy
Laurea Magistrale in Applied Physics

COMPLEX NETWORKS

Studying a network composed of tweets in the social media Twitter regarding the vaccination for COVID-19

Ferraioli Alessio Giuseppe, Cuticchia Giancarlo

Bologna, July 2022

Abstract
In this project, we build a network with tweets from the social network Twitter. The tweets concern the topic of Covid-19 vaccination in Italy in 2020-2021. The network is built by considering any retweet as a link between two users. We analyze the characteristics of this network and examine how three different groups of users, AntiVax, ProVax, and Neutrals, behave on this network. By measuring typical characteristics of a network such as degree, closeness, and betweenness, we note that this network tends to behave as it is expected for a social network. It follows typical distributions of centralities, showing the property of being "scale-free", which is common for social networks. We see that most "celebrities" are AntiVax users, who both get retweeted the most and by the largest number of users and both retweet the most. Also, we take a look at the community structure of the network with the Leiden algorithm. We note that being labeled AntiVax, ProVax or Neutrals is not uncorrelated with the membership in the community. We note that AntiVax users and ProVax users both mix with Neutral users but they do not mix with each other.
Keywords: antivax, betweenness, centrality, closeness, communities, degree, covid-19, links, network, provax, retweet, social media, tweet, Twitter, vaccination.

---

1 Introduction

Social network analysis employs mathematical structures to build and study graphs made of nodes (also called vertexes), representing people, and links (also called edges), representing a connection between people. This helps to represent potentially a very large amount of data and to describe the overall social structure. Analyzing the network means identifying characteristics of a social structure, such as finding the most influential people in a group, locating different groups of users (communities), and observing how they interact with each other, by keeping track of how information flows from person to person.

Social network analysis is a useful tool to gather information about our increasingly connected world. An enormous volume of data is available due to online social media networks such as Facebook or Twitter. In this work, we will analyze data from the social media website Twitter, concerning the topic of Covid-19 vaccines, and use the techniques of network analysis to take a look at the users posting comments (called "tweets") in order to study the connections between them and the groups they form.

2 The data

The database in question is a recollection of different information from over 2 million tweets, which was collected during late 2020 and early 2021. The dataset is composed of 2,067,483 rows and 12 columns, one row for each tweet and each column providing different information about the corresponding tweet. The tweets regard the topic of COVID-19 vaccination in Italy.

The database was provided by prof. Daniel Remondini and PhD Francesco Durazzi, from the Dipartimento di Fisica e Astronomia "Augusto Righi" at the Università di Bologna (Italy). Each row of the dataset corresponds to the following:

1. **created_at**: if it's a tweet, the time when it was created; if it's a retweet, the time when it was made.
2. **id**: an identification number within Twitter for the tweet or the retweet.
3. **text**: the content of the tweet.
4. **user.id**: if it's a tweet, an identification number within Twitter for the user who made the tweet; if it's a retweet, for the user who made the retweet.
5. **user.screen_name**: if it's a tweet, the username within Twitter of the user who made the tweet; if it's a retweet, from the user who made it.
6. **place**: the location from which the tweet was made.
7. **url**: any link to other websites contained within the text of the tweet.
8. **retweeted_status.id**: if the tweet is a retweet, the identification number within Twitter of the original tweet. If the tweet is not a retweet, this value is left empty.
9. **retweeted_status.user.id**: if the tweet is a retweet, the identification number within Twitter of the user who made the original tweet. If the tweet is not a retweet, this value is left empty.
10. **retweeted_status.url**: if the tweet is a retweet, the url address to the original tweet. If the tweet is not a retweet, this value is left empty.
11. **annotation**: a label of "AntiVax", "Neutral" or "ProVax" to the content of the tweet, regarding its position to the COVID-19 vaccine.
12. **user_annotation**: a label of "AntiVax", "Neutral" or "ProVax" to the user who made the tweet or the retweet, regarding its position to the COVID-10 vaccine.

3 Preprocessing the database to build a network

The aim is to analyze the network structure composed of the users posting these tweets. To build a network, each node will represent a user and each user is linked to another via a retweet. There is a directed edge from node A to node B if user A retweeted user B; if user A retweeted user B multiple times, the weight of the edge will be taken as the number of retweets.

Note that any tweet that is not a retweet would not provide an edge in the network, because it would not link the author of the tweet to any other. For this reason, all the tweets that are not retweets were ignored, since they don't provide information about the connection between the users in the network. For this reason, all the rows from the

data that have an empty value under "retweeted_status.id" were ignored, since such rows would correspond to a tweet that is not a retweet.

Additionally, each node will be given a label, or polarity, depending if the user is in favor of the vaccination for COVID-19 (ProVax), against it (AntiVax) or neutral (Neutral). Rows of the data for which there was no value of this label or polarity for the user who made the tweet were also ignored[1].

The database was reduced to 151,508 rows this way.

3.1 Removing low-weight edges

In order to improve the network analysis and visualization, only links with a weight of 2 or bigger were kept. This criterion is taken under the statement that a link (or edge) of a weight less than at least 2 would correspond to a single instance of interaction between two users, so won't carry much information but only contribute to noise within it. So, these low-weighted edges were safely discarded. By doing so, the dataset was reduced to 96,424 rows. Getting rid of low-weight edges also reduces the relative quantity of unlabeled users: the labeled users go from being $\sim 20\%$ of the total to being $\sim 37\%$ of the total. This could be due to the fact that a higher weight means a higher number of tweets and so a larger quantity of information that leads to the assignment of the polarity to the user. (Remember that a user gets a polarity assigned if it has retweeted one or more tweets that have been labeled after individual examination by an operator).

3.2 Weakly and strongly connected components

By using the free library NetworkX for Python [2], it's possible to build a weighted directed network from the dataset described before, in which each user is a node and there is a link from user A to user B if user A retweeted user B. The weight of the link corresponds to the number of times user A retweeted user B. What we get is a directed graph made of 3,549 nodes and 20,798 edges connecting them.

We can now take a look at the connected components within the network. A component is simply a maximal subset of nodes from the network where each node is reachable from any other node within the same subset. The component is weakly connected if we are not taking into consideration the direction of the edges, and strongly connected if we do. [7]

The network consists of 60 weakly-connected components and 2,856 strongly-connected components. The biggest weakly-connected component contains 3,444 nodes (around 97% of the nodes), while the second largest weakly-connected component only contains 6 nodes.

[1]It must be noted that, even after removing rows with unlabeled users, there are still nodes within the network without a label of polarity. This is because the information on the column "user_annotation" corresponds to the user who performed the retweet, not the one who was retweeted. So, if user A retweets a tweet from user B, this interaction will be a row within the dataset from which we know the polarity of user A, but in that row is not provided the polarity of user B. If the information regarding the polarity of user B is not provided in any other row of the dataset, then it will be missing. When we discard all the rows with unlabeled users, we are discarding almost always rows in which neither of the users was labeled; such a row would correspond to a part of the network we are not interested in since it involved only unlabeled users. On the contrary, if we were now to discard all the remaining unlabeled nodes, which by construction are always connected to at least one labeled node, we would significantly alter part of the network we are actually interested in. For this reason, we will keep these unlabeled users for our analysis. Still, from unlabeled users being more than 99% of all nodes, we get to them being about 80% of all nodes.

This comes as no surprise since real-world social networks usually show this feature of possessing a single very large component and several very small ones ([7], Par. 8.1). Since the other components won't provide much information about the network, they are easily discarded, and only the biggest weakly-connected component will be considered for the analysis in the next sections.

The biggest strongly-connected component contains 641 nodes (around 18% of the nodes), while the second largest strongly-connected component only contains 6 nodes. As we might expect, the largest strongly connected component is considerably smaller than the largest weakly connected component.

3.3 The Network

In conclusion, the network built and to be analyzed is composed of a single component made of 3,444 nodes and 20,702 edges connecting them. Of these nodes, 488 of them are labeled as AntiVax, 230 are labeled as ProVax, 712 are labeled as Neutral, and the rest (2,014 nodes) are unlabeled.

4 Centrality measures

We can start analyzing the network by measuring some of its centrality measures. The centrality measures are quantities that give an indication of the importance of each node in the network. There are many different points of view to define "importance", therefore there could be many centrality measures, each with its own definition of importance.

4.1 Degree centrality

The simplest centrality measure is the degree centrality, which is just the number of other nodes a node is connected to. In Figure 1, we can see the histogram of degrees over the whole network.

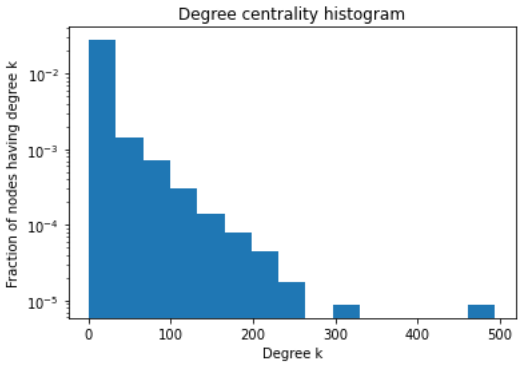

Figure 1: Histogram of the degree centralities of the network.

