CANCEL CULTURE AND REDFLAG HASHTAG: A SOCIAL NETWORK ANALYSIS APPROACH TO SOCIAL NORMS AND MORALITY GENERATED IN TWITTER

Dacia Braca Michele M. Crudele Lucia Depaoli Alessandro Marcomini
Simone Mistrali Camilla Quaglia Jacqueline Pamela Padilla Torres
Gala Pradillo Diaz

February 3, 2022

Contents

1 Introduction **3**

2 Theoretical background **3**

3 Goal of the study **6**

4 Tweets Download and Pre-Processing **6**
4.1 Tweets Download: the *snscrape* Library 7
4.2 Pre-Processing 7

5 Networks Creation **9**
5.1 Network of Hashtags 9
5.1.1 Network of Hashtags Creation 9
5.1.2 Network of Hashtags Visualization 9
5.2 Selection of Topics 11
5.2.1 The Six Topics 11
5.3 Semantic Networks 12
5.3.1 Semantic Networks Creation 12
5.3.2 Semantic Networks Visualization 13
5.3.3 WordCloud 19

6 Control Group Analysis **21**
6.1 Control Group Analysis 21

7 Degree distribution **28**
7.1 Maximum likelihood estimation of γ 28
7.2 CCDF for estimating γ 28

8 Assortativity analysis **31**
8.1 Pearson correlation coefficient . . . 32
8.2 Assortativity coefficient through fitting . . . 33

9 Robustness **36**
9.1 Random Failures and Targeted Attacks . . . 36

10 Centrality analysis **38**
10.1 Centrality measures . . . 38
10.2 Communities' centrality measures . . . 40
10.2.1 Covid community . . . 41
10.2.2 Climate . . . 45
10.2.3 Abuse . . . 47
10.2.4 Mental health . . . 49
10.2.5 Dating . . . 51
10.2.6 Market . . . 53
10.3 Clustering coefficient . . . 55

11 Sentiment Analysis **58**
11.1 Vader Sentiment . . . 58
11.2 Correlation with neighboring sentiments . . . 63
11.3 Text Blob . . . 64
11.4 Linguistic Inquiry and Word Count (LIWC) . . . 69
11.5 Positive & Negative Nodes Removal . . . 70

12 Summary of the Results **75**

13 Discussion **75**

14 Limits & future directions **77**

Abstract

Social Networks are virtual spaces where people can almost freely express their interests, experiences and opinions in always more and new different ways. Whether or not this is positive for society is still a topic of discussion, but for sure nowadays they influence many sociological aspects on global scale, as they are part of the public social sphere. In fact, on one hand they reflect the dominant ideology and culture, on the other hand, they create new and alternative areas of discussion. Actually, while platforms like Instagram, TikTok or Facebook are more related to the sharing of multimedia, Twitter is particularly used for sharing thoughts and opinions and as a consequence it can be scraped to analyze many different sociological trends and behaviors. In this project, we try to understand whether there exist a connection between the sociological phenomenon of *cancellation* and the RedFlag topic, trending on social media from September to November 2021. First, we understand which are the topics related to the RedFlag trend studying the network of *hashtags* linked to it (since on social media, especially on twitter, hashtags indicates topics). Using this network, we look for different communities inside it that represent different themes related to RedFlag and we choose six of them to be analyzed further. Afterwards, we build semantic networks of the *words* used inside each of those communities, trying to understand what is considered as a RedFlag in the different topics. For each of them, many Network Science tools are used to determine their technical features: studies about degree distribution, centrality and robustness show the scale-free nature of all the semantic networks we consider, meaning the presence of high-degree nodes (hubs). A control group of

tweets not including the hashtag #RedFlag is then built to confirm the influence of that hashtag on the words used in each topic. Finally, a sentiment analysis is performed using different tools, giving sensible and consistent results about the general feelings that characterize the semantic networks. The results are eventually analyzed from a psycho-social perspective, and compared with cases of cancellation, with the purpose of verifying if this type of hashtags contribute to generating social norms and shared morality in today's society.

1 Introduction

During the last few years it has been observed, in some way, an hegemonic thinking in social networks, specifically in Twitter. This has led to practices, more and more frequent, such as the *cancellation*. This concept is used as a synonym for ostracism in social media. The cancel culture involves behaviors such as group shaming, and ejection of social and professional circles. Usually, this kind of internet humiliation is done to public figures and companies when they have done or said something considered for the public sphere as politically incorrect, objectionable or offensive [14].

This research work aims to analyze the morality of this hegemonic thought, if it exists, through the latest trend "#RedFlag", since during September 2020 many videos began to appear on TikTok accompanied by the red flag emoji and hashtag, where users pointed out actions considered as toxic in a couple's relationship. However, this trend quickly jumped to twitter, and the red flags were used for all kinds of issues, including giving opinions about gastronomic and musical tastes, or daily actions such as greeting the bus driver or thanking the waiters. The entire network was commenting on the behaviors that seemed inadequate to them and therefore pointing them as "toxic". The trend went so viral that brands took advantage of it and started using it for advertising purposes, doing their own red flags. Over time, the hashtag has also included ironic red flags, memes and even green flags (the opposite to red flags, to point out positive actions).

To this end, all the topics related to the RedFlag trend will be examined, in order to extract popular opinions about social issues. To achieve this, students from the Social Network Analysis course, in collaboration with some from the Network Science one, will look over Twitter's network searching for that hashtag.

2 Theoretical background

Gala Pradillo Díaz & Jacqueline Pamela Padilla Torres

Within the academic field of sociology and psychology, many authors have studied the forms and spaces in which human beings rationally deliberate on matters that concern them. This area of social life is known as the *public sphere*. One of the most influential authors in the study of this field is Jürgen Habermas, who elaborate a great interpretation and development of the concept. According to this philosopher the public sphere establishes a space for the construction of public opinion which is open to all citizens. Through conversation, private individuals constitute themselves as public individuals, becoming part of the public space. In this space, rational arguments are exchanged to reach an agreement – something like a common will – regarding matters that concern the general interest. The author claims that the bourgeois public sphere originated in the XVI century when mercantile capitalism, together with the press and the proliferation of meeting rooms, created the conditions for a new public domain to emerge. The bourgeois public sphere emerged, then, as a space in which private individuals met to discuss among themselves the regulation of civil society (Castrelo, 2018 p. 73-74). In his own words:

"By "the public sphere" we mean first of all a realm of our social life in which something approaching public opinion can be formed. Access is guaranteed to all citizens. A portion of the public sphere comes into being in every conversation in which private individuals assemble to form a public body. They then behave neither like business or professional people transacting private affairs, nor like members of a constitutional order subject to the legal constraints of a state bureaucracy. Citizens behave as a public body when they confer in an unrestricted fashion–that is, with the guarantee of freedom of assembly and association and the freedom to express and publish their opinions-about matters of general interest. In a large public body this kind of communication requires specific means for transmitting information and influencing those who receive it. Today newspapers and magazines, radio and television are the media of the public sphere."(Habermas, 1974, p.49)

In addition, the author includes the Theory of the Communicative Action characterized by mutual understanding, which operates as a harmonizing mechanism of divergent interests around common problems. Therefore, he believes that there is not a system which regulates the public sphere communication, because they are mainly spontaneous, and its administration cannot be organized by whoever holds power. Nevertheless, this theory considers the possibility that public opinion can be distorted and manipulated. Anyway, Habermas has a strong conviction that the public sphere and communicative action are key to the right performance of deliberative democracies (Castrelo, 2018 p. 75).

These days, the public spheres have changed, they are not any more the London tea houses or French cafes that Habermas referred to. In fact, the philosopher's theory has some limits and should be revised taking into account the recent social changes. To start with, according to the author, the interaction of the public sphere has to be face to face. Anyway, in the last few decades this social space has been amplified, and the production and ways of inhabiting it are increasingly linked to social networks and digital platforms. Thus, the social sphere is not anymore just a physical place, but also a virtual one. Furthermore, nowadays it is hard to believe that the public sphere is based on the search for agreement, as we can see in the polarization of the topic discussed among society (anti-vaccines, the flat earth society, pro-life, etc.). Lastly, the collective action has also changed. The groups created on the Internet are more flexible and volatile, also, the requirements to be part of them are less strict in terms of identity and affiliation.

Regarding the polarization of the topics discussed in the public sphere, it has to be noticed that it is due to the fact that now it is a global area. Therefore, we can not expect a high consensus as it involves a wide variety of identities. In fact, some authors are already studying about ¡¡counter-public space¿¿, alternative areas of discussion. By alternative we mean complementary, or even sometimes contrary, to the mainstream. The term counter-public spheres has two dimensions. On the one hand, it refers to some social groups that feel outside the mainstream or not represented by it, so they create different spaces of discussion where their identities, interests and needs are reflected. "An important reason for the construction of counter-public spheres is the subjective feeling of those affected, that information, messages, news, etc they produce do not find the way into the mainstream media" (Wimmer, 2005 p.96). As an example we can mention the Black Twitter, an informal community inside twitter focused on the interest of the black community. On the other hand, the notion of counter-public spheres implies new political processes and forms or organizations. For instance, all the actions made regarding the black life matters movement, such as the publication of black post on Instagram, or the boycott that the K-pop community made to the Dallas Police Department (the institution asked citizens to send them videos of illegal activity from the black live matters protest through a special app called iWatch Dallas, instead, "K-pop

fans flooded the software with content from their favorite artists and seemingly overloaded the reporting system in the process" (Alexander, 2020)).

As we can see, the digitization and globalization of the public sphere has a great impact on its access, participation and production (text, hypertext, multimedia, interactivity). This communicational paradigm shift has undoubtedly changed the way in which political and social ideologies are exchanged. Information on the internet now flows without limit. However, this has resulted in the saturation of individualized currents of opinion, which has caused the loss of firm positions that existed before. In the same way, the supposed freedom of the Internet has led to total exposure of individuals, which sometimes means that people do not share their ideas freely for fear of repression, or cancellation (Wimmer, 2005).

This last idea is closely linked to sociological concepts such as cultural hegemony and symbolic violence, both similar to each other. The notion of symbolic violence was coined by Pierre Bordiuer to designate the cultural domination of society by the ruling class. Their beliefs, morals, explanations, perceptions, institutions, values or customs become the accepted cultural norm and the dominant, valid and universal ideology. The idea of cultural hegemony was developed by Gramsci. It justifies the social, political and economic status quo, and shows it as something natural and inevitable, as well as beneficial for everyone. The philosopher makes it clear that in reality it is a social construct that benefits only the ruling class ("Hegemonia cultural", 2021).

As we have mentioned, the Internet and social media have caused a new scenario of communication, characterized by the redefinition of participation and new possibilities for the creation of a common culture. However, we cannot ignore that in this process of communicative transformation, information does not flow in a vacuum, but in a political space that is already occupied, organized and structured in terms of power. The fact that power is decentralized or diffuse does not mean that there is less power, that we have more freedom. Within social networks, large companies, economic agents and public figures are the ones who focus attention and have greater visibility. We cannot fall into the idealization of the participation that occurs in these virtual spaces. We cannot ignore the great commodification that they have: on the one hand through publicity, and on the other hand, through the sale of user data. After all, Twitter is a for-profit company that stratifies the visibility of messages, profiles and trends in favor of advertisers (Peña, 2018). In short, the Internet is not outside of cultural hegemony.

In addition, as we have learned in the course, within social media there are several phenomena that increase social and political polarization. For example, ¡¡echo chamber¿¿, which refers to situations in which beliefs are amplified or reinforced. "By participating in an echo chamber, people are able to seek out information that reinforces their existing views without encountering opposing views, potentially resulting in an unintended exercise in confirmation bias" ("Echo chamber (media)", 2022). Something similar happens in the phenomena of the ¡¡filter bubble¿¿. This concept is used to describe "state of intellectual isolation that can result from personalized searches when a website algorithm selectively guesses what information a user would like to see (...) As a result, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles" ("Filter bubble", 2022).

In brief, the internet is also structured in classes where large companies and other central actors dominate attention and symbolic, social and material benefits. However, this does not mean that there are no elements of counterpower within social networks. The Internet makes it easier

to challenge institutionalized values: it allows greater interaction of citizens, and favors a better organization of social movements. As Gramsci claimed, the cultural hegemonic is a social construct, therefore, there is possibility of change. Through counter-hegemonic actions, new ethical-political forms can be created, where conditions of marginalization and exclusion are denounced and attempted to be overcome, as we have seen before with movements such as black life matters (Peña, 2018).

3 Goal of the study

The main goal of the study is to deeply analyze the topic develop of trend of the red flag hashtag, specifically in the social media "Twitter". Through extracting the necessary data from the period of September 2021 to November 2021. This will allow us to build networks of hashtags that determine social spheres involved. The main topics, distributed in communities, allow the study to identify the acts that are being pointed out as unacceptable in the society, and therefore dictating social norms of how people in a specific situation should act or what to avoid.

The develop of the study turns around the hypothesis that, there is an hegemonic public opinion that has being built based on the phenomena of the trends in social media, having influence in different communities, this can be reflected in the way social norms are translated from the digital world to the behaviors of people in real life communities, excluding and discrediting those who do not comply with the behaviors that are seen as acceptable.
As a consequence of not following the actions considered as accepted, there is a tendency to appeal to the culture of cancellation. Where a member of society is punished by means of exclusion and discrediting.