From the histogram, we note that the degree centrality follows a right-skewed distribution. This means that there are a small number of nodes with a very high degree, while most of the nodes have a small degree. This is to be expected in most real networks. [7]

We could also plot the degree distribution of the network in a histogram with logarithmic binning, as in Figure 2. A histogram with logarithmic binning is a histogram on a log-log scale in which the size of each bin is equal to the size of the previous bin times a

constant factor ([7], Par. 8.4.1). The bins' heights are normalized to account for the fact that they have different widths.

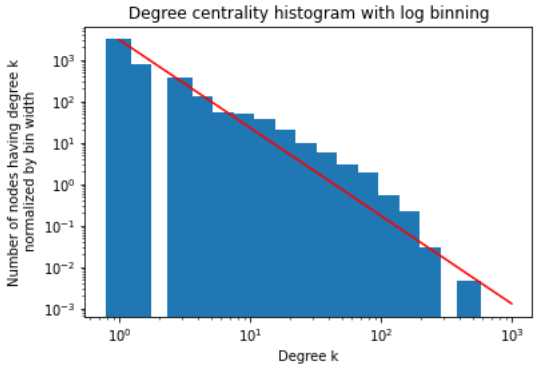

Figure 2: Histogram with logarithmic binning of the degree centralities.

We can also take a look at the cumulative degree centrality, pictured in Figure 3. The fact that the degree distribution in the log binned histogram and the cumulative degree distribution follows roughly a straight line, might suggest that the degree distribution of this network approximately follows a power law. [7] [2] Networks with such properties are called scale-free networks and tend to have a structure consisting of a central core that contains most of the nodes, outside of which there are "tendrils" of more peripheral nodes. In a scale-free network, it is more common to find nodes of a much larger degree than the other, as opposed to a random network. These special nodes, called "hubs", act as celebrities in the network, being much more connected than most other nodes. [7]

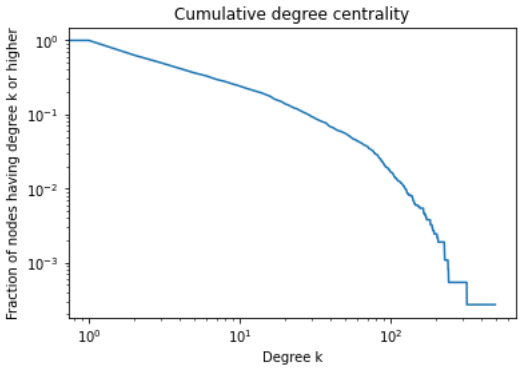

Figure 3: Cumulative degree centrality.

In Figure 4 we can see a histogram similar to the one shown in Figure 2, but this time showing it for each different polarity group instead of the whole network. We can see that each of the groups has a distribution of similar shape.

[2]Note that we ignore the behavior at very low or very high degrees because the power-law relation usually does not hold at the extremes. Also, since the bins are not independent from each other, the red line plotted over the histogram is not directly fitted, but it is evaluated analytically with the formula: $\alpha = 1 + N\left[\sum_i \ln \frac{k_i}{k_{min}-\frac{1}{2}}\right]^{-1}$ ([7], Par. 8.4). In this case, we get $\alpha \approx 2.12$, which is a typical value for a social network, for which α usually ranges between 2 and 3 ([7], Par. 8.4).

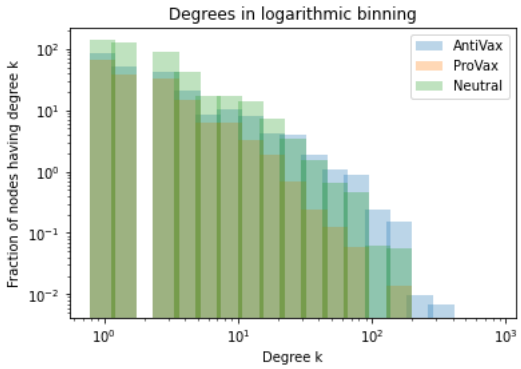

Figure 4: Histogram with logarithmic binning of the degree centralities of the network with the three user groups highlighted.

4.2 In-degree and out-degree centralities

In our network, we have information on the direction and weight of the edges. This means that we could take into consideration the measure of a weighted degree, in which the degree of a node is not just the number of nodes it is connected to, but the sum of the weights of the edges it is linked to. Moreover, we can measure two kinds of directed degree: inward degree which counts only the edges that link to a node, or outward degree which counts only the edges starting from the node.

What do these different degree measures mean in our network? Remember that each node of our network represents a Twitter user and a weight-one edge from user A to user B represents a tweet in which user A reposted user B. Therefore, the unweighted outward degree of user A is the number of users it has retweeted, while the unweighted inward degree is the number of users that retweeted user A. On the other hand, the weighted outward degree is the total number of tweets user A has retweeted, while the weighted inward degree is the total number of tweets that reposted user A. It could be interesting to see to which of the three groups ("AntiVax", "ProVax", "Neutral") these high-centrality users belong.

In the following, we will show tables with the information regarding the top 5 highest outward and inward degrees within the network, both for the weighted and unweighted case. Each table is accompanied by its corresponding histogram with logarithmic binning.

Table 1 shows the users with the highest weighted outward degree, which are the users that retweeted the highest number of tweets; the corresponding histogram in logarithmic binning is shown in Figure 5. Table 2 shows the users with the highest weighted inward degree, which are the users that were retweeted by the highest number of tweets; the corresponding histogram in logarithmic binning is shown in Figure 6

Table 3 shows the users with the highest unweighted outward degree, that are the users that retweeted the highest number of different users; the corresponding histogram in logarithmic binning is shown in Figure 7. Table 4 shows the users with the highest unweighted inward degree, that are the users that were retweeted by the highest number of different users; the corresponding histogram in logarithmic binning is shown in Figure 8.

Highest Weighted Outward Degree

	polarity	degree k
1	AntiVax	3158
2	AntiVax	1136
3	AntiVax	1021
4	N/A	1001
5	AntiVax	972
The first AntiVax user is at position:	1	with k = 3158
The first ProVax user is at position:	23	with k = 615
The first Neutral user is at position:	9	with k = 814

Table 1: Polarity and degree of the users with highest weighted outward degree. These are the users who retweeted the highest number of tweets.

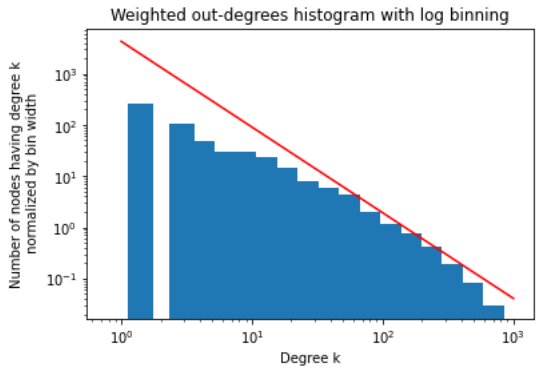

Figure 5: Histogram with logarithmic binning of the weighted outward degree centralities.

Highest Weighted Inward Degree

	polarity	degree k
1	AntiVax	1830
2	AntiVax	1742
3	AntiVax	1652
4	Neutral	1500
5	ProVax	1041
The first AntiVax user is at position:	1	with k = 1830
The first ProVax user is at position:	5	with k = 1041
The first Neutral user is at position:	4	with k = 1500

Table 2: Polarity and degree of the users with the highest weighted inward degree. These are the users who have been retweeted by the highest number of tweets.

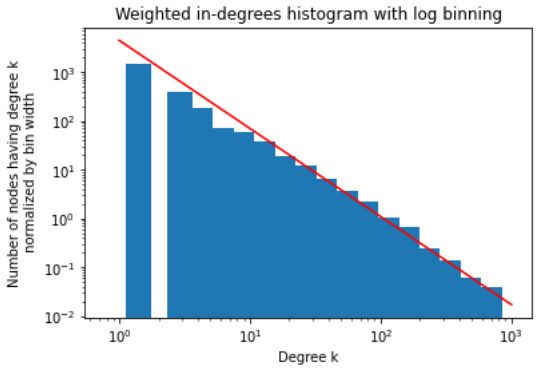

Figure 6: Histogram with logarithmic binning of the weighted inward degree centralities.

Highest Unweighted Outward Degree

	polarity	degree k
1	AntiVax	429
2	N/A	203
3	AntiVax	190
4	AntiVax	189
5	AntiVax	183
The first AntiVax user is at position:	1	with k = 429
The first ProVax user is at position:	24	with k = 107
The first Neutral user is at position:	8	with k = 166

Table 3: Polarity and degree of the users with the highest unweighted outward degree. These are the users who retweeted the highest number of different users.

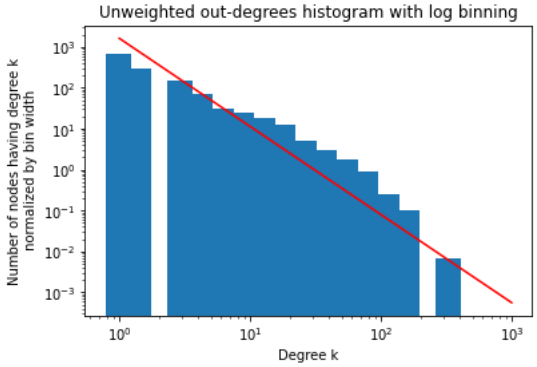

Figure 7: Histogram with logarithmic binning of the unweighted outward degree centralities.