Having mentioned the hypothesis, we deduce the main research question as: How does social media influence the social norms of a society?
Around this research question with the extraction of the data from twitter, it is intended to analyze the information obtained based on the standards of social behavior and focusing on which specific communities are being targeted and therefore influenced.
Once the standards imposed through specific topics have been recognized, the consequences for society of not complying with the minimum expected behavior, its relationship with freedom of expression and the concept of private and public spheres should be considered.

With the study conducted, it should be possible to recognize whether there is a relationship between the norms imposed by means of the network flags and the cases that come to the culture of cancellation of individuals in society who, due to some behavior considered to be in breach of the imposed standards, are discredited and excluded from the corresponding community.

4 Tweets Download and Pre-Processing

Before any kind of analysis, the first step is of course the download of the data. For the aim of this project, we are interested in all the tweets whose topic is *Red Flag*. In social networks and especially on Twitter, the topics are represented by hashtags. As a consequence, we look for all the tweets using *#RedFlag*. The time window is chosen from September 1 to November 30 2021, since this is the period in which the hashtag *#RedFlag* was trending on Twitter. We look for tweets in English only.

4.1 Tweets Download: the *snscrape* Library

Alessandro Marcomini

Twitter allows to load data using its API, that is made available only for developer accounts. It is very simple to use but is limited in many aspects: maximum 30 requests per minute and 10 per second, with a maximum of 100 tweets per request. Moreover, there is a limit on the number of tweets that one can download in a month.

All the previous limitations disappear using *snscrape* [15], a library that allows anyone to scrape tweets without requiring personal API keys. It can return thousands of tweets in seconds, without limitations, and has powerful search tools that allows for highly customizable searches. The cost to pay is that it is not an official library and so it is currently lacking in detailed documentation. Still, once the main features are understood, it is really fast and convenient to use. For each tweet, we are only interested in its date, text, the user who wrote it and the hashtags it contains. We collect all these features in a dataframe, built as it is shown in Figure 1.

```
# Creating list to append tweet data
tweets_list1 = [] # Using TwitterSearchScraper to scrape data and append tweets to list
for i, tweet in enumerate(sntwitter.TwitterSearchScraper("#redflag lang:en since:2021-09-01 until:2021-12-01")\
                          .get_items()):
    tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.user.username, tweet.hashtags])

# Creating a dataframe from the tweets list above
tweets_df1 = pd.DataFrame(tweets_list1, columns=['Datetime', 'Tweet Id', 'Text', 'Username', "Hashtags"])
tweets_df1.head()
```

	Datetime	Tweet Id	Text	Username	Hashtags
0	2021-11-30 23:11:46+00:00	1465820879661711367	Alert students realize this is not the languag...	BigFish3000	[RedFlag]
1	2021-11-30 22:16:40+00:00	1465807015901483009	A 🤡 doctor that says he's been doing this over...	NetertAsetRe	[RedFlag, NotEveryonesBodyIsTheSame, NoWonderY...
2	2021-11-30 22:16:39+00:00	1465807010188775429	Let's add that there are no educational poster...	NetertAsetRe	[RedFlag]
3	2021-11-30 22:16:38+00:00	1465807007777054728	Also, seeing little papers on the walls asking...	NetertAsetRe	[RedFlag]
4	2021-11-30 22:16:37+00:00	1465807002546761733	I wasn't going to join this #RedFlag discussio...	NetertAsetRe	[RedFlag]

Figure 1: Usage of the *snscrape* library to download tweets containing a specific hashtag in a given language and time window. By including *#redflag* in the query, the function will return all the tweets containing that hashtag, making no difference in capital or lower case letters (it will download also the tweets containing #REDFLAG, #RedFlag and so on). For each tweet, we collect information about date, tweet ID number, text, username and used hashtags.

In this way we build a dataset made up of 5554 tweets related to the RedFlag topic, but before analyzing them it is necessary to pre-process their texts.

4.2 Pre-Processing

Michele M. Crudele

The goal of the pre-processing of the data is to make the tweets as clean as possible, removing all the non-informative content. We use two different cleaning procedures for two different goals: the sentiment analysis and the creation of the networks.

For the former we used tools that can work with almost original texts, so the pre-processing procedure consists only in transforming the *"u"* characters in *"you"* (we understood this operation

was needed *a posteriori*, after having noticed the strong presence of the "u" character in the networks of words) and removing all the emojis from the tweets.

For the creation of the semantic networks of words instead, we would like to have only nouns, adjectives, adverbs and state verbs. As a consequence, in addition to the previous operations about emojis and "u" characters, we remove mentions and links from the text of each tweet, in addition to the hashtags, since we have a dedicated column for them in our dataset. Then, everything is put in lower case. At this point we can use POS-Tagging and Lemmatization to keep only nouns, adjectives, adverbs and state verbs, bringing every word to its root form (plurals to singular and all the verbs in their infinite form). All the punctuation is eliminated. Finally, we remove all the words that are not in the English vocabulary and the stop-words (pronouns, conjunctions, prepositions, auxiliary verbs, ...). We do all these operations thanks to the *re* library [17], that allows to remove mentions, hashtags and URL links, and the *NLTK* library [18], that provides a full list of English words, stop-words and the functions to perform POS-Tagging and lemmatization. An important point to take care of is to perform the transformation in lower case and the lemmatization *before* removing the words that are not in the English vocabulary; this is fundamental because the latter contains only singular nouns and lower case words, so for example all the plural words would be removed from the tweets otherwise. We show an example of the output of the pre-processing function in Figure 2.

```
ORIGINAL TWEET: ►►►Red Flag in a Relationship►►►

When you are constantly made to feel like you are stupid, illogical, or wrong for thinking a certain way.

#redflag #toxicrelationship https://t.co/Wg3gOME8cI

PRE-PROCESSED TWEET: red flag relationship constantly make feel like stupid illogical wrong think certain way
```

Figure 2: An example of the output of the cleaning function for the pre-processing of the tweets.

Let us recap the performed operations in bullet points.
Pre-processing for sentiment analysis:

- Emojis removal
- $u \rightarrow you$ transformation

Pre-processing for networks creation:

- Emojis removal
- $u \rightarrow you$ transformation
- Mentions removal
- Hashtags removal
- URLs removal
- Lowercase transformation
- POS-Tagging and Lemmatization
- Non-English words removal
- Stopwords removal

With these cleaned tweets, we are now ready to build the networks.

5 Networks Creation

5.1 Network of Hashtags

Michele M. Crudele

The first goal of our project is to understand which topics *RedFlag* is related to. To reach this goal, we can build a network of hashtags since once again, in social networks they represent topics.

5.1.1 Network of Hashtags Creation

In order to do that, we collect all the unique hashtags from the tweets in our dataset, building a dictionary in which we collect both the hashtags and the total number of times it is used.

From them, we remove the hashtag #RedFlag because every tweet we have downloaded contains it and so it would result in a huge hub, connected to all the other unique hashtags. In addition to it, we also remove those hashtags that are very often used together with it, but does not contain any significant information about the topic of the tweet, either because they are used as synonymous of the hashtag *#RedFlag* or because they are exploited only to make a tweet more visible on Twitter.

The list of hashtags we remove is then: #RedFlag, #RedFlags, #RedFlagTrend #Trend, #Trends, #TwitterTrend, #TwitterTrends, #TrendigNow.

Another operation we perform is the merge of similar hashtags: #Meme & #Memes & #Memes-Daily &TrendingMemes; #Covid & #Covid19; #Relationship & #Relationships.

Finally, one should also consider the fact that sometimes on social media one unique user can make a lot of posts using the same hashtags, just to make those hashtags and its account visible and trending. As a consequence, we decide to remove from the dictionary of unique hashtags all those hashtags that are used only by one single account. By doing so, we are able to capture in a better way the trending topics on twitter, reducing the presence of nodes and connections due to single accounts only.

In this way we obtain the nodes of the network of topics. Then, it is necessary to create the links between them: we decide to connect two hashtags if they are used together in the same tweet with a link weighted based on the number of times the connection is established. For example, if #toxic and #relationship are used together in 20 different tweets, in our network they will be connected with a link whose weight is 20.

By doing so, we are creating an undirected weighted network, made up of 681 hashtags and 2223 links.

5.1.2 Network of Hashtags Visualization

In order to visualize the network of topics, we use Gephi [19], an open-source and free visualization and exploration software for all kinds of graphs and networks. We have to provide it a list containing the nodes with their labels and a list with the links and their weights, that we build starting from the dictionary of unique hashtags created in section 5.1.1. Once these two lists are imported, we can also compute some statistics about our network: in particular, we use the PageRank algorithm [20] (with damping factor $p = 0.85$ and precision $\epsilon = 0.001$) to make the size of each label a function of its "importance" in the network, while we use Modularity, in particular the Louvain method [21] (with resolution equal to 1 and randomization) to divide them in different colors that represent communities.

By using these two algorithms we are able to assign two different attributes to each node (its PageRank score and the community it belongs to), but given the size of the network, the most difficult task is its visualization. We would like a good-looking and easily-understandable graph,

with well-recognizable communities and strongly connected neighboring nodes. To this end, we use the ForceAtlas algorithm [23] implemented in Gephi. However, this tool is not sufficient to display such a big network with the previously listed features, especially if we also want to display the labels of the nodes. This is why we use ForceAtlas to make a first separation of the nodes and then we move single nodes by hand in such a way to put them inside their community and near to the nodes they are most strongly linked to. With this procedure, we are able to obtain the network shown in Figure 3, where the communities are very well separated and each hashtag is put as close as possible to the nodes it is most strongly connected to. Looking at it, we can distinguish the main topics related to RedFlag (the different colored communities) and inside each one of them we can also spot the most important hashtags based on their size; moreover, the thickness of the links tells us how strong the connection between two nodes is.

We can notice the presence of the very expected *dating* topic in blue, whose main hashtags are *#dating*, *#datingadvice* and *#relationships*. In light green there is the Formula1 community, present because in Formula1 a literal red flag is used to stop a race. A literal red flag is also used in meteorological warnings about fires, whose community is colored in olive in the bottom-right side of Figure 3. The biggest communities are mental health (in orange), business and marketing (in black), crypto (in sea green), social media (in green), Covid-19 (in pink) and abuse (in dark red), where *#2a* refers to the Second Amendment of the United States Constitution, that is about "the right of the people to keep and bear arms". Smaller communities are about climate change (in red), and cybersecurity (in dark blue). There is also a light blue community, but it is not very definite and its topic is not very clear, even though it seems to be about entertainment (music, films, reading, ...), in addition to small and non significant communities, like the dark pink one at the bottom.

In section 5.2 we will choose six of these topics to be analyzed more in details.

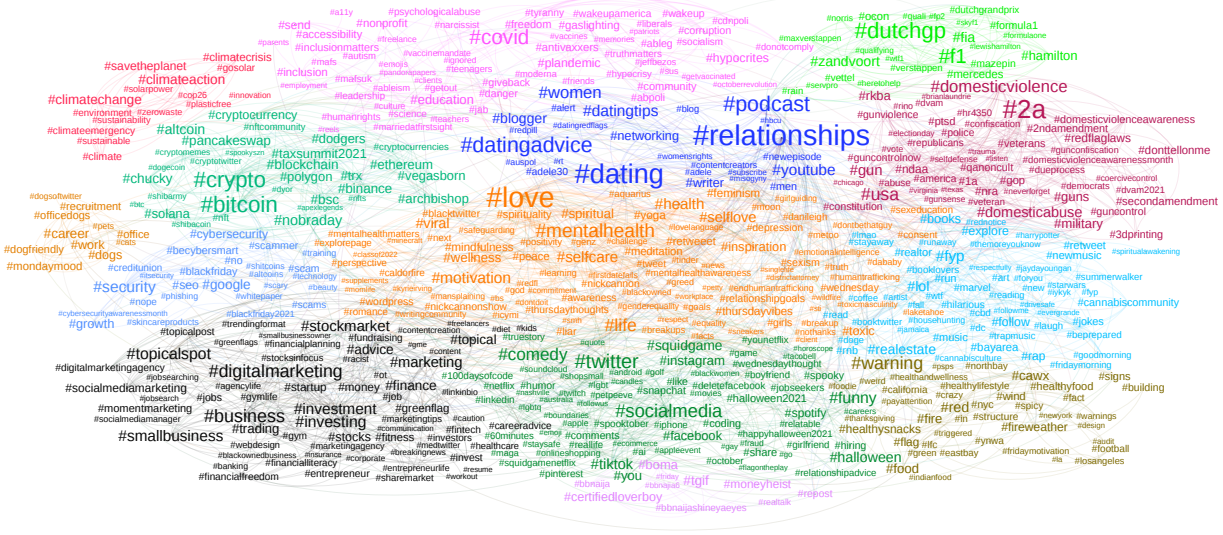


Figure 3: The network of hashtags visualized using Gephi. The size of each label is proportional to its PageRank centrality measure (a measure of its "importance" in the network), while different colors represent different communities found with the Louvain method. The thickness of each edge is proportional to its strength.

5.2 Selection of Topics

Michele M. Crudele

Once the network of hashtags is displayed, we can use it to choose the most interesting topics that we want to analyze deeper.

However we have to consider the following facts: since the Louvain algorithm contains elements of randomness, it will not divide the nodes always in the same way; moreover, an hashtag can be strongly linked to nodes that are in different communities, making difficult to understand the one it belongs to. This is exactly what happens, for example, to the hashtag *#love*, strongly connected both to the *dating* and *mental health* communities. For these reasons, we decide to run the Louvain method a few times: in this way we can understand which are the hashtags that belong for sure to a community and which are the ones that are frequently used in more than one topic.

5.2.1 The Six Topics

Jacqueline Pamela Padilla Torres & Gala Pradillo Diaz

With this in mind, we are able to select wisely the most significant hashtags for the communities we decide to analyze further:

- COVID: #covid19 #gaslighting #pandemic #freedom #vaccines #truthmatters #covid #nonprofit #antivaxxers #community #education #science
- CLIMATE: #climatechange #climatecrisis #suistainable #savetheplanet #zerowaste #climatecrisis #climate #plasticfree
- ABUSE: #domesticviolence #domesticabuse #abuse #guncontrol #2a #guns #gunsense #redflaglaws #coercivecontrol
- MENTAL HEALTH: #mentalhealth #health #feminism #toxic #selfcare #metoo #mentalhealthmatters #facts #girls #motivation #mentalhealthawereness #selflove #depression #run #meme #memes #blacktwitter #life #selfguarding #swipeleft #tinder #wtf #mindfulness #awereness #cannabiscommunity #cbd
- DATING: #dating #relationship #relationships #relationshipadvice #reallife #flag #relatable #stayway #funny #humor #datingadvice #datingtips #comedy #women #men #datingredflags #alert #lgtb #horoscope
- MARKET: #digitalmarketing #entrepreneur #advice #facebook #fraud #trading #trend #money #instagram #twitter #business #linkedin

5.3 Semantic Networks

Lucia Depaoli & Simone Mistrali

After having selected the communities we are most interested in, we can build a network of words for each one of them. The goal is to analyze the words inside each community from a semantic point of view, trying to understand what is considered as a RedFlag in the different topics we analyze. Also, the distribution of the words and, most importantly, the community of the words, can make us understand which are the relevant sub-topics in the tweets.

5.3.1 Semantic Networks Creation

In order to create the network of words, we follow the exact same procedure we did before for the network of hashtag. The only difference is that now we do not consider all the tweets in our dataset, but only those that are related to the community we are investigating. To extract only a set of tweets, we filter the full dataset selecting only those tweets that contain *at least one* of the hashtags of that community. In other words, in this way we are selecting all the tweets related to each topic.

In order to build the semantic networks, we create a dictionary of `unique_words` from the preprocessed cleaned texts and count the number of occurrences for each one of them. We remove the words *red* and *flag* from this dictionary for the same reason we removed some hashtags in the networks of topics: they would be big but non-informative nodes. Each unique word represents a node of the semantic networks. The links are created between the words that are used together in the same tweet, weighted by the number of times they appear in the same tweet. Also in this case we are creating undirected weighted networks.

In table 1 we report the main features of our networks. Regarding the number of tweets, for the second topic there are very few. This surely will make our network less informative than, for example, the DATING one, because all the tweets are probably correlated to the same argument, or maybe written by the same few people. Given the number of nodes and links, we manage to show everything without having to perform cut at any levels. The problem is that the biggest

networks will appear very confused, with a huge number of nodes. Therefore, we divide manually the communities, breaking down the connection between words belonging to different communities.

Community	Tweets	Nodes	Links	Words Communities
COVID	39	234	1306	12
CLIMATE	15	86	443	7
ABUSE	58	396	3261	13
MENTAL HEALTH	136	523	3882	15
DATING	179	688	6637	10
MARKET	80	336	1759	21

Table 1: Values for community networks.

5.3.2 Semantic Networks Visualization

In order to visualize the semantic networks, we proceed using the same rationale explained in section 5.1.2 and here we show the networks that we obtain for each one of the six topics.

Figure 4: Semantic network of COVID topic.

In figure 4 we have the COVID community. Here we report the words communities.

- a VACCINE community (bottom-right, red) → words such as: *vaccination card, vaccine company, falsify (vaccination card)*. The tweets could be about "vaccine is bad", which is a #redflag, or "#redflag if someone claims that vaccine is bad".

- a TOTALITARIAN community (bottom-left, purple) → this community seems to be against vaccine. We can understand this from the words *totalitarian, racism* (maybe against people who do not want to vaccinate themselves), *biologically* (not necessary) and so on. Tweets could be "warning, we are going toward a totalitarian government #redflag".
- a STORE and WORKER community (center, dark green) → words such as: *essential store, safety, mask, crowded (church), pregnant*. The biggest word, *month*, could be related to the others by sentences such as: "stores have been closed for months #redflag".
- a EVERYONE (MUST) community (top-right, brown) → sentences such as: *everyone (must be) allowed to study, everyone (must) accept* and so on.
- a SCHOOL POLICY community (top, light blue) → words such as: *school policy, incorrect, agree, friend*. #redflag could be related to the fact that school should not close due to the pandemic.
- a NO-VAX community (top-left, brown) → words such as: *giant scam, big, inject chemical, smell*. They are probably no-vax tweets who are using #redflag to warn people about the danger of vaccine or about the non-existence of Covid.

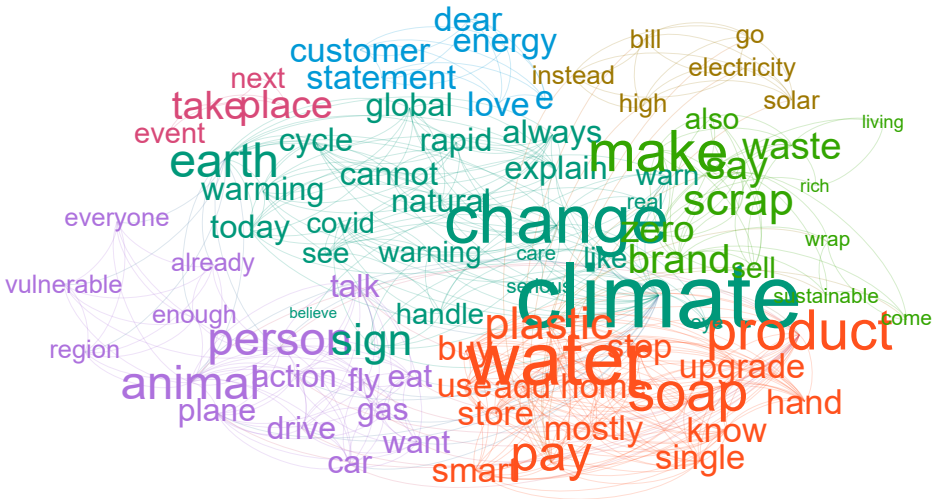


Figure 5: Semantic network of CLIMATE topic.

In figure 5 we have very few nodes compared to the others network, but we have more defined communities.

- a CLIMATE CHANGE community (center, green) → words such as: *climate change, earth change, rapid, global warming*. #redflag about global warming, self-explanatory.

- a PRODUCT community (bottom-right, red) → words such as: *plastic product, hand soap, pay water, buy plastic water.* Probably tweets to incentivized the usage of green recyclable products. #redflag is, as before, used to warn.

- a SUSTAINABILITY community (right, green) → sentences like: *make (product from) scrap, sustainable waste, living sustainable, sustainable brand.* Maybe the tweets are #redflag about products that claims to be sustainable but in reality are not.

- a ELECTRICITY community (top-right, brown) → tweets could contain sentences such as: "electricity bill go high #redflag".

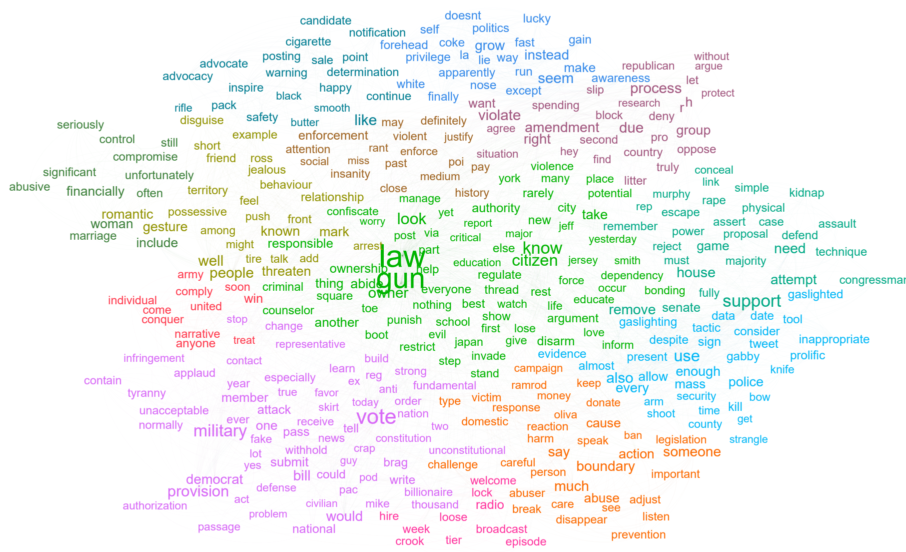


Figure 6: Semantic network of ABUSE topic.

The community in figure 6 is mainly dominated by "gun and law" groups of words. What we can notice is that here we have two main different topic: gun and violence (abuse). Words naturally separate themselves in these two categories.

- a GUN and LAW community (center, green) → words such as: *ownership, school, restrict, educate, inform*, suggest us that they are probably making a law about gun or they are protesting in order to make one. Probably tweets are like "#redlag, we need a law for gun control".

- a VOTE community (bottom, purple) → words such as: *law, military, tyranny, unconstitutional, attack, democrat, news, constitution.* This is a general vote community that seems to include a lot of different topics.

- a CAMPAIGN community (bottom, orange) → sentences such as: *donate money to victim, prevention, abuser say, speak.* They are related to the tweets #domesticviolence #domesti-

cabuse #abuse. Tweets could be something like "#redflag, abuses are increasing, donate to this campaign to help victims".

- a CAPITOL HILL ASSAULT community (right, dark green) → words such as: *kidnap, congressman, support, murply (law)*. In the period of time we have considered, there was a trial against a leader of the assault, so it is possible that these words are related to this.

- a AMENDMENT community (top-right, purple) → sentences such as: *violate the second amendment*. Probably pro-gun tweets, and the #redflag is about the gun laws.

- a JEALOUSY community (top-left, gold) → words such as *friend, jealous, relationship, push, romantic gesture, threaten*. Related to "domestic abuse" hashtags. Tweets could be something like "#redflag if your boyfriend/girlfriend is possessive in front of people".

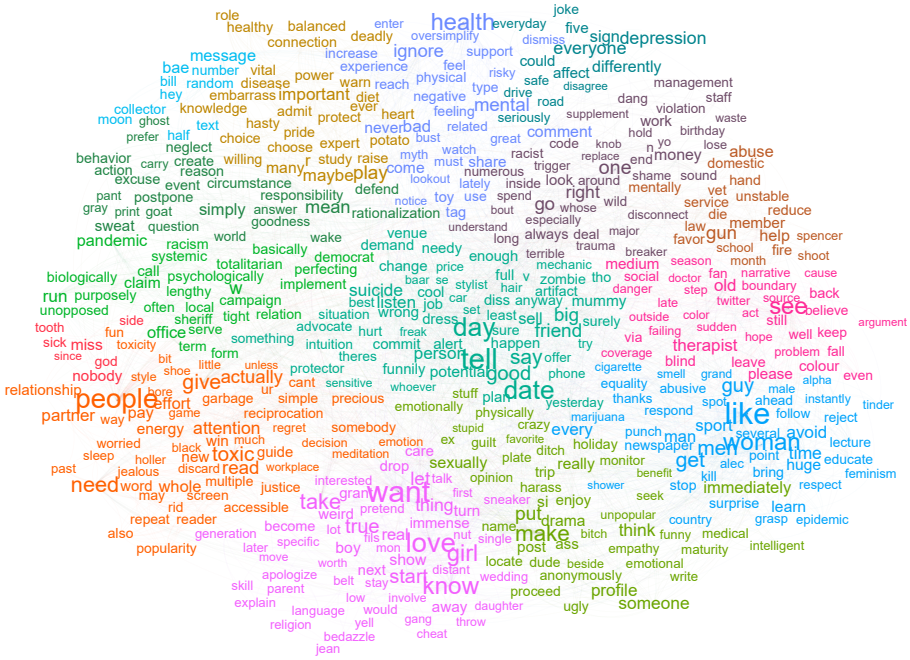


Figure 7: Semantic networks of MENTAL HEALTH topic.

In figure 7 we can distinguish two big community: mental health and girls related topics.

- a WHAT GIRLS/BOYS LIKE community (bottom-right, light blue) → words such as: *tinder, cigarette, educate, equality, marijuana, feminism.* #redflag could be, for example, about cigarette, or feminism.

- a SEE A THERAPIST community (right, pink) → sentences such as: *problem, argument, medium, danger, old.* But it is not a well-defined community. Maybe tweets are about "#redflag if he/she sees a therapist", but it does not seem so clear.

- a ABUSE community (right, brown) → words such as: *gun fire, domestic abuse, law.*

- a DEPRESSION and MENTAL HEALTH community (top, blue) → words such as: *everyone could (manifest) sign (of) depression, mental health, ignore mental health, support* and so on. #redflag is used to show a sign of depression or mental health problem.
- a WEIGHT community (top-left, brown) → words such as: *pride, (couch) potato, diet, plat, healthy, balanced*. #redflag could be used to report a controversial diet.
- a COVID community (left, light green)→ words such as: *pandemic, totalitarian, biological, racism.*
- a TOXIC RELATIONSHIP community (bottom-left, orange) → words such as: *people, give attention, need attention, toxic relationship, effort, jealous, popularity.*
- a GIRL community (bottom, pink) → tweets could be something like "#redflag if a girl says she want true love" or "#redflag if she says she wants a marriage and children".