Highest Unweighted Inward Degree

	polarity	degree k
1	AntiVax	213
2	N/A	191
3	AntiVax	186
4	Neutral	179
5	AntiVax	169
The first AntiVax user is at position:	1	with k = 213
The first ProVax user is at position:	6	with k = 164
The first Neutral user is at position:	4	with k = 179

Table 4: Polarity and degree of the users with the highest unweighted inward degree. These are the users who have been retweeted by the highest number of different users.

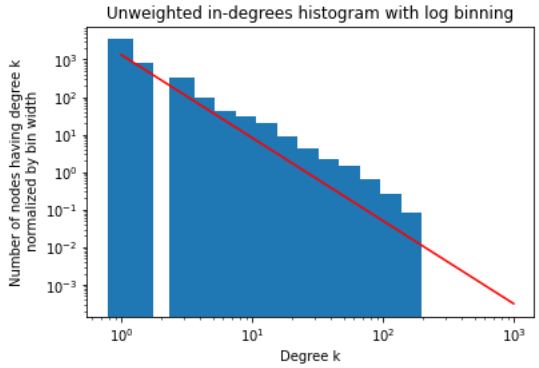

Figure 8: Histogram with logarithmic binning of the unweighted inward degree centralities.

The fact that unlabeled users are under-represented in the highest degrees, while they make up more than half of the network, could be due to the construction of the network. Since the polarity label is assigned when a user retweets a labeled tweet, it is more likely that a high-degree user would be labeled. So it could be expected that unlabeled users have on average a lower degree.

Still, among the labeled users the AntiVax polarity is over-represented in the highest degrees even though they are just one-third of the labeled users (while Neutrals are half of all the labeled users). This could mean that AntiVax users tend to be the more active, retweeting more tweets, but also the ones who get retweeted more. These two aspects could be as well be correlated: it could be that AntiVax users are retweeting mostly other AntiVax users. In Section 6.2, we will discuss how the different groups mix and form communities.

We can also compare out-degree and in-degree by plotting them on a 2D graph and highlighting the three groups with different colors. In Fig. 9, 10, 12, and 11 we can see the plots for unweighted out-degree and in-degree, while in Fig. 13, 14, 16, and 15 we can see the plots for weighted out-degree and in-degree. As before, we choose to examine both unweighted and weighted measures because they refer to different aspects of the data: unweighted measures depend on the number of users a node is connected to, while weighted measures depend on the number of tweets. We could expect that the connectivity of a node depends also on the total number of users/tweets of the same group. For this reason, in these plots we normalized the unweighted measures by the total count of users belonging to the same group, while for weighted measures by the total weight

of users of the group. Looking at the plots, we don't notice a clear distinction between users of the three groups. Examining the graphs, we could see that Neutral users tend to be more crammed near the origin, meaning that they make on average fewer connections than the others, while Antivax users are the most popular nodes in the top right areas of the graph, meaning that they tend to be the users that make more connection in both ways (by tweeting and being retweeted a lot). This feature of AntiVax users is slightly more apparent in the plots for unweighted degrees. Still, there is no obvious separation between the three groups, because for the most part we have areas crammed with all three types of users. In Section 4.5, we will employ principal component analysis (PCA) to highlight the variation among users in a clearer and more insightful way.

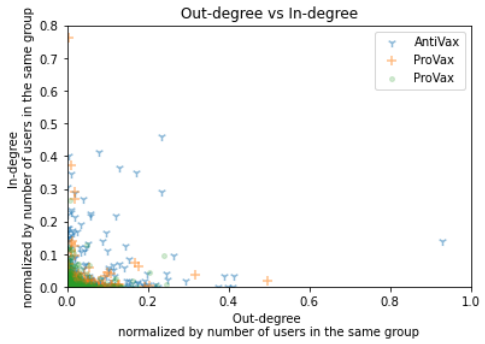

Figure 9: Plot of unweighted out-degree vs in-degree with colors highlighting the three groups.

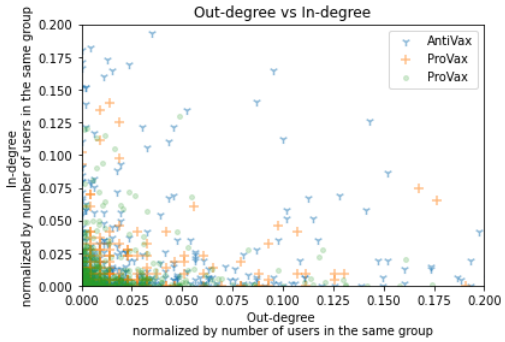

Figure 10: Zoomed-in plot of unweighted out-degree vs in-degree with colors highlighting the three groups.

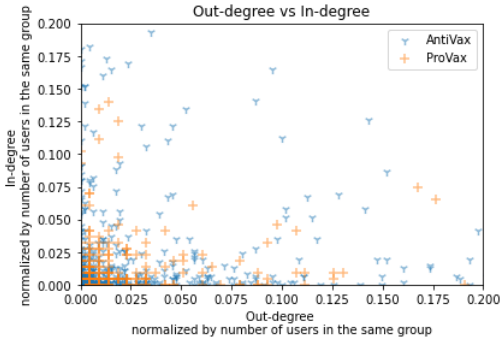

Figure 11: Plot of unweighted out-degree vs in-degree for AntiVax and ProVax users.

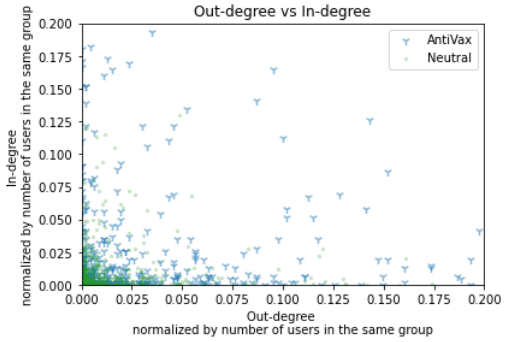

Figure 12: Plot of unweighted out-degree vs in-degree for AntiVax and Neutral users.

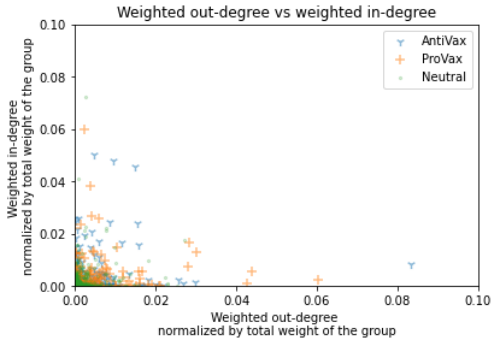

Figure 13: Plot of weighted out-degree vs in-degree with colors highlighting the three groups.

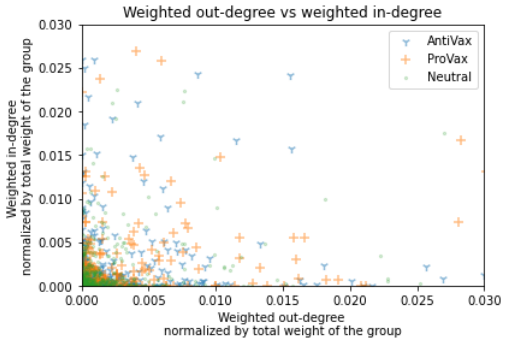

Figure 14: Zoomed-in plot of weighted out-degree vs in-degree with colors highlighting the three groups.

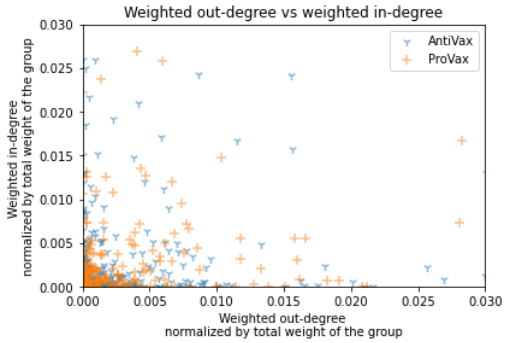

Figure 15: Plot of weighted out-degree vs in-degree for AntiVax and ProVax users.

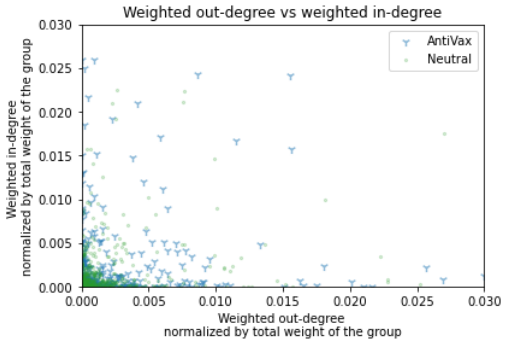

Figure 16: Plot of weighted out-degree vs in-degree for AntiVax and Neutral users.