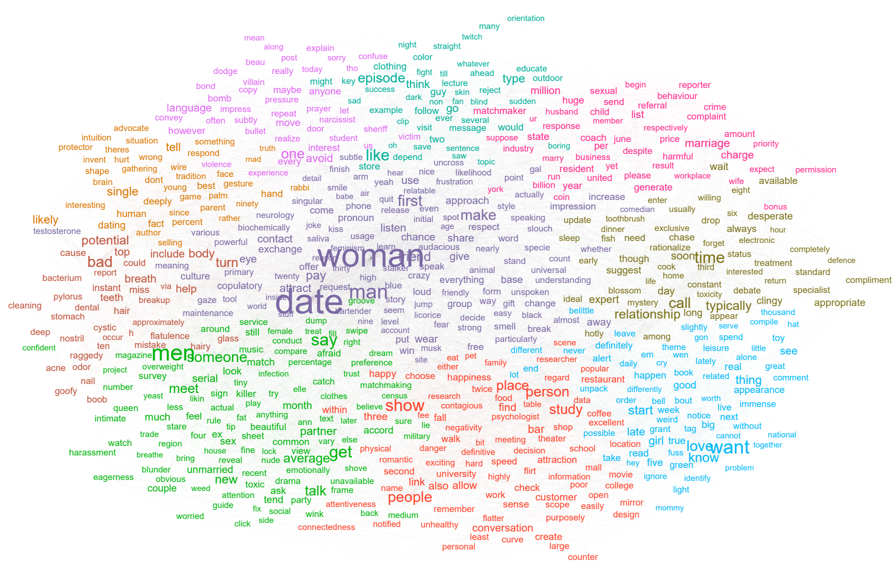


Figure 8: Semantic network of DATING topic.

This topic in figure 8 is dominated by the words *men, woman, date.*

- a FIRST DATE community (center, purple) → words such as: *first date, contact exchange, man (does not) offers pay, impression, man (does not) listen.* Tweets seem to be about what is a #redflag on a first date.
- a PLACE community (bottom, red) → Words such as: *show, place, bar, conversation, poor restaurant.* Maybe talking about when the first date is a #redflag based on the place.
- a TIME community (right, gold) → sentences such as: *time (to) call, desperate, six hours, appropriate.* Probably things as "if she/he calls you after six hours from the first date, she/he is desperate and this is a #redflag".

- a RELATIONSHIP ADVICE communiy (top-right, pink) → words such as: *harmful behaviour, change behaviour, husband, child, price, permission.*

- a BAD and BODY community (left, brown) → words such as: *bad breath, bad body, potential body, teeth, stomach.* Self-explanatory.

- a MEN community (bottom-left, green) → sentences such as: *men say, men talk, fat men, average men unmarried, men look (for) someone (just for) sex.*

Figure 9: Semantic network of the MARKET topic.

In figure 9 we have the words network of the market topic. This is more difficult to analyze, since probably tweets are talking about a particular event.

- a BUSINESS community (center, light blue) → words such as: *business, job, company.* This is a general community.

- a ENVIRONMENTAL community (right, green) → words such as: *food, specie, water, support, date, warning, indicator.* #redflag could be about environmental issues.

- a PRINCE community (top, light blue) → sentences such as: *perfect prince, charming prince, perfect relationship, nobody, read, article.* Maybe the tweets are talking about "if he seems the perfect prince, then it is a #redflag". Or maybe "if he is too charming, it is a #redflag". These could be related to hashtags #advice, #money, #fraud.

- a MARRIAGE community (top-left, red) → sentences such as: *husband allow permission, suppose.* They could something like: "if your husband has to give you permission, then this is a #redflag".

- a COMPUTER community (left, dark green) → words such as: *computer use, screen use, intern screen use, everybody use (the) computer*. Tweets could be about "#redflag about the time spent on computer".

- a TREND community (left, red) → words such as: *today trend child, twitter, dream (of) children, make (a) trend, invest (in) twitter*. The tweets could be about "be careful about what are the new trends" or "#redflag about investing in children (young influencers)".

5.3.3 WordCloud

Alessandro Marcomini

In the following paragraph we report the words clouds we created for the semantic analysis carried out in section 5.3. In each case, we present both the original and cleaned version of the cloud. The latter has been achieved by performing the pre-processing analysis as illustrated in section 4. There is a noticeable improvement in the visualization for the cleaned version: in fact, filtering out unwanted words allows to highlight at best the community-leading terms.

Figure 10: Words cloud of COVID topic (left: original, right: cleaned).

In figure 10 we can notice how there is a strong correspondence (how it is expected) between the highest frequency words in the cloud and the dominant ones the the communities. In particular, apart from the two hubs "red flag" and "covid", we recognize topics like "school", "vaccine", "everyone"/"someone" which are leading in the community analysis.

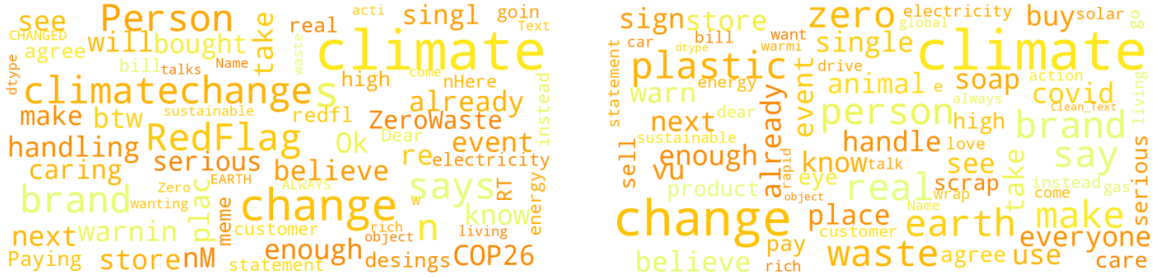


Figure 11: Words cloud of CLIMATE topic (left: original, right: cleaned).

As for figure 11, the situation is similar to the one above: "climate" is a dominant term, together with topic words like "change", "plastic" (product) and "waste".

Figure 12: Words cloud of ABUSE topic (left: original, right: cleaned).

For the "abuse" topic, we see "law" and "gun" among the most used words. This is in correspondence with the first community. Other much used terms are the ones referring to elections: "vote", "support", "democrat" and "republican".

Figure 13: Words cloud of MENTAL HEALTH topic (left: original, right: cleaned).

Here, in line with figure 7, the vast majority of references are from the relationships world: "woman", "relationship", "respect", "reciprocation", "toxicity" tower among others. We also find correspondence of secondary communities, such as "school" and "holidays".

Figure 14: Words cloud of DATING topic (left: original, right: cleaned).

Coming to dating, we see some similarities with the previous topic (in words as "woman" and "respect") but also activities and objects related to relationship life ("episode", "toothbrush", "date",...).

Figure 15: Words cloud of MARKET topic (left: original, right: cleaned).

Finally, the "market" topic displays a large variety of widely-used words belonging to very different communities. This is probably the field in which we see the most differences among dominant terms, as we could expect given the numerous declination this topic can have. We move from people-related words ("child", "someone") to technical terms ("finances", "client"), from active verbs ("evaluate", "research", "support") to temporal information ("today", "tomorrow", "date").

6 Control Group Analysis

Simone Mistrali & Lucia Depaoli

Is #redflag a relevant hashtag in the period of time we have considered? In other words, does it create a bias in the words networks? We want to investigate this aspect, in order to understand if tweets with and without #redflag are similar or not.

We download 6 "Control Group", meaning sets of tweets containing the main hashtags for each community, but that do not contain the hashtag #redflag. This is useful to see if the hashtag #redflag changes the text of the tweets, i.e. if the Twitter users in the period we have considered talk about the same topics regardless of the hashtag #redflag.

Each Control Group have been cleaned from the #redflag in order to take out the redundancies in the two datasets. Moreover, the duplicates of the tweets have been removed, because same tweets may occur more than one time in the same community.

6.1 Control Group Analysis

As before, in table 2 we report the main features of our networks. We do not report all the nodes and edges because then it would be impossible to visualize and also meaningless. We filter the edges weight, selecting links with $\geq$ than 10 and 15 (for the biggest network). After applying this filter, we remove the nodes with 0 degree. Finally, we eliminate the smallest communities. The numbers in parenthesis are the original values of the feature, before the filtering.

Community	Nodes	Links	Words Communities
COVID	634 (19425)	4543 (2173863)	14
CLIMATE	1124 (9658)	9524 (510275)	11
ABUSE	1235 (9477)	12634 (547274)	7
MENTAL HEALTH	1245 (16026)	11645 (1193382)	10
DATING	778 (11419)	3050 (446165)	7
MARKET	1438 (10961)	14547 (569530)	10

Table 2: Values for Control Group network.

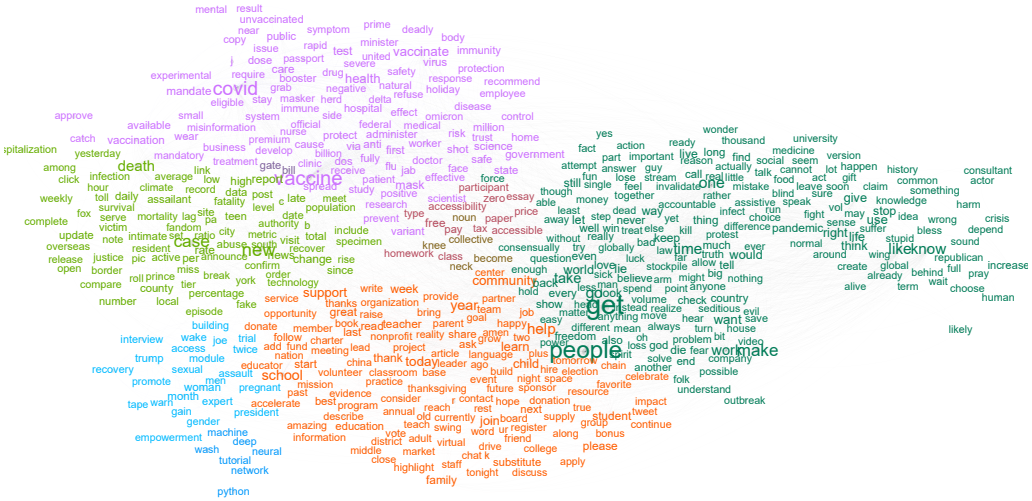


Figure 16: Semantic network of COVID Control Group topic.

In figure 16 it is presented the Control Group of the VACCINE topic. Here are presented the communities that are partially similar to the ones in the #redflag tweets:

- a VACCINE community (top, pink) → words such as: *vaccine, covid, health, misinformation, vaccination, science, government, variant, safety, flu.* A lot of words appear also in the #redflag tweets, but the topics can be very different.

- a SCHOOL and TEACHING community (bottom, orange) → words such as: *teacher, support, child, ask, help, education, best, school, program.* This topic is present also in the #redflag community, but with different words.

Here topics not related to the #redflag tweets:

- a PEOPLE community (right, dark green) → words such as: *work, university, money, time, single, history* and so on. Maybe the topics are about how the pandemic changes the way of living.

- a TRUMP community (bottom-left, light blue) → words such as: *interview, Trump, sexual, president, assault, pregnant.*

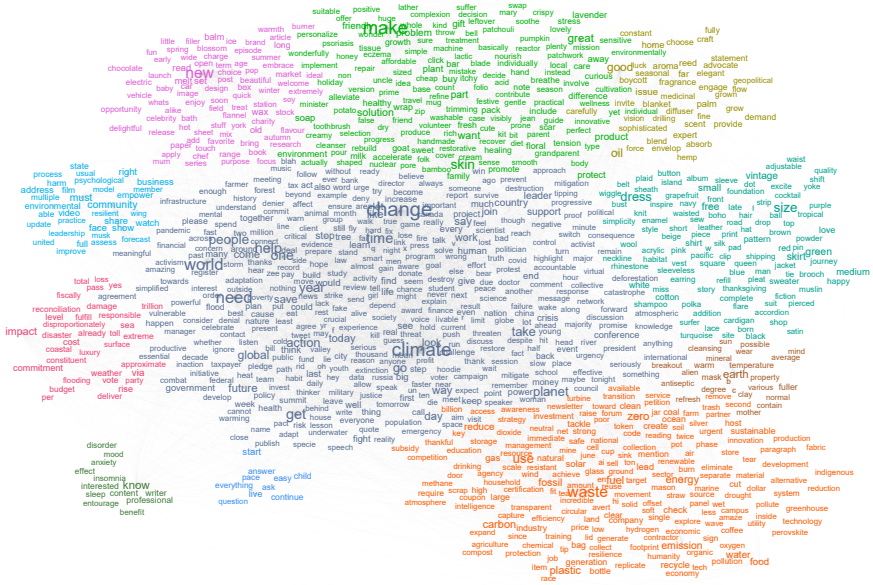


Figure 17: Semantic network of CLIMATE Control Group topic.

For the CLIMATE topics, we have the network shown in figure 17. The communities similar to the #redflag tweets are:

- a PLANET CHANGE community (center, blue) → sentences such as: *time to change, planet needs you, the world needs you, the future need you, global catastrophe* and so on. This topic is very similar to the one found in the previous section.

- a GREEN PRODUCT community (top, green) → words such as: *skin, solution, cream, soap, product.* Similar to the PRODUCT community and the SUSTAINABILITY community.

- a ENERGY and SUSTAINABILITY community (bottom, orange) → words such as: *waste of energy, carbon, reduce, zero emission, recycling, zero waste.* Like the one above, this is a mix of PRODUCT and SUSTAINABILITY community.

And here we have the communities not related to the ones found in the previous section:

- a BOYCOTT community (top-right, gold) → words such as: *oil issue, advocate, absorb (emission).*

- a CLOTHES community (right, blue) → words such as: *vintage clothes, shirt, green dress.*

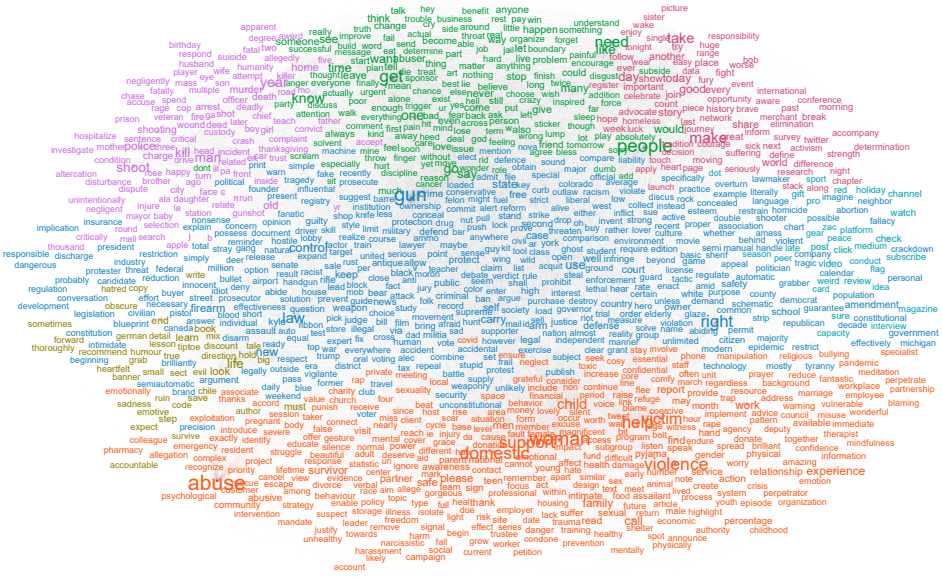


Figure 18: Semantic network of ABUSE Control Group topic.

In figure 18 is shown the ABUSE topic network. The graph is very large and we can clearly distinguish the two main topics (also found above): abuse and gun. Here we have the similar communities:

- a GUN and LAW community (center, blue) → words such as: *control, right, court, constitutional, racism, law, new, firearm, homicide, school.* This community is very similar to the GUN and LAW community.
- a WOMEN and DOMESTIC ABUSE community (bottom, red) → words such as: *abuse, violence, survivor, call, experience, relationship, signal, call, workplace.* This has some similarity with CAMPAIGN community.

Whereas, here we shown the communities not present in the previous section:

- a POLICE VIOLENCE community (top-left, purple) → words such as: *shoot, kill, year, station, accuse, prison, veteran.*
- a PEOPLE community (top, green) → this is a confused community, so many different words and we can not extract any relevant information.

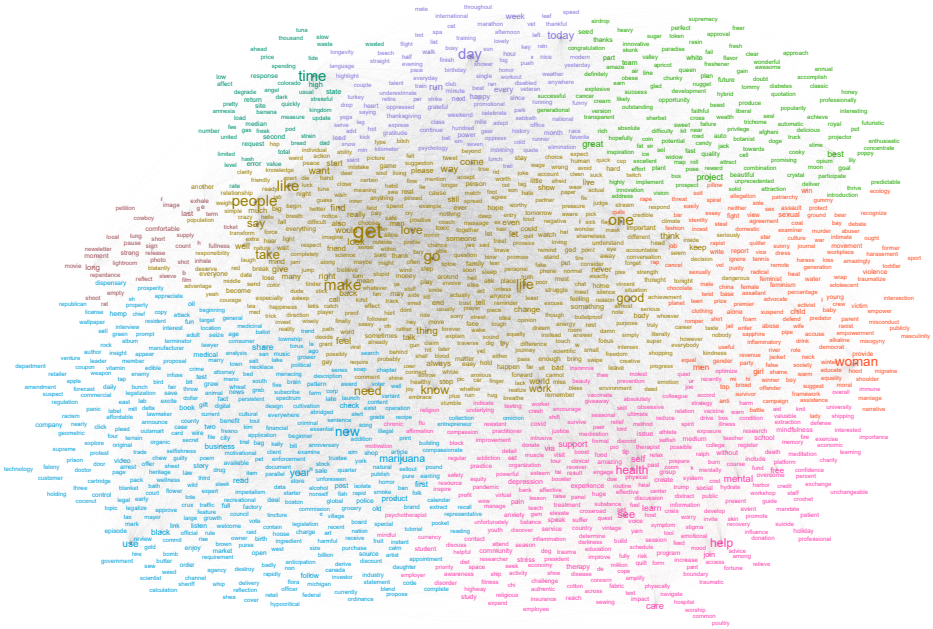


Figure 19: Semantic network of MENTAL HEALTH Control Group topic.

In the tweets containing #redflag, this community has so many more words topics. In figure 19 we shown the MENTAL HEALTH community without the #redflag hashtag. The only similar topic is:

- a MENTAL HEALTH community (bottom-right, pink) → words such as: *find support, help, care, therapy, pain, lesson, experience, depression.* This is very similar to the DEPRESSION and MENTAL HEALTH communities find in the #redflag tweets.

Whereas, the other topics are:

- a PHYSICAL ACTIVITY community (bottom-right, pink) → sentences such as: *run, be happy, run every day, workout, wake up at seven.*
- a FEMINISM community (right, red) → words such as: *women, man, child, feminism, violence, movement, radical, patriarchy, sexual assault* and so on.
- a LIFE community (center, gold) → this is a very confusing community. Words such as: *coach, message, positive, happiness, choice, keep* seems to suggest a positive lifestyle, but it is a very huge community.
- a MARIJUANA community (left, blue) → words such as: *new year, oil, share medical drug, police.*

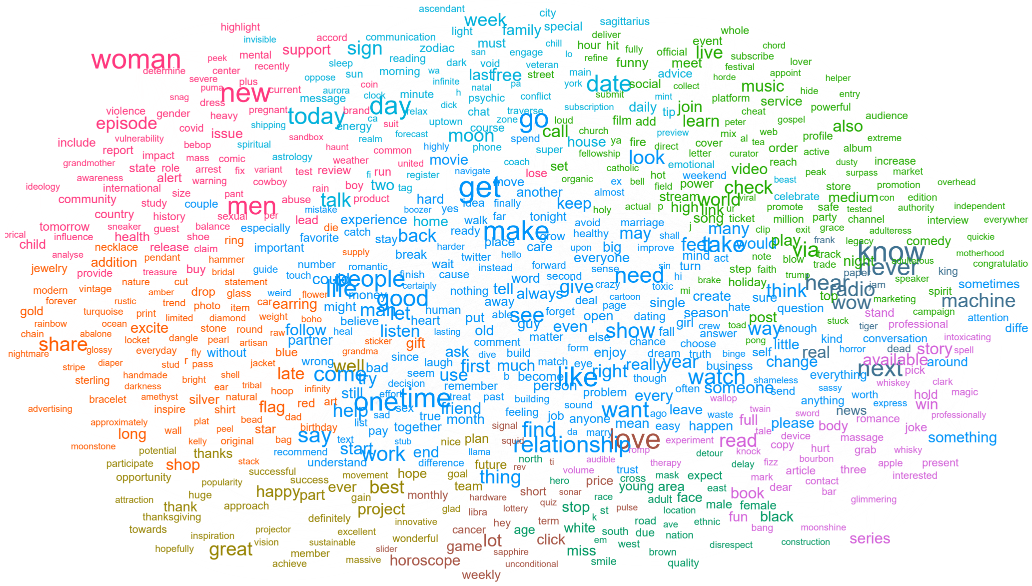


Figure 20: Semantic network of DATING Control Group topic.

Here in figure 20 is the DATING community. As before, the only similarity with #redflag tweets is:

- a RELATIONSHIP community (center, blue) → words such as: *relationship, like, make, people, ask first date , take, need find relationship, first, movie.* This is similar to FIRST DATE community and RELATIONSHIP ADVICE community in the #redflag tweets.

And here we have the different topics:

- a WOMEN and MEN community (top-left, pink) → the main words are: *woman, new, men, episode, community, issue.*
- a ZODIAC SIGN community (top, light blue) → words such as: *today, day, week, sign, zodiac, ascendant, sagittarius, spiritual, astrology.*
- a INTERESTS community (top-right, green) → words such as: *music, live, event, check, world, song, comedy, stream.*
- a RACISM community (bottom, green) → words such as: *black, white, mask, face, stop, race, ethic.*
- a FUTURE PROJECT (bottom-left, gold) → words such as: *project, plan, future, great, happy, success, best part, opportunity, popularity.*
- a PRESENT community (left, orange) → words such as: *share, earring, silver, bracelet, vintage, gold, gift.*

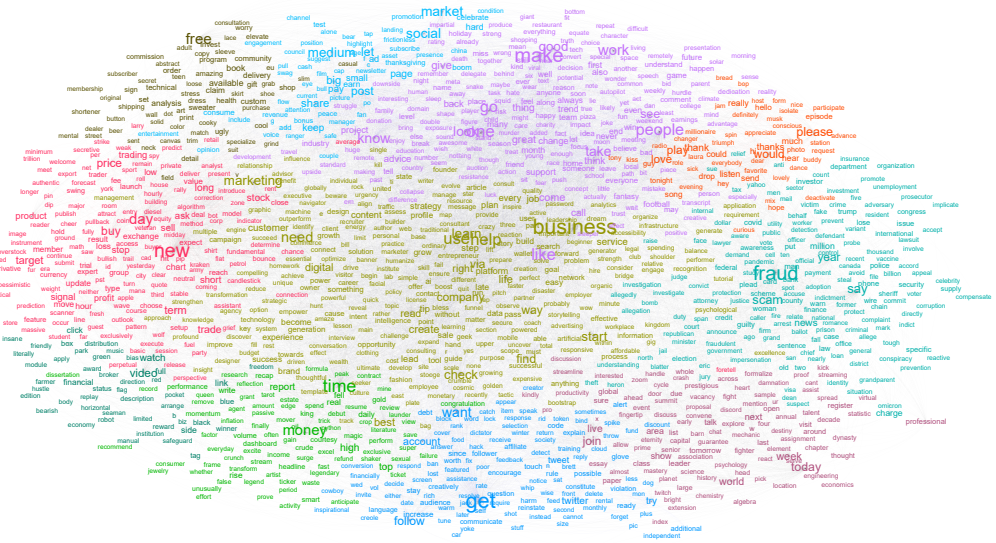


Figure 21: Semantic network of MARKET Control Group topic.

In 21 is shown the MARKET words networks. As before, we have few similar communities:

- a BUSINESS community (center, gold) → words such as: *business, marketing, company, create, start, customer, find, data, service, goal.* Similar to the BUSINESS community of #redflag tweets.
- a TWITTER community (bottom, light blue) → words such as: *get account, twitter, follower, want, feedback, society, code.* It could be compared to the TREND community.

And many different communities.

- a MONEY community (bottom, green) → words such as: *high, money, top, income, daily report, rise.*
- a FRAUD community (right, teal) → words such as: *fraud, scam, control, million, investigation, covid, news, criminal, security.*
- a SOCIAL and MARKET community (top, light blue) → words such as: *market, social, medium, share, big, pay, small, earn, subscribe.*
- a FREE community (top-left, black) → sentences such as: *free book, free program, amazing program, amazing book, free shipping, free membership.*
- a NEW community (left, red) → sentences such as: *new day, new product, buy, result, signal, profit, stock, value, price, trading.*

At first sight, we can say that the topics that have more similarities between the #redflag tweets and the without-#redflag tweets are: VACCINE (but in a different way), CLIMATE (this is the most similar one), ABUSE (main arguments, such as gun, law and domestic abuse). For the others, the topics are consistently different.

7 Degree distribution

Camilla Quaglia

An useful way to have more insight about the properties of a network is to visualize its degree distribution. The latter, denoted in the following as P_k, provides the probability that a randomly selected node in the network has degree k. In particular we visualize the degree distribution of the words networks in each of the aforementioned communities. This allows us to inspect if the networks are scale-free and it is linked to a following analysis about *network robustness*.
In figures 22 and 23 the degree distributions, in log-log scale, are shown.
In particular in the left panel one can see a similar trend in every community: the presence of a plateau in correspondence with high degree nodes and a power-law behavior i.e. $P(k) \sim k^{-\gamma}$.

7.1 Maximum likelihood estimation of γ

In the left panel of figures 22 and 23 also a vertical line is plotted. It represents the k_{min}, the lowest extreme of the interval where the behavior of the log-log plot is approximately linear. The value is chosen *ad hoc* for each community and displayed in the corresponding plot. The purpose of this choice lies in the approach for estimating γ discussed in this section: the maximum likelihood.

In the region in which the log-log plot P_k vs k is approximately linear, the behavior can be described by the equation

$$P_k = Ck^{-\gamma} \longrightarrow C = (\gamma - 1)k_{min}^{\gamma-1} \tag{1}$$

where C is determined by the normalization condition i.e. $\int_{k_{min}}^{\infty} P_k dk \stackrel{!}{=} 1$.