The normalized degrees used in the plot are shown in Tables 5, 6, 7, 8, 9, 10, which show the 20 highest values for each measure. From these tables we can see that AntiVax users still rank the highest in terms of degree centrality. However, now that we took into account a normalization by the number of users of the same group, we see that we also have some ProVax users in the top rankings, whereas before ProVax users ranked lower because they suffered from being in the smallest group of users.

degree	norm_degree	labels
494	1.071584	AntiVax
165	0.767442	ProVax
320	0.694143	AntiVax
242	0.524946	AntiVax
240	0.520607	AntiVax
111	0.516279	ProVax
228	0.494577	AntiVax
227	0.492408	AntiVax
206	0.446855	AntiVax
194	0.420824	AntiVax
190	0.412148	AntiVax
188	0.407809	AntiVax
183	0.396963	AntiVax
82	0.381395	ProVax
173	0.375271	AntiVax
165	0.357918	AntiVax
165	0.357918	AntiVax
76	0.353488	ProVax
156	0.338395	AntiVax
227	0.334808	Neutral

Table 5: Highest values of degree normalized by the number of users of the same group.

out_degree	norm_out_degree	labels
429	0.930586	AntiVax
107	0.497674	ProVax
190	0.412148	AntiVax
189	0.409978	AntiVax
183	0.396963	AntiVax
179	0.388286	AntiVax
173	0.375271	AntiVax
68	0.316279	ProVax
135	0.292842	AntiVax
121	0.262473	AntiVax
115	0.249458	AntiVax
113	0.245119	AntiVax
166	0.244838	Neutral
162	0.238938	Neutral
108	0.234273	AntiVax
107	0.232104	AntiVax
104	0.225597	AntiVax
150	0.221239	Neutral
101	0.219089	AntiVax
98	0.212581	AntiVax

Table 6: Highest values of out-degree normalized by the number of users of the same group.

in_degree	norm_in_degree	labels
164	0.762791	ProVax
213	0.462039	AntiVax
191	0.414317	AntiVax
186	0.403471	AntiVax
80	0.372093	ProVax
169	0.366594	AntiVax
161	0.349241	AntiVax
160	0.347072	AntiVax
141	0.305857	AntiVax
63	0.293023	ProVax
134	0.290672	AntiVax
132	0.286334	AntiVax
125	0.271150	AntiVax
58	0.269767	ProVax
179	0.264012	Neutral
109	0.236443	AntiVax
105	0.227766	AntiVax
103	0.223427	AntiVax
102	0.221258	AntiVax
101	0.219089	AntiVax

Table 7: Highest values of in-degree normalized by the number of users of the same group.

Wdegree	norm_Wdegree	labels
3468.0	1.475117	AntiVax
2218.0	0.943428	AntiVax
2100.0	0.893237	AntiVax
2009.0	0.854530	AntiVax
1583.0	0.673330	Neutral
1467.0	0.623990	AntiVax
1211.0	0.515100	AntiVax
1184.0	0.503615	AntiVax
1179.0	0.501489	Neutral
1164.0	0.495108	AntiVax
1054.0	0.448320	AntiVax
1051.0	0.447044	AntiVax
1047.0	0.445342	AntiVax
1044.0	0.444066	ProVax
1001.0	0.425776	None
976.0	0.415142	AntiVax
962.0	0.409188	None
942.0	0.400681	AntiVax
918.0	0.390472	AntiVax
916.0	0.389621	AntiVax

Table 8: Highest values of weighted degree normalized by the total weight of users of the same group.

out_Wdegree	norm_out_Wdegree	labels
3158.0	1.343258	AntiVax
1136.0	0.483199	AntiVax
1021.0	0.434283	AntiVax
1001.0	0.425776	None
972.0	0.413441	AntiVax
962.0	0.409188	None
881.0	0.374734	None
820.0	0.348788	AntiVax
814.0	0.346236	Neutral
812.0	0.345385	None
796.0	0.338579	AntiVax
762.0	0.324117	AntiVax
758.0	0.322416	Neutral
687.0	0.292216	Neutral
687.0	0.292216	Neutral
686.0	0.291791	AntiVax
669.0	0.284560	Neutral
663.0	0.282008	AntiVax
628.0	0.267120	AntiVax
623.0	0.264994	None

Table 9: Highest values of weighted out-degree normalized by the total weight of users of the same group.

in_Wdegree	norm_in_Wdegree	labels
1830.0	0.778392	AntiVax
1742.0	0.740961	AntiVax
1652.0	0.702680	AntiVax
1500.0	0.638026	Neutral
1041.0	0.442790	ProVax
942.0	0.400681	AntiVax
942.0	0.400681	AntiVax
907.0	0.385793	AntiVax
883.0	0.375585	AntiVax
877.0	0.373033	AntiVax
854.0	0.363250	Neutral
821.0	0.349213	N/A
820.0	0.348788	N/A
786.0	0.334326	AntiVax
761.0	0.323692	AntiVax
753.0	0.320289	N/A
698.0	0.296895	AntiVax
670.0	0.284985	AntiVax
631.0	0.268396	N/A
624.0	0.265419	AntiVax

Table 10: Highest values of weighted in-degree normalized by the total weight of users of the same group.

4.3 Closeness centrality

Another way to define the importance of a node is whether or not that node is closely connected to the others. The smaller the mean distance from a node to other nodes is, the more important that node could be. Suppose d_{ij} is the length of the shortest path from node i to node j, i.e. the number of edges along the path the shortest path. This is called the geodesic distance. Then the mean geodesic distance from i to j, averaged over all vertices j in the network, is:

$$l_i = \frac{1}{n-1} \sum_{j(\neq i)} d_{ij} \tag{1}$$

where we have excluded from the sum the effect of node i on itself. [7] The inverse of this quantity is the closeness centrality C of a node:

$$C_i = \frac{1}{l_i} \tag{2}$$

This number is higher if the node is on average at a short distance from the others, so it would be considered more central.

As it usually is for social networks [7], these values tend to spam a very small interval and so are cramped together, as we can see from the histogram in Figure 17, that shows the distribution of closeness centrality for the nodes within the network. Note that the highest closeness centrality is 0.394 while the smallest is 0.127, so the values are really close to each other.

In Figure 18 we can also note that AntiVax users are the ones with the highest average value of closeness. This means that the average distance between two AntiVax users is shorter than the average distance between two Neutral or ProVax users, so AntiVax users are in a way more closely connected together than the other groups.

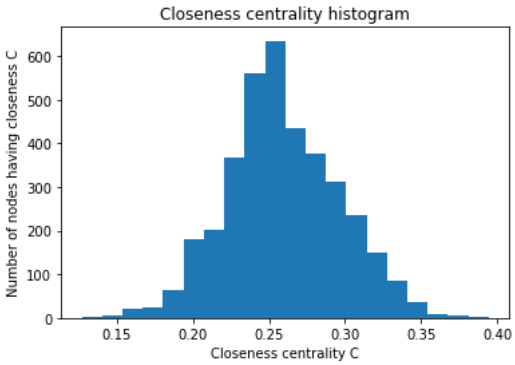

Figure 17: Closeness centrality histogram.

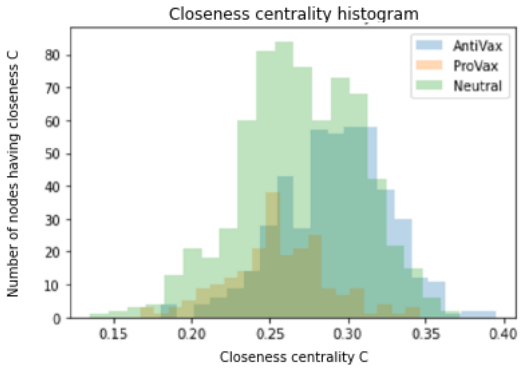

Figure 18: Closeness centrality histogram for the three groups.

If we take a look at the mean of l_i over all users, we find that it is $l = 3.94$. This means that, if we take any pair of nodes at random, on average the shortest distance between them will be just about four edges. The fact that thousands of nodes are on average all very close to each other is quite a counter-intuitive result, but one that is widely occurring in networks. This phenomenon is called the "small-world effect" [6], and it is widely documented This is due to the fact that path length typically is a very slow function of the number of nodes, usually growing as the logarithm of the number of nodes ([7], Par. 8.2). This effect is of crucial importance for an online social network like Twitter, which benefits from a fast flowing of information, that should reach as many people as possible in the shortest time.

As we saw in Section 4.1, this network could be considered a scale-free network. The presence of "hubs" or celebrities within the network, i.e.: users that are connected to a very large number of other users, may be a factor that contributes to this small-world effect. The network could be structured in dense communities in which nodes with a low degree are well connected to each other. In Section 6.2, we will try to inspect the communities formed in the network and compare them to the data we have about the user polarity.

4.4 Betweenness Centrality

In the previous section, we mentioned that a social network benefits from having a fast flowing of information. For this reason, any node which can influence this flow of information can be considered as very important in the system.

A measure of centrality that takes into account how important a node is for the flowing of information is the betweenness centrality, which measures how well a node falls "between" others. The betweenness centrality of a node is the number of shortest paths a node lies on. If we assume that information from node to node flows at a constant rate and follows the shortest path, the number of shortest paths a node lies on is directly proportional to the information that flows through it. For example, if a node i lies upon the shortest path between two other nodes, then the removal of node i would disrupt this path and the flow of information would be different and perhaps slower.