Therefore the target pdf to which apply the maximum likelihood approach to find the best parameter γ is

$$P(k|\gamma) = (\gamma - 1)k_{min}^{\gamma-1}k^{-\gamma} \tag{2}$$

The goal is to maximize the logarithm of the target pdf above

$$\sum_i lnP(k_i|\gamma) \tag{3}$$

where k_i is the degree of node i. From the calculation one gets

$$\boxed{\gamma = 1 + \frac{\sum_i 1}{\sum_i ln(k_i/k_{min})}} \tag{4}$$

Applying the boxed equation to each community we get the results shown in table 3.

7.2 CCDF for estimating γ

To estimate the *gamma* exponent of the power-law in an another way, the complementary cumulative distribution function is computed and represented in the right panel of the figures 22 and 23. The plot is always in log-log scale and the CCDF is fitted. The coefficients of the linear fit are displayed in each graph. The latter representation is particularly useful for the sake of the fit, since it allows to eliminate the presence of the plateau. To retrieve γ from the coefficient of the fitting line μ the following rule is used:

$$CCDF(k) \sim k^{-(\gamma-1)} \longrightarrow \mu = -(\gamma - 1) \tag{5}$$

Community	Number of nodes	γ (CCDF fit)	γ (ML estimation)
Covid	234	2.97	2.97
Climate	86	2.69	2.59
Abuse	396	2.85	2.67
Mental health	523	2.85	2.74
Dating	688	2.72	2.77
Market	336	2.91	2.91

Table 3: Communities and γ parameter

From the table 3 one can conclude that the values of γ lie in the typical range for real networks (e.g. Science collaboration ($\gamma = 3.35$), Protein interaction ($\gamma = 2.89$), Actor network ($\gamma = 2.12$)..), that is $\gamma \in$ [2,4]. Furthermore one can notice the compatibility between the two estimations of γ obtained with the aforementioned approaches. To conclude, all the words networks lie in the scale-free regime being $\gamma \in [2, 3]$, as the *robustness* analysis confirms. Moreover in the scale-free regime one expects the presence of large hubs, as it is indeed confirmed by the *centrality* analysis.

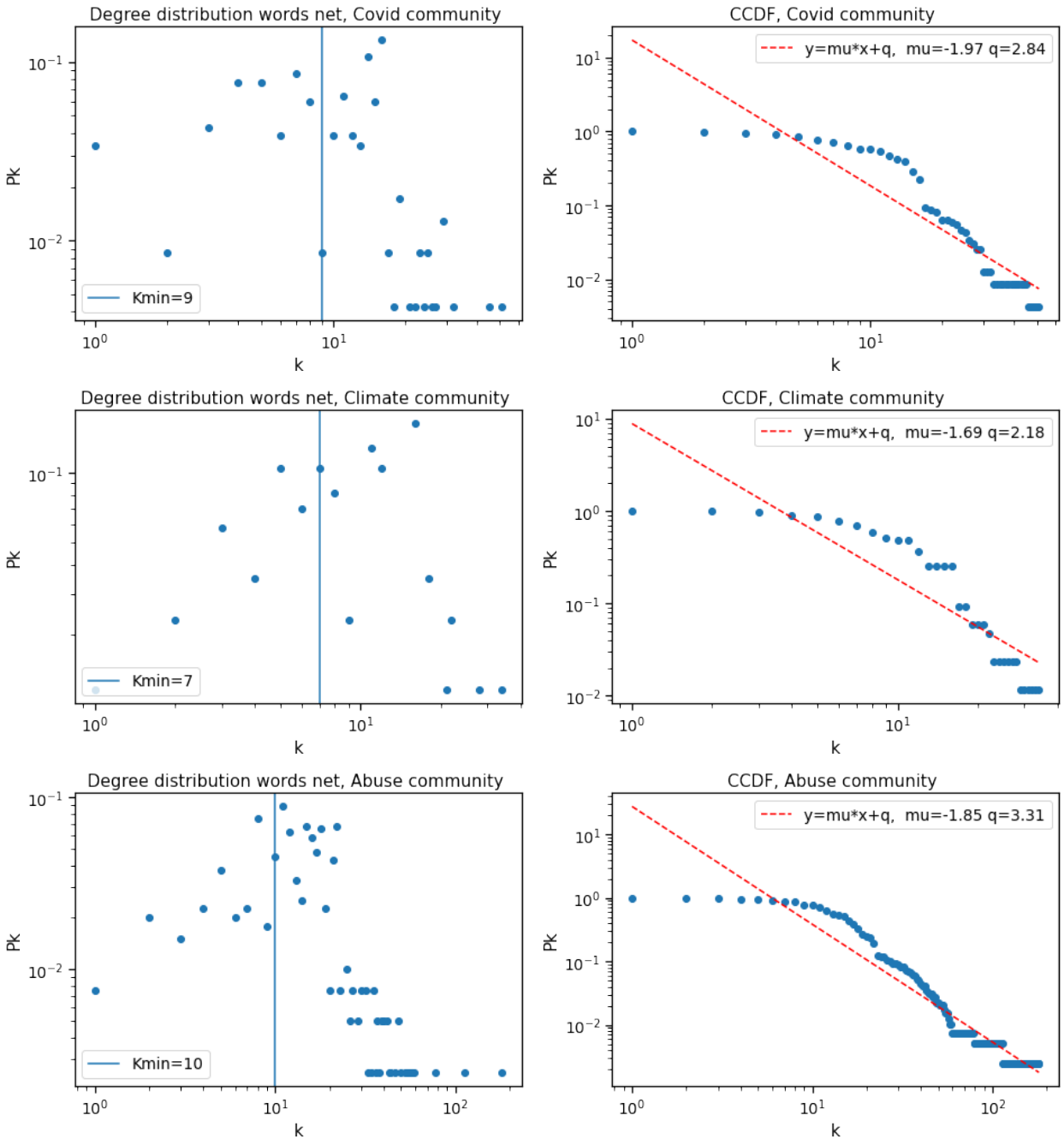


Figure 22: Degree distribution (left panel) and CCDF (right panel). Both the plots are in log-log scale.

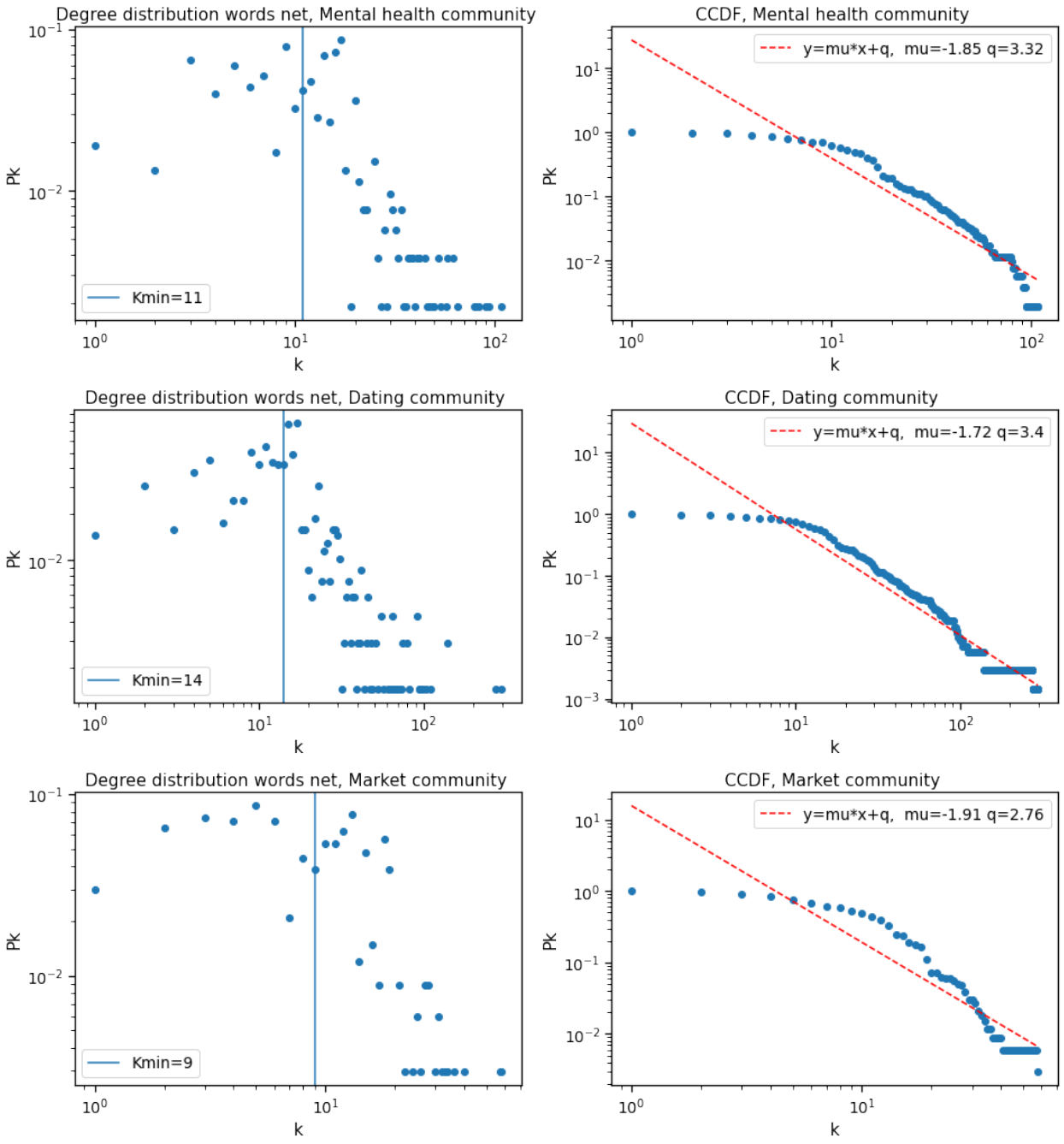


Figure 23: Degree distribution (left panel) and CCDF (right panel). Both the plots are in log-log scale.

8 Assortativity analysis

Camilla Quaglia

In this section we inspect the assortativity in the words networks for each of the six communities. By assortativity we mean the *degree of homophily* in the network. In an assortative network high degree nodes connect with each other, avoiding low degree nodes. In a disassortative network one

can observe the opposite trend, while a neutral network shows a random behavior.
Since we work with social networks we expect our networks to be assortative. The property is measured in two ways:

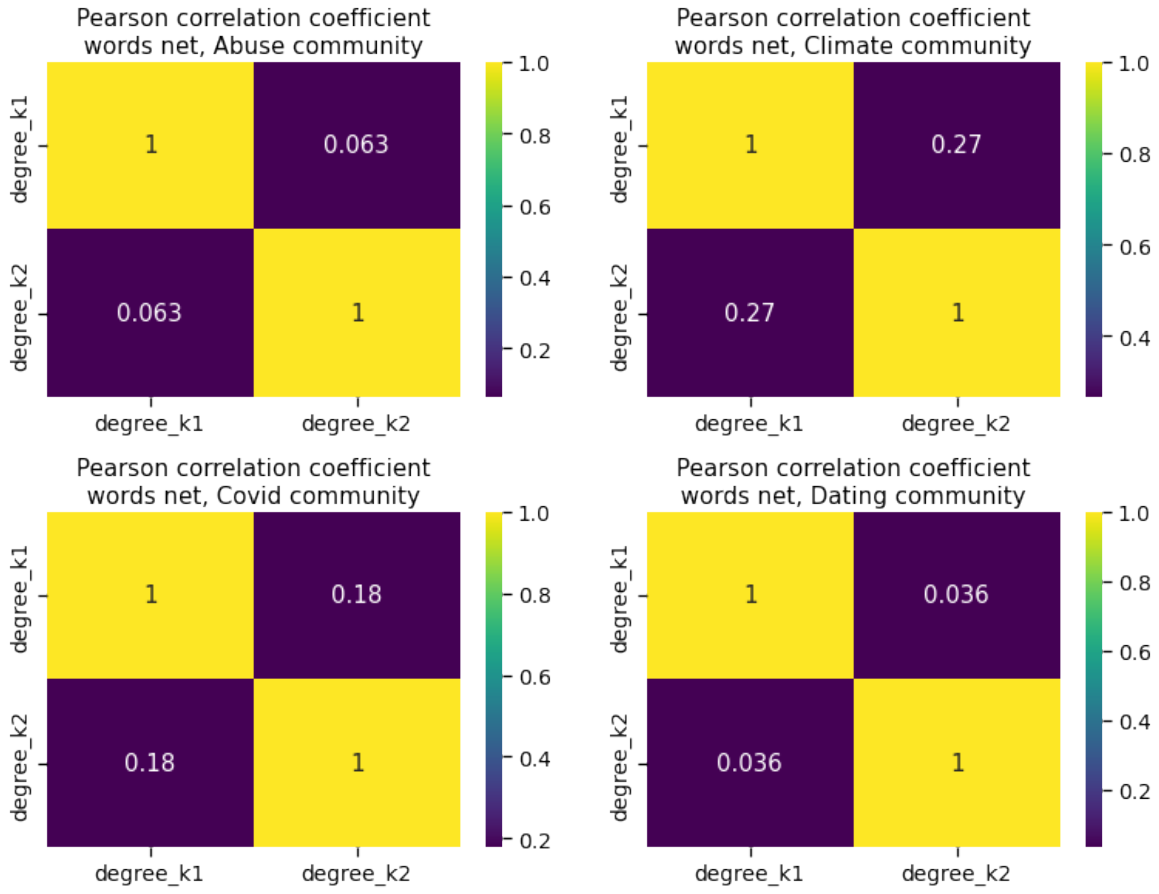
- through the Pearson correlation coefficient
- through fitting

8.1 Pearson correlation coefficient

One first way to measure if the networks are assortative, disassortative or neutral is to inspect the Pearson correlation coefficient among the degree of the i-th node and its neighbor degrees. Considered the pair of words x_i and y_i the Pearson coefficient is computed as follows

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

where n is the total number of nodes. The results are shown in figure 24



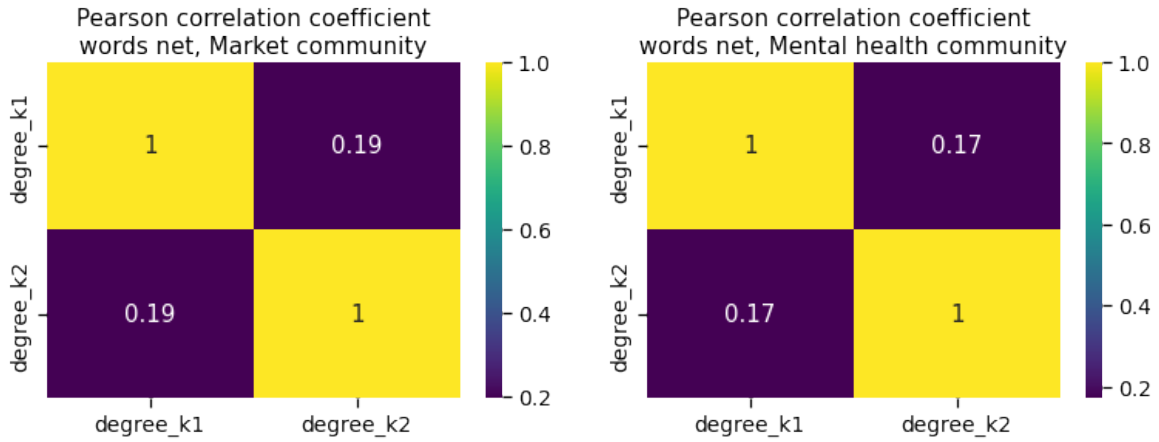


Figure 24: Pearson coefficient for each community.

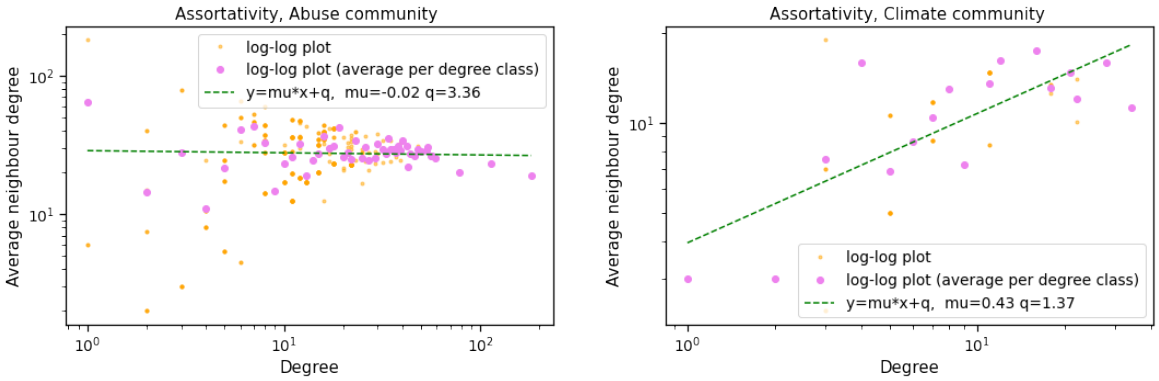
8.2 Assortativity coefficient through fitting

The idea of this way of measuring assortativity is to inspect the degree of neighboring nodes of the vertices of our network. In figure 25 , for each community, the average neighboring degree (y axes) vs the degree (x axes) is plotted, showing the behavior represented by the orange points. In order to fit the points, all the degree of the nodes are grouped per "degree classes" and the corresponding y value is obtained averaging the previous values. With the latter method the violet points are obtained and fitted following the rule.

$$ln(K_{nn,i}) = \mu \cdot ln(k_i) \tag{7}$$

where $K_{nn,i}$ and k_i are respectively the y and x points of our plots.
Now, if

- $\mu > 0$, the network is assortative
- $\mu = 0$, the network is neutral (random)
- $\mu < 0$, the network is disassortative



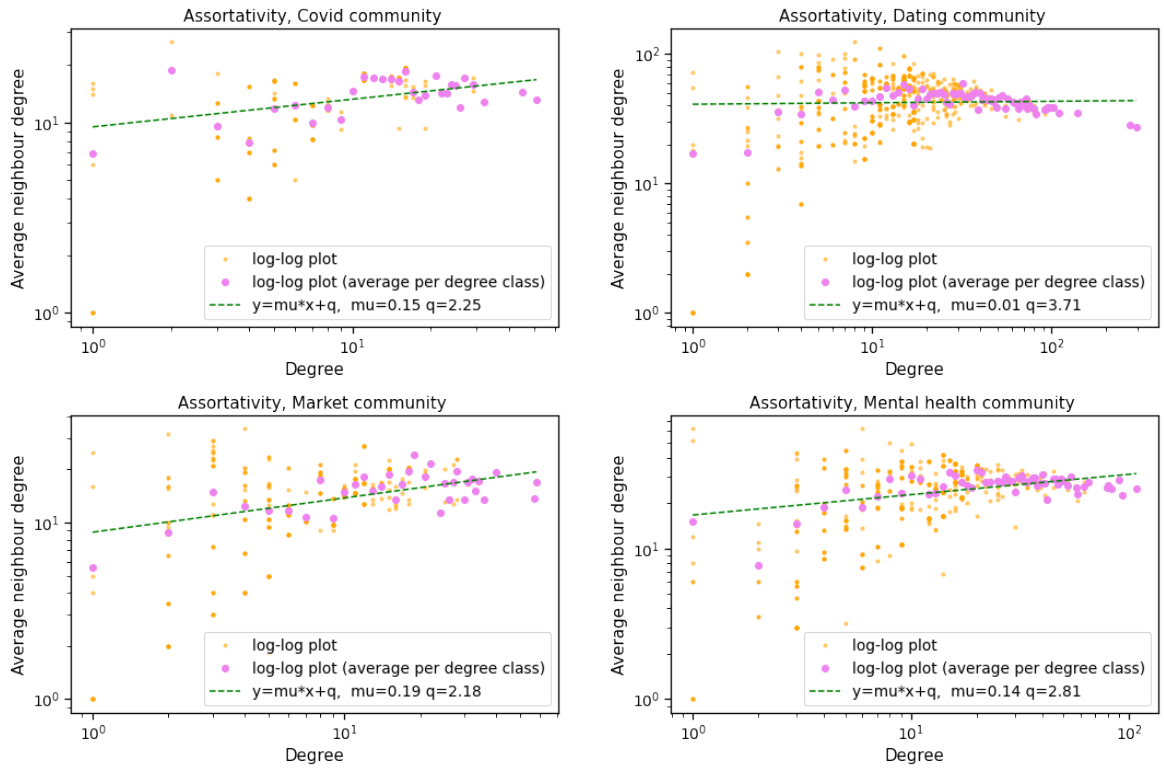


Figure 25: Average neighboring degree vs degree (log-log plot).

From the values of the coefficient μ of the fitting line in figure 25, can be noticed that all the networks of words are assortative, with the exceptions of *Abuse* and *Dating* communities.

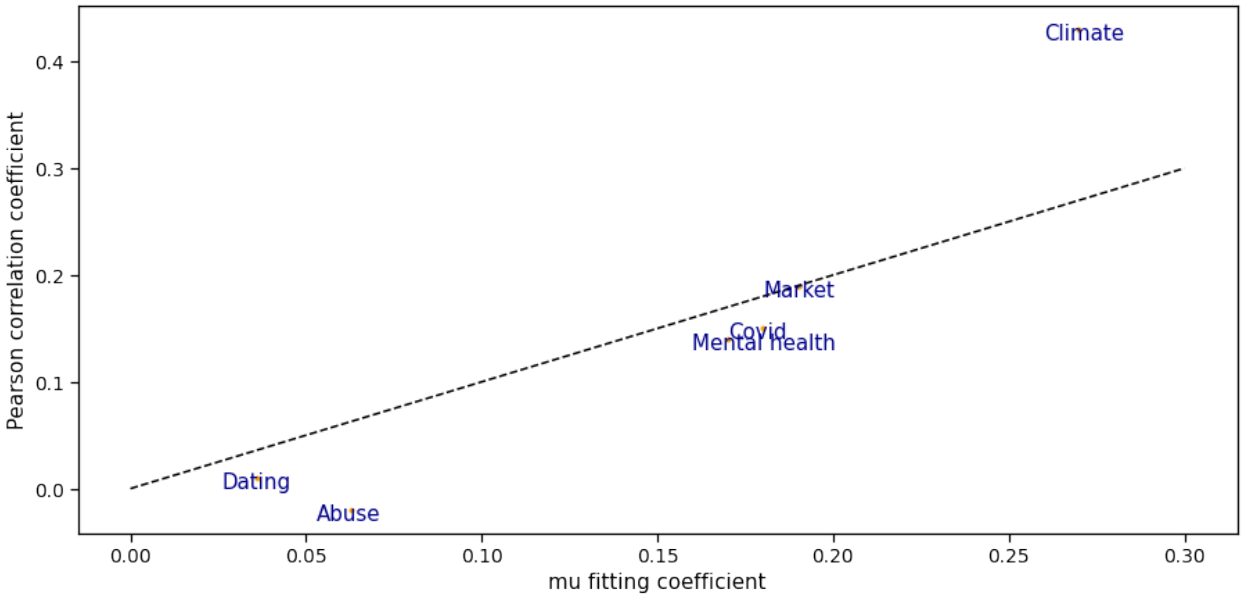


Figure 26: Comparison between the two methods

The assortativity coefficient μ estimated through fitting and the Pearson correlation coefficient

assume coherent values, despite the fact they are not exactly the same, except for the *Market* community, as figure 26 shows.

To inspect also structural disassortativity the networks are randomly rewired (with multiple seeds) preserving node degrees and the results of one random trial are shown in figure 27. I specify that the behavior of the random rewired points (in green) is qualitatively the same for all the seed tried, 3 in total.
Looking at the plot we can conclude that:

- The *Abuse* community is neutral and this seems a property of the degrees. Same for the *Dating* community.
- The *Climate* community is assortative and this is a property of the network. Same for the *Covid*, *Market* and *Mental health* communities.

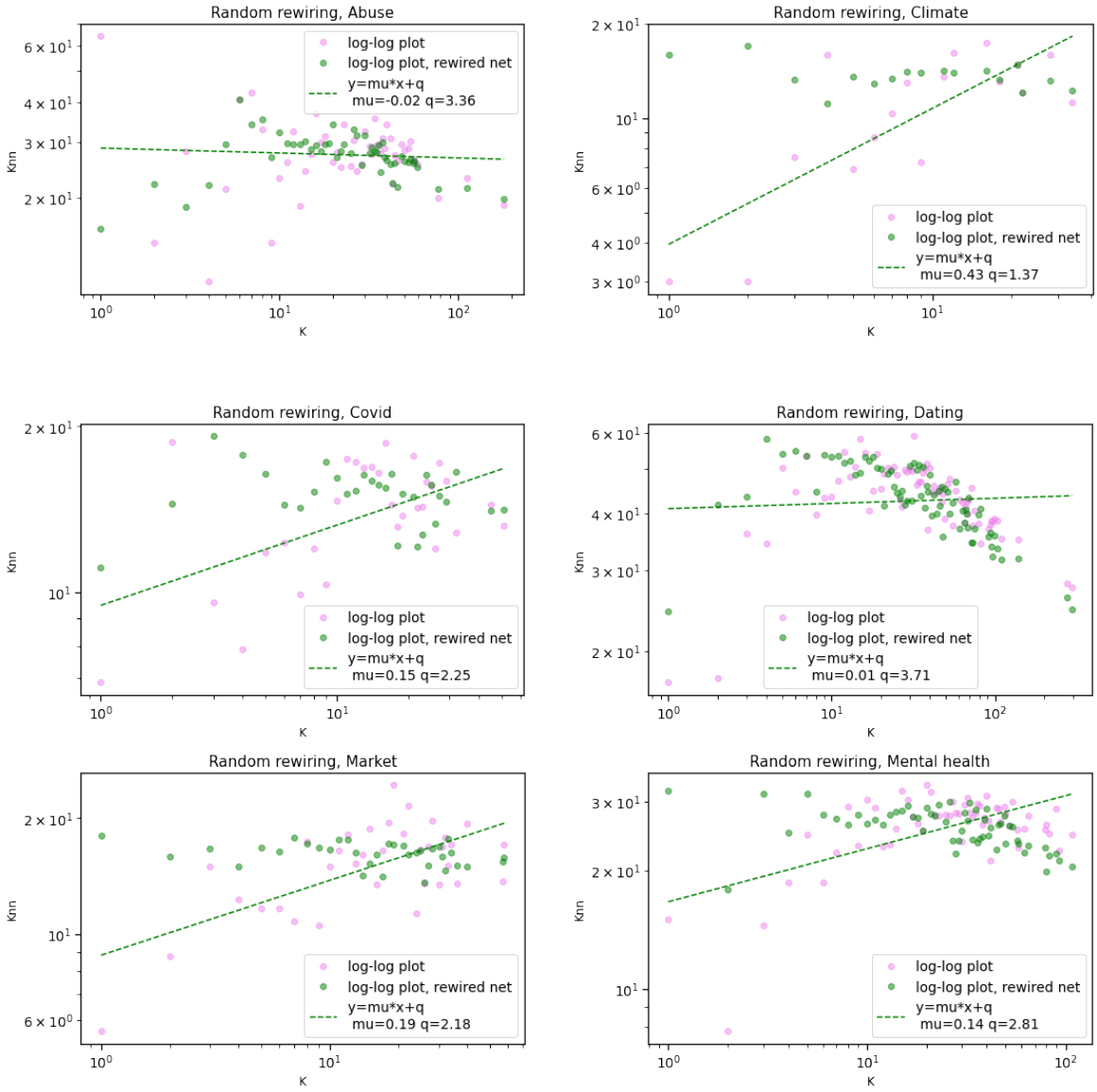


Figure 27: Average neighboring degree vs degree (log-log plot). Random rewiring version.