In Figure 19 we can see a histogram of betweenness centrality with colors highlighting the three polarity groups. As we can see from the histogram, values of betweenness span over a very large range, with the highest value of ~ 476076, but having also about ~ 3000 users with betweenness $= 0$ (the smallest value bigger than zero is ~ 0.14). In Table 11 we can see the polarities of the users with the highest betweenness. As for the other centrality measures, AntiVax users are in the top ranks of betweenness, but we find also notable ProVax and Neutral users.

Highest Betweenness

	polarity	betweenness B
1	AntiVax	476076
2	Neutral	437611
3	ProVax	384537
4	ProVax	384094
5	AntiVax	364784
The first AntiVax user is at position:	1	with B = 476076
The first ProVax user is at position:	2	with B = 384537
The first Neutral user is at position:	2	with B = 437611

Table 11: Polarity and degree of the users with the highest betweenness.

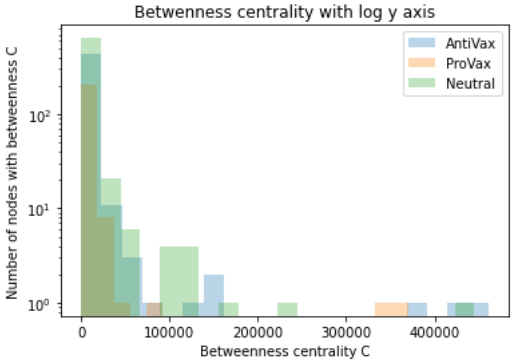

Figure 19: Betweenness centrality histogram for the three groups.

4.5 PCA of the centralities

Principal Component Analysis (PCA) is a statistical analysis of multidimensional data that aims to extract features of a dataset to highlight the distinctions among the data points. PCA works by defining a new set of variables, the "Principal Components", as linear combinations of the original variables. The coefficients of the linear combination are chosen in order to give more weight to variables over which there is greater separation between the data points. The coefficients of the linear combination are called loading scores. The new set of variables, i.e. the principal components, are built in order of importance: PC1 is the linear combination that maximizes the separation among the data points, PC2 is the best linear combination that is orthogonal to PC1, PC3 is the best one that is orthogonal to PC1 and PC2, and so on. By measuring the variation of the data around the origin for each PC, we can also quantify how much of the data variation a specific PC accounts for. Usually, the first few PCs account for the vast majority of the variations: this means that by examining just a few PC variables we get a picture of the entire multi-variate data. Hence, PCA is useful as a dimensionality reduction technique.

In our case, the variables we are considering are the centralities evaluated in the previous sections. We have eight total variables, which are:

1. Unweighted (undirected) degree
2. Unweighted out degree
3. Unweighted in degree
4. Weighted (undirected) degree
5. Weighted out degree
6. Weighted in degree
7. Unweighted betweenness
8. Unweighted closeness

From these eight variables, with Python library scikit-learn [1] we build eight principal components. In Fig. 20 we can see the amount of variation in the data each principal component accounts for.

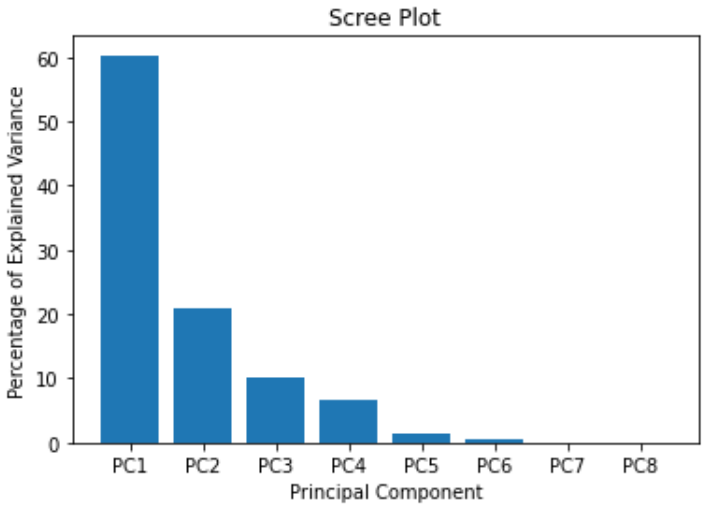

Figure 20: Scree plot that shows the percentage of variation in the data each principal component accounts for.

The first two principal components combined account for 80% of variation in the data. So, by plotting these two axes we get useful insights into data that, in the original variable space, were impossible to appreciate on a 2D graph.

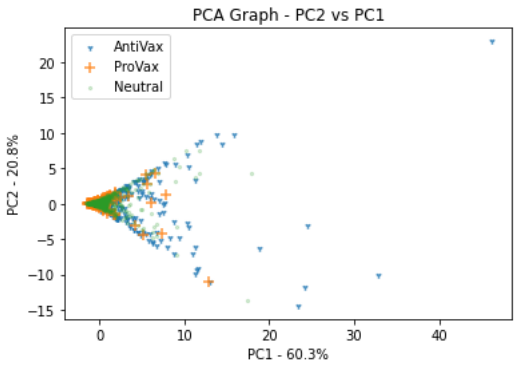

Figure 21: Plot of PC1 and PC2 with colors highlighting the three groups.

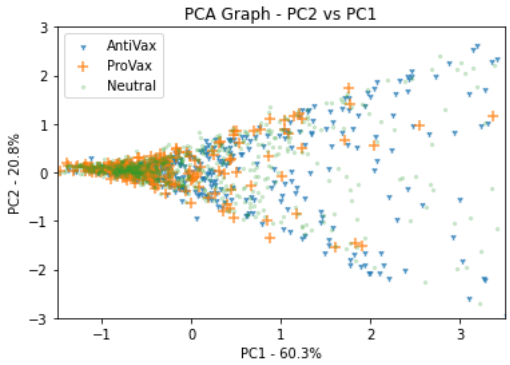

Figure 22: Zoomed-in plot of PC1 and PC2 with colors highlighting the three groups.

As we can see in Fig. 21 and Fig. 22, there is not a clean separation between the three groups of users. In fact, in large areas of the plot we see the three types of users overlapping each other. However, we can also note that as we move to higher values of PC1 (moving to the right in the graph) we see that the relative quantities of users tend to change. In particular, we encounter fewer ProVax users, as we can more clearly see in Fig. 23 and Fig. 24. If we compare instead AntiVax and Neutral users, as in Fig. 25 and Fig. 26, we see that there is no visible separation.

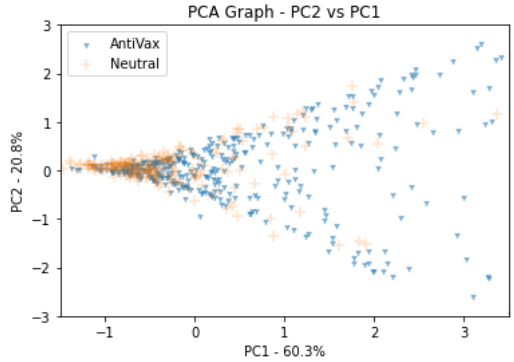

Figure 23: Plot of PC1 and PC2 for AntiVax users and ProVax users.

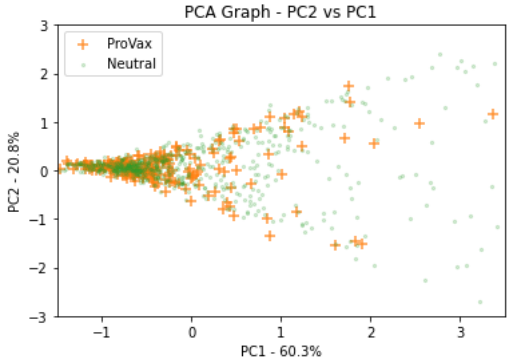

Figure 24: Plot of PC1 and PC2 for ProVax users and Neutral users.

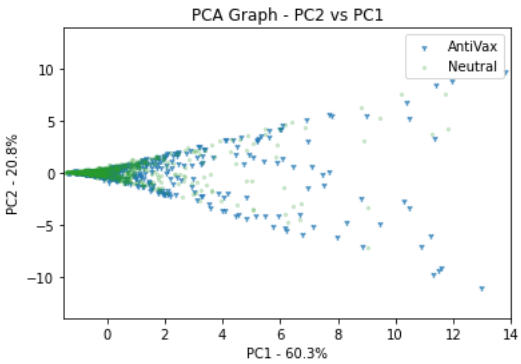

Figure 25: Plot of PC1 and PC2 for AntiVax users and ProVax users.

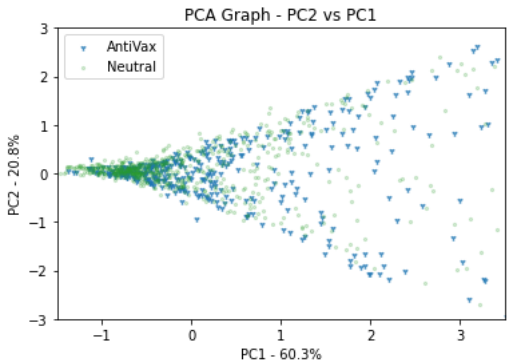

Figure 26: Zoomed-in plot of PC1 and PC2 for AntiVax users and ProVax users.

Centrality	PC1 loading score
Unweighted Degree	0.445367
Weighted Degree	0.441610
Weighted Out-degree	0.349295
Unweighted Out-degree	0.345501
Weighted In-degree	0.325417
Unweighted In-degree	0.321617
Closeness	0.298680
Betweenness	0.258279

Table 12: PC1 loading scores for the centralities.

This situation is not too different from the one we encountered in Section 4.2, where we compared out-degree with in-degree. To get an idea of how much each centrality affects the PCA graph, we can take a look at the loading scores for PC1 in Tab. 12. We can see that the variables that have the greater effect in separating the data are unweighted and weighted degrees, while closeness and betweenness have the least effect.

5 Other characterizations of the network

5.1 Number of tweets per day

We can take a look at the number of tweets per day made by the users of the three polarity groups and compare them to see if there are any significant differences. In Figure 27 we can see the number of tweets per day for the three groups. To compare them to each other, we should normalize by the total number of tweets per group, as in Figure 28. Here we can see that there is not a significant difference between the three groups. The peaks in the graphs represent days of high activity by the user and might be correlated to some specific events that triggered the users to post more tweets. For example, we find the highest peak on the 27 of December 2020, the day of the first vaccination in Italy [8].

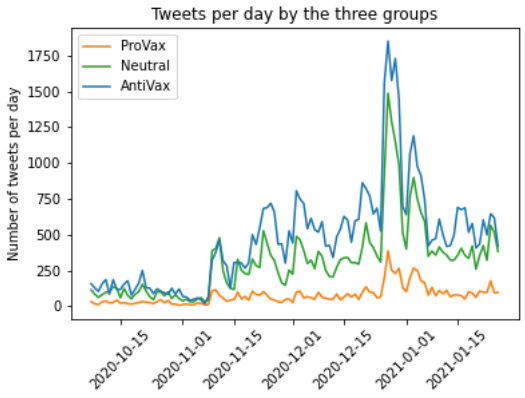

Figure 27: Number of tweets per day for the AntiVax, Neutral and ProVax users.

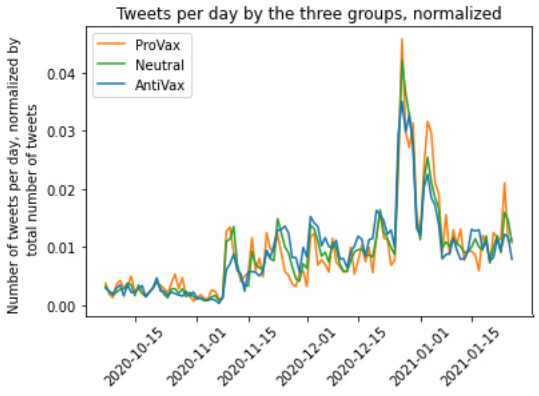

Figure 28: Number of tweets per day for the AntiVax, Neutral and ProVax users, normalized by the total number of tweets per group.

We can also show the activity (number of tweets per day) in a heat-map representation, by presenting the days in a calendar fashion. These representations are shown in Figures

29, 30, and 31. From these figures, it's easy to see which day marked moments of high activity.

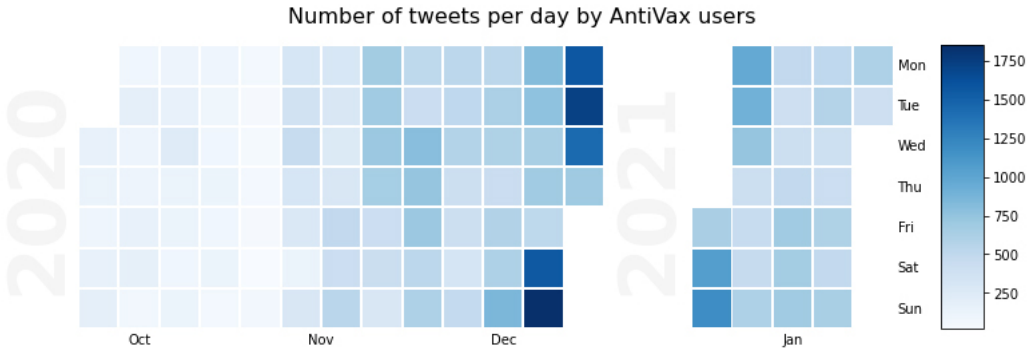

Figure 29: Heatmap showing the number of tweets per day for the AntiVax users.

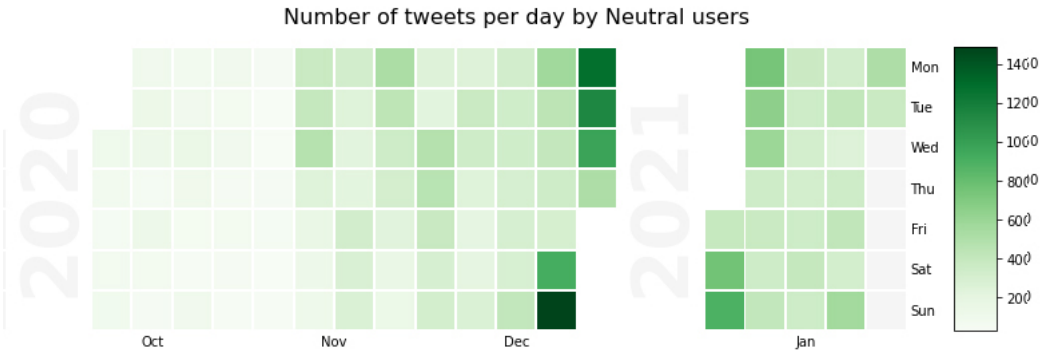

Figure 30: Heatmap showing the number of tweets per day for the Neutrals users.

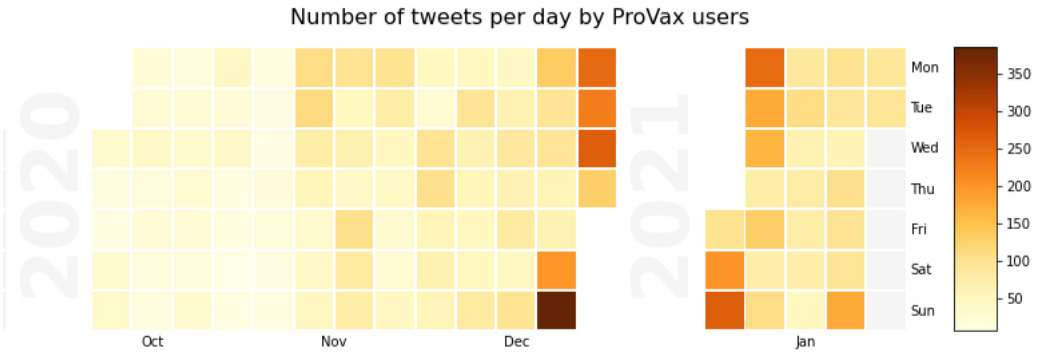

Figure 31: Heatmap showing the number of tweets per day for the ProVax users.

5.2 Length of the tweets

We can also inspect the length of the tweets posted by the users of the three polarity groups. We define length as the number of characters of the tweets, spaces and punctuation included. In Figure 32 we can see a histogram of the length of the tweets.

We can see that the shapes of the histograms for the three groups are similar, but for a fair comparison we should normalize the histogram since we have a much lower number of ProVax tweets than the others. In Figure 33 we have the normalized histogram with

respect to the number of nodes in each group, and we can see that the three groups follow more or less the same distribution for the length of the tweets. From this, we don't see any particular difference between AntiVax, Neutral, or ProVax users.

Looking at the histogram, the first features we see are two big spikes, that are centered on L=140 and L=280. Regarding L=280, this is the maximum length of a tweet allowed on Twitter as of today. This means that users who post long tweets tend to "cram" their tweets within the character limit, resulting in an over-representation of tweets with L=280 due to users trying to write as much as possible per tweet.

Regarding L=140, we can see a pretty narrow spike instead of a distribution around it. This could suggest that there could be a particular reason people posted 140-character long tweets. Upon inspection, we see that almost all of the tweets falling on this length are truncated, and end with "..." (three dots). This could mean that in the process of acquisition of the data, for some reason some tweets were perhaps not collected in their entirety but only the first 140 characters.

In addition to that, we see a small but noticeable bump around $L \sim 130$. The overall shape of the histogram is not unusual as a distribution of tweet lengths, in fact, we see the same kind of behavior in articles from the Twitter engineering blog (for example [4]), in which a much larger pool of tweets was analyzed.[3]

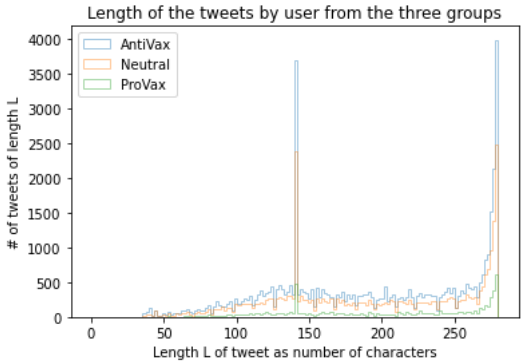

Figure 32: Histogram showing the length of the tweets by the three groups of users.

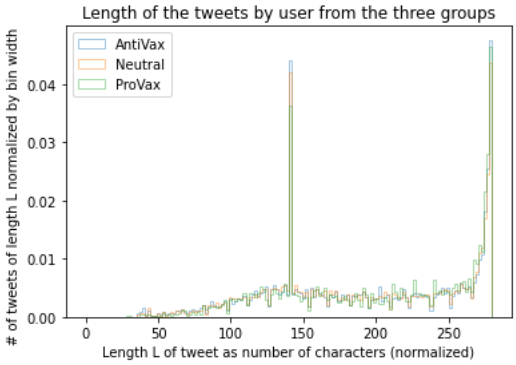

Figure 33: Histogram showing the length of the tweets by the three groups of users, normalized by the bin width.

6 Assortativity and Community detection

We can take the analysis one step further by introducing the concept of mixing within a network. In our network, we have 3 distinct types of users, "ProVax", "AntiVax" and "Neutrals". We might expect that these 3 types of users (or nodes) would not be connected completely at random to each other, but maybe the links would highlight the fact that there are 3 groups of users. It could be possible, perhaps, that nodes are more likely to be connected to nodes of the same type (this situation is known as assortative mixing), or, on the other side, a node is more likely to be connected to a node of a different type (disassortative mixing). [7]

[3]The tweets analyzed in the article were previous to the year 2017, when Twitter updated the maximum number of characters allowed for a tweet from 140 to 280. Also, the exact positioning of the first peak depends on the language of the tweet. Still, what matters to us is that the shapes of the histograms are similar to the ones we found, so we have no reason to think that the users in our analysis behave differently from the standard with regard to tweet length.

6.1 Assortative mixing

To measure assortativity, we count the number of how many edges lie among nodes of the same type c minus how many edges we would expect if we added edges randomly. Normalizing by the total number of edges N in the network we get a quantity known as modularity Q ([7], Par. 7.13):

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{3}$$

where m is the number of nodes, k is the degree, δ is the Kronecker delta and A is the adjacency matrix of the network. Normalizing for the maximum modularity in a given network we get a number between -1 and 1 called the assortativity coefficient a: [7]

$$a = \frac{Q}{Q_{max}} = \frac{\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)}{2m - \sum_{ij} \left(\frac{k_i k_j}{2m} \right) \delta(c_i, c_j)} \tag{4}$$

The assortativity coefficient is 1 if all the connections stay within categories and -1 if all the connections join nodes of different categories.

With Networkx we measured the assortativity coefficient, which resulted to be $a \approx 0.0882$. As it is often the case for social networks, the assortativity coefficient is positive. This indicates that people tend to connect with people within the same group more than with those from other groups. For example, "AntiVax" users are more likely to retweet another "AntiVax" user than a "ProVax" user. [7]

This could suggest that users tend to form communities within the network, which are subsets of the network with nodes more strictly connected to other nodes within the subset rather than to nodes outside the subset.

6.2 Community Detection

It could be interesting to use a community detection algorithm in order to find communities, based solely on network properties, and to compare them to the data from the user labels. In other words, we look at the structure of the network to locate different communities, and then we compare them with the 3 groups we have to see how they overlap.

From the assortativity coefficient, we expect these communities to be "closed" more often than not, that is, communities in which there are many users from the same group, instead of having the three groups equally represented.

The basic concept of community detection is to divide the network into sub-networks (subsets) in such a way that there will be fewer edges between these subsets rather than we would expect if we placed edges at random. In other words, we should search for the communities that have the highest assortativity we can find.

The problem can be tackled in many ways, and in literature there are many different algorithms for community detection, some that use heuristic methods and/or stochastic methods. The one we chose to use is the Leiden algorithm ([10]). The Leiden algorithm is a refinement of the very famous and widespread Louvain algorithm ([3]), which is known to be very reliable and fast ([5]), and it also is suitable for both weighted and directed graphs.

The Louvain method is a straightforward iterative process that starts by considering each node a community. Then the individual nodes are moved stochastically from one community to another in a way to locally optimize assortativity. After a partition is identified in this way, communities are replaced by supernodes (agglomerate of nodes) of different weights. The process is iterated on the new smaller network of supernodes by again moving nodes. The process is repeated until the assortativity cannot be increased. [7]

The Leiden method is pretty much the same, but it adds a small step to ensure that the communities found are well connected within themselves: before replacing communities with supernodes, it undergoes a phase of refinement, in which a node could be moved at random to another community if in doing so it gets a better-connected community and an increase in assortativity.

6.3 Tuning the algorithm

One of the authors of the original paper on the Leiden algorithm ([10]), V. A. Traag, implemented the algorithm to work with the free software package iGraph for Python in 2020 ([9]), which we will use for the community detection.

For a moment let us ignore the information we have about the direction and weight of the edges, and just consider the network as an unweighted undirected one. If we run the algorithm a couple of times, we see that the results vary wildly. For example, for the first five runs, we get the following numbers of communities found: 35, 22, 42, 38, 40. A degree of randomness is to be expected since the algorithm has a certain number of stochastic steps, but in this case it played a significant role.

However, if we take into account the additional information coming from the direction and weight of the edges, the results of the algorithm are much more reliable from run to run. In fact, testing it with multiple runs, we get a number of communities varying from about 35 to 42; the communities found for various runs also look pretty similar to each other in terms of user memberships.

The statistical fluctuations do not result in a drastic change, since the differences from run to run concern at worst $\sim 5\%$ of the users for the large communities. Of course the smaller the community is the larger the relative error, because with a community of a few users the displacement of just one user could impact a lot. However, excluding the very small communities with less than 10 users, differences from run to run on small communities ($\sim$ 10 to 50 users) are always within $\sim 10\%$, usually within $\sim 5\%$.

It is also possible to iterate the whole Leiden algorithm more than one time, feeding back the output of the algorithm into the input, which gives a finer structure to communities with more communities of smaller sizes. However, the partition we get from one iteration is fine enough and a further division of the communities would not convey substantial greater information since we already have lots of very small communities. For this reason, we will not feed the algorithm back into itself.

In conclusion, we will use the Leiden algorithm in the version implemented by V. A. Traag, taking into account the information about the weights and direction of edges, without feeding the algorithm back on itself.

7 Visualizing the communities

Our interest is to see in which way the three known groups ("AntiVax", "ProVax", "Neutrals") are spread over the communities. First of all, it is useful to visualize the three groups over the whole network, as in Figure 34.

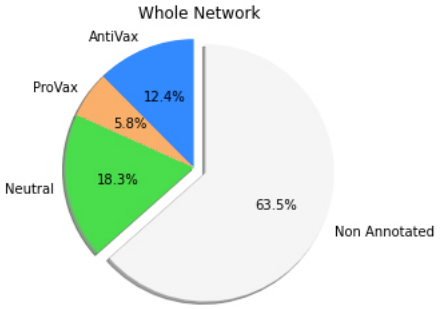

Figure 34: Pie chart showing the polarity of users over the whole network

From this pie chart, we can see the relative size of the groups within the network, plus a big part of the network which is made of users we don't have information about. This, as noted in Section 3 is due to the fact that for most edges of the network we know the polarity of just the posting user and not the retweeted user.

If we were to suppose that the polarity would have no correlation with the community structure of the network, we would expect to find about the same percentage of the three groups in each detected community. On the contrary, if the percentage for the different communities would look different from this we could suppose that a user being in a group or the other actually has an influence on the community it forms.

We found out that for each community detected, the percentage of non-annotated users was about the same of the percentage over the whole network (which was 63,5%). This could indicate that the relatively small sample we have of the labeled users is representative enough of the whole network and the relative proportions of the three polarity groups would be more or less the same even if we knew the information of polarity for these non-annotated users.

With the Leiden algorithm, we detected 37 communities, indexed from the largest (0) to the smallest (36). Communities 15 to 36 consist of just a few users each (community 15 has 6 users, community 20 just 3 users, and so on), so they are negligible for our analysis. The other communities range from having more than 800 users to a couple dozen users.

Figures 35 to 39 show pie charts with the relative quantities of AntiVax users, ProVax users, Neutral users, and unlabeled users for the 5 biggest communities. In Figures 40 and 41 it is shown in a stacked-bar char the absolute and relative distribution of users of each polarity, respectively, among the biggest 15 communities detected.

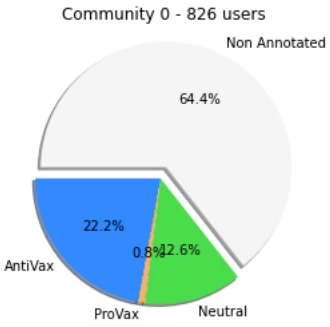

Figure 35: Pie chart showing the polarity of users of community 0.

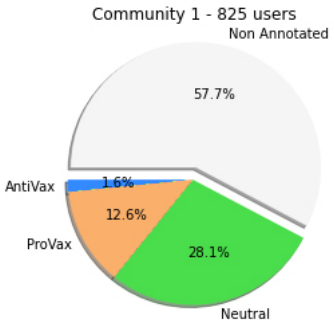

Figure 36: Pie chart showing the polarity of users of community 1.

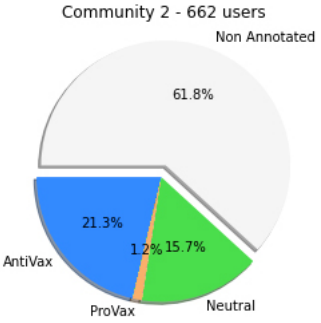

Figure 37: Pie chart showing the polarity of users of community 2.

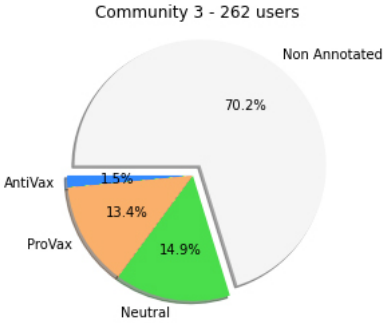

Figure 38: Pie chart showing the polarity of users of community 3.

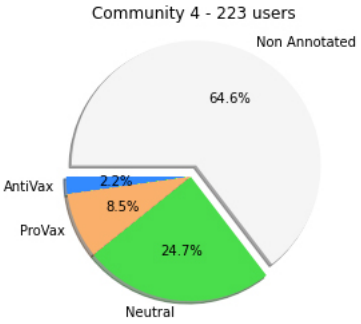

Figure 39: Pie chart showing the polarity of users of community 4.

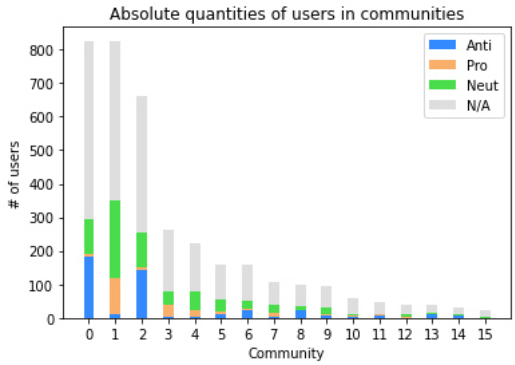

Figure 40: Stacked bar chart showing the absolute number of users for each polarity for communities 0 to 14.

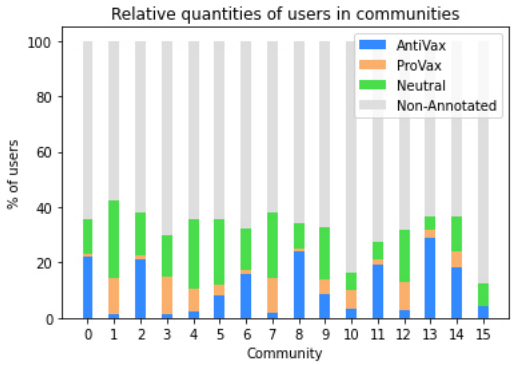

Figure 41: Stacked bar chart showing the percentage of users for each polarity for communities 0 to 14.

As we can see, the proportion of AntiVax, ProVax, and Neutral in each community is not exactly the same as over the whole network. This could mean that the user polarity has actually an influence on how the users interact with each other.

One thing we note is that in almost every community there is a substantial fraction of either AntiVax or ProVax users and the rest are Neutral users. It never happens that we have a community with both a substantial AntiVax fraction and ProVax fraction. The two groups do not mix with each other: an AntiVax user would rarely connect with a ProVax user and vice-versa. On the contrary, Neutral users tend to interact both with AntiVax users and ProVax users: in almost every community we have a certain percentage of AntiVaxxers or ProVaxxers and more or less the same quantity of Neutrals.

In every community but two, the fraction of non-annotated users is more or less the same as the fraction over the whole graph. As we said, this could mean that if we had had the information about every user, the result would have perhaps maintained approximately the same proportions. The only exceptions are communities 10 and 15, in which more than 80% of the users are non-annotated, which could be due to an effect that is unknown to us. Still, the fact that almost every community is made of about 60% of non-annotated users could suggest that our sample could be ad adequate representation of all the users. In Figure 42 we can see a representation of the whole network with the communities that were detected.

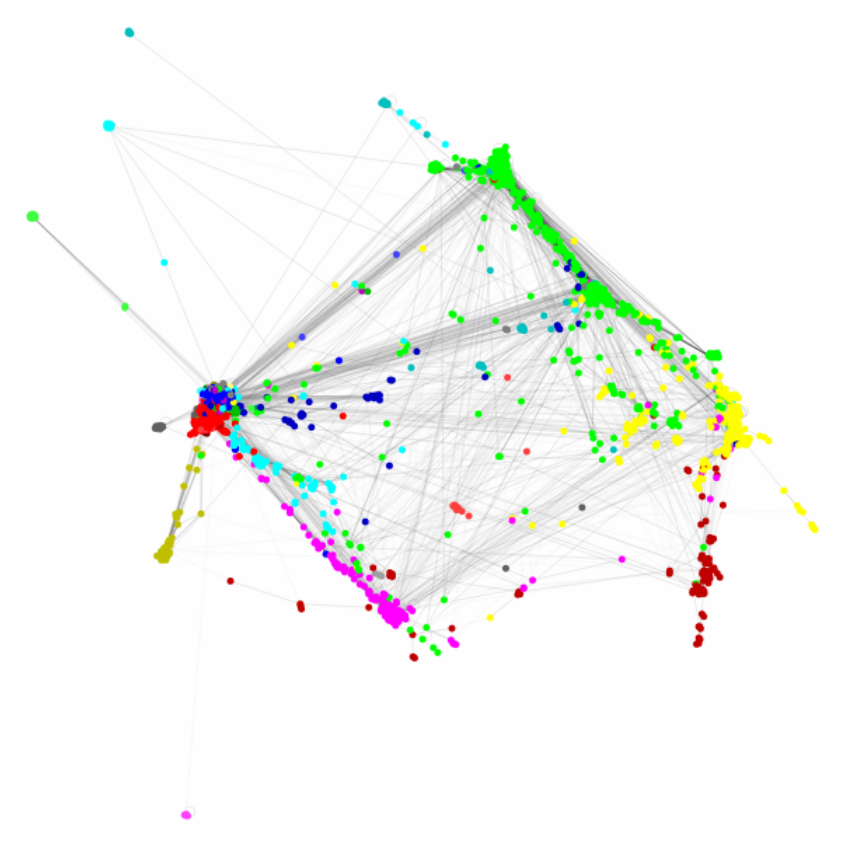

Figure 42: Visualization of the network with the different communities found highlighted in different colors.

8 Conclusions

In this project, we analyzed a network of tweets and examined how three different groups of users, AntiVax, ProVax, and Neutrals, interact on this network. By measuring typical characteristics of a network such as various types of degree centrality and closeness centrality, we noted that this network tends to behave as it is expected for a social network and shows typical distributions of centralities.

We noted that the degree distribution seems to follow a power law, which gives the network the property of being "scale-free", not unusual for social networks. This means that there is a large fraction of users who are connected to a number of users remarkably larger than the average; these users are the "celebrities" in the network and their influence is substantial. We saw that most "celebrities" belong to the AntiVax group, meaning that the largest hubs are almost always AntiVax users, being both the users that get retweeted the most and by the largest number of users and both the users who retweet the most.

We also took a look at the betweenness, a measure of how important a node is to bridge different users. Still, we have AntiVax users in the top ranks of betweenness, but we find also notable ProVax and Neutral users with very high betweenness.

In order to have a more insightful look at the centralities measured, we performed a principal component analysis on them. Since the first two principal components account for more than 80% of the variation within the data, examining them gives a useful picture of how the three groups compare to each other. From this, we saw that there is no clear separation between the three groups, since in most areas of the graph we have all three types of users combined. Still, there are some tendencies that emerge from this analysis,

most notably that ProVax users are confined to a smaller area of the graph than AntiVax and Neutral users.

Regarding the tweet-per-day rate and tweet length, we saw that there are not any significant differences between the three groups and the overall behavior is exactly in line with what we would expect.

We then examined the community structure that arises in the network and how the three groups mix. By measuring the assortativity coefficient, we noted that users tend to mix mostly with users of the same group. We then used the Leiden algorithm to detect communities formed by the users on the network we compared these communities to the data previously gathered about the three groups. We noted that, as expected, being labeled AntiVax, ProVax or Neutrals has an impact on the presence in a community. We noted that most communities have a component of Neutral users and either Antivax users or Provax users. In summary, AntiVax users and ProVax users both mix with Neutral users but they do not mix together in any of the communities found.

References

[1] Scikit-learn. URL `https://scikit-learn.org/stable/index.html`.

[2] Networkx documentation, 2022. URL `https://networkx.org/`.

[3] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008. doi: 10.1088/1742-5468/2008/10/p10008.

[4] Ikuhiro Ihara. Our discovery of cramming, 2017. URL `https://blog.twitter.com/engineering/en`$_{u}s/topics/insights/2017/Our - Discovery - of - Cramming$.

[5] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5), 2009. doi: 10.1103/physreve.80.056117.

[6] S. Milgram. The small-world problem. *PsycEXTRA Dataset*, 1967. doi: 10.1037/e400002009-005.

[7] M. E. J. Newman. *Networks an introduction*. Oxford University Press, 2010.

[8] Redazione Sky TG24. Vax day, le immagini dei primi vaccinati in italia. foto, Jan 2021. URL `https://tg24.sky.it/cronaca/2020/12/27/covid-primi-vaccini-italia-vax-day`.

[9] V. Traag. Vtraag/leidenalg: Implementation of the leiden algorithm for various quality functions to be used with igraph in python., 2022. URL `https://github.com/vtraag/leidenalg`.

[10] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-41695-z.