9 Robustness

Alessandro Marcomini

In this section we test the robustness [22] of the semantic networks to two different types of nodes removals: random failures and attacks.

9.1 Random Failures and Targeted Attacks

Studying the robustness of a network to failures allows to understand if it behaves more like a random or a scale-free network, verifying the results obtained in section 7. Indeed, while random networks react almost in the same way to random failures and attacks, scale-free networks are much robust to the former type of collapse, while being much fragile to the latter due to the presence of hubs. When we talk about robustness, we mean the ability of a network to maintain its connected structure even after the removal of some of its nodes: the higher the percentage of nodes needed to destroy the structure of a network, the higher its robustness.

In practice, what we do here to test the robustness is to use the inverse percolation method, plotting the size of the giant component as a function of the fraction of nodes we remove from the network. In case of random failures, the eliminated nodes are chosen at random, while for simulating targeted attacks we remove the nodes with the highest degree first. This is a clever way to test how robust the structure of a networks is.

We do this with the *NetworkX* library [24], building the networks using the edgelists created in section 5.3, removing different fractions of nodes (randomly chosen or hubs) and checking the size of the giant component after the removal. In this way, we are able to draw the plots in Figure 29.

To draw the curves related to random failures, we perform inverse percolation 20 times and then we take the average curve, since a single result can be not representative of the general behavior. On the contrary, this is not needed when simulating attacks, since we remove nodes based on their degree in descending order.

We can exploit the plots to confirm the scale-free nature of the semantic networks, already demonstrated in section 7 using the γ values of their degree distributions. When using robustness, this is deducible by the presence of a noticeable separation between the curves related to random failures and targeted attacks, clearly visible from the following plots. All the semantic networks are much robust to random failures, since to completely destroy their giant component we would need to remove at random a percentage of nodes above 90%; on the contrary, with targeted attacks we are able to do that removing less than 30% or even 20% of their nodes: this is an evidence of the presence of hubs. The community that seems to have the less scale-free behavior is the climate one, but the results related to that community are not very significant, due to its small size.

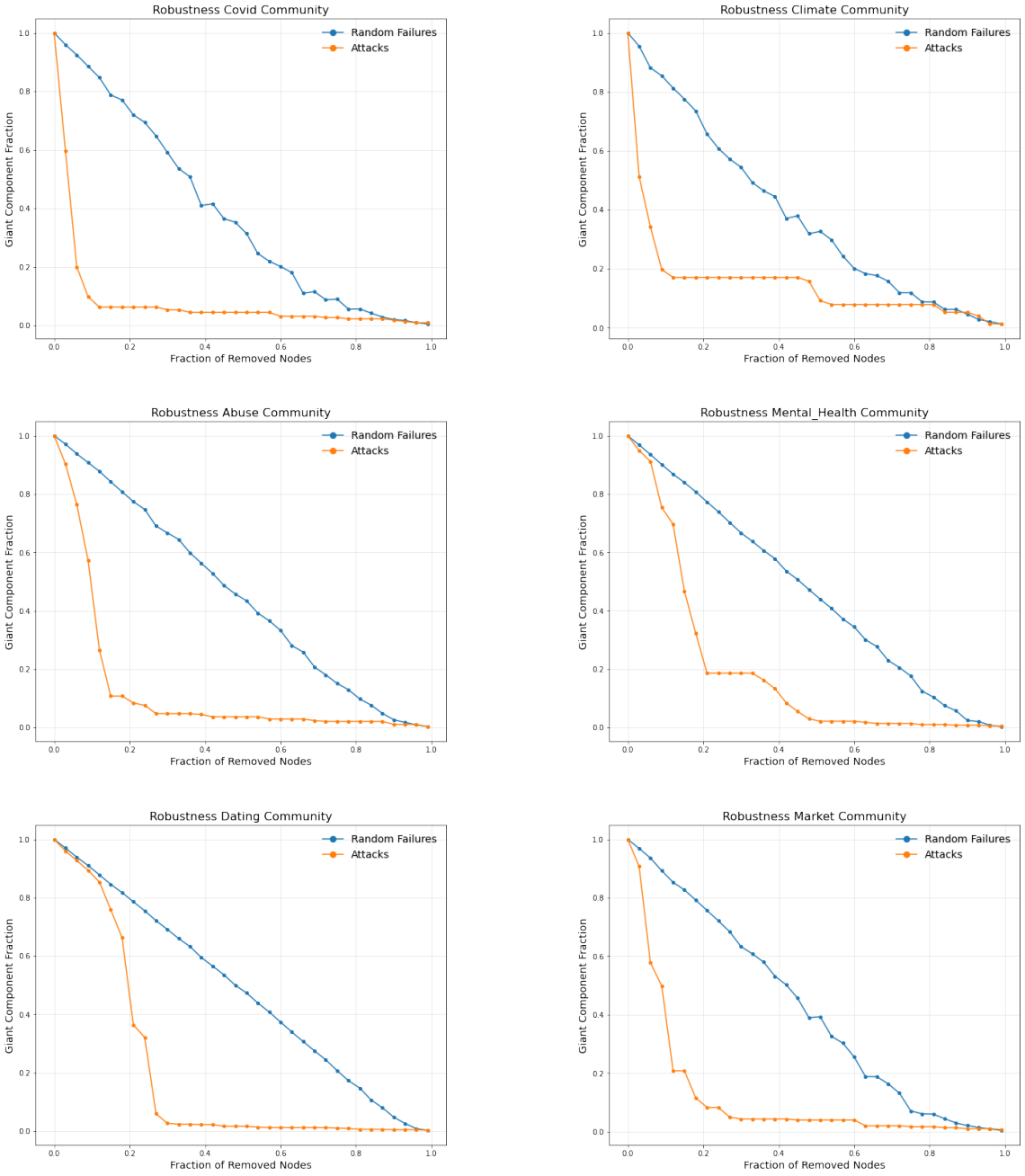


Figure 28: Robustness of the six communities we consider to random failures and targeted attacks. We can see that all of them behaves as scale-free networks.

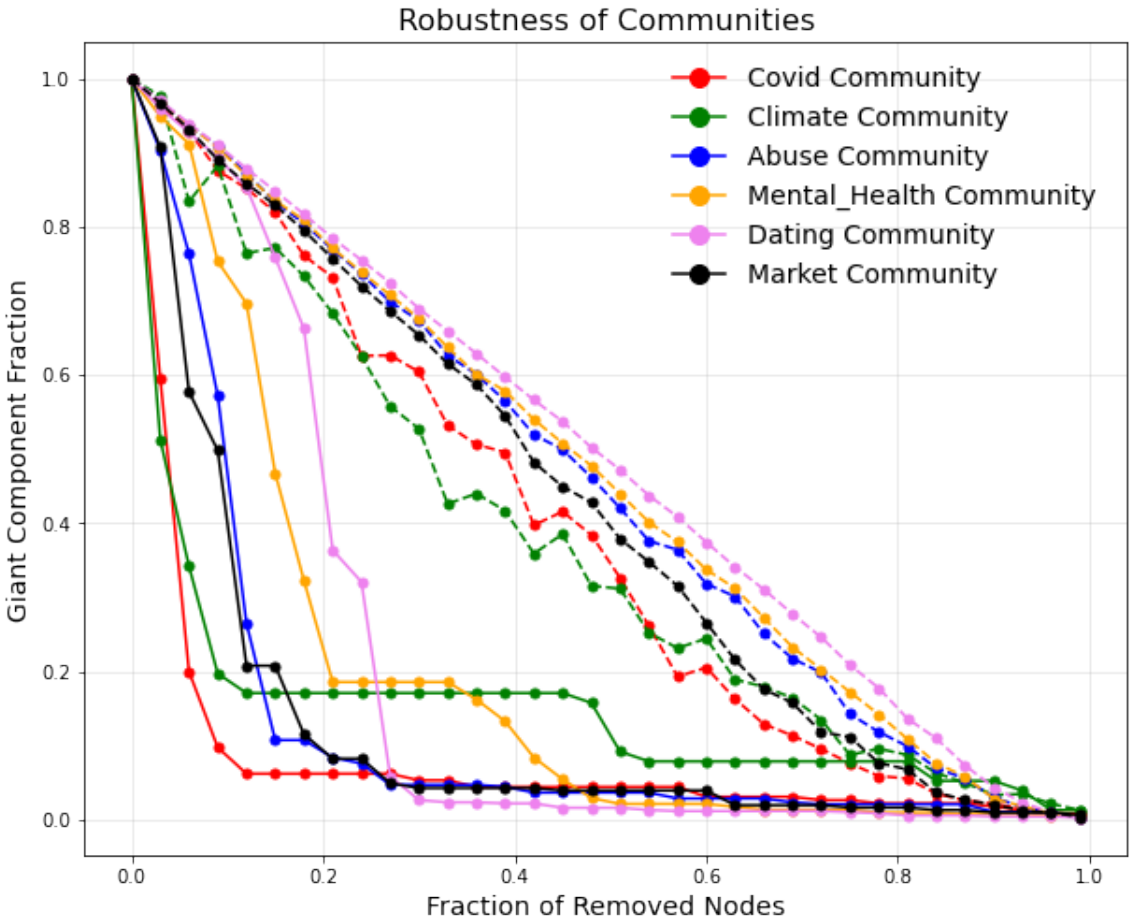


Figure 29: Robustness of different communities to random failures (dashed lines) and attacks (solid lines). The clear separation between dashed and solid lines indicates that the networks are scale-free, while the slightly different behaviors of the communities to random failures are due to differences of their degree distributions: the straighter the curve, the lower the γ value of their degree distribution.

10 Centrality analysis

Dacia Braca

Each community of words is obtained by selecting a portion of the tweets in the starting dataset and refers to a particular semantic context. Given a word community, it is possible to study the importance of each node (i.e. the word) by considering various estimates of centrality. Each measure of centrality aims to highlight a characteristic of the node and quantify its importance within the network, based on the attributes that characterize it.

10.1 Centrality measures

To carry out the centrality study, seven measures are considered for each of the six communities [27].

- **Degree centrality**: measures the number of incoming and outgoing relationships a node has

and uses it as a way to quantify the associated popularity. Given a graph G and a node u, the degree evaluates the fraction of nodes it is connected to. A normalization is then applied to each node of word community network by dividing by the maximum possible degree $N-1$, where N is the number of nodes in G.

- **Hits**: this is an acronym for *Hyperlink Induced Topic Search* and the associated algorithm allows to compute two scores of centrality for a node. In a directed graph, *authority* estimates the node value based on the incoming links while *hub* estimates the node value based on outgoing links. In the undirected graphs - that is our case - each edge can be thought as a double directed one, in other words as a pair of arrows pointing in the opposite direction. The measure of centrality thus calculates is reduced to one since hubs and authorities coincide [28].

- **Pagerank centrality**: computes a ranking of the nodes in the network counting the number and quality of incoming edges to determine a rough estimate of how important it is. This centrality measures is typically used to evaluate the importance of a website just looking at the number and quality of links to and from which it is connected. The underlying assumption is that more important nodes are likely to receive more links from the others. Then, pagerank algorithm is originally thought for directed graphs but it could be applied also on undirected ones just by converting each edge into a double directed one. At a certain time t we have a probability $p_{t,u}$ of being in the node u, which is connected to n_i nodes. At time $t+1$ it is possible to move to one of the neighboring nodes and the associated probability is equally divided among the available options, so given by $p_{t,u}/n_u$. The final probability $p_{t+1,v}$ at time $t+1$ for a new node v will therefore be given by the sum of all the probabilities associated with the previous nodes to which it is connected. This identifies a Markov chain which, by iterating the procedure on the various nodes of the network, returns an equation for the displacement probability:
$$p_{t+1} = cMp_t + (1-c)q \tag{8}$$
where M is the normalized adjacency matrix, $c \in [0,1]$ is the *dumping factor* which adds the contribution of the random component in the displacement choices and the associated q is the independent probability vector, typically uniform. In our analysis the dumping factor assumes the usual value of $c = 0.85$.

- **Betweenness centrality**: it measures the number of shortest paths passing through a node and it is a way of detecting the amount of influence a node has over the flow of information or resources in a graph. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. Betweenness centrality of a node v is mathematically given by the sum of the fraction of all-pairs shortest paths that pass through v:
$$C_B = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{9}$$
where V is the set of nodes, $\sigma(s,t)$ is the number of shortest connecting s and t while $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s and t. If $s = t$, $\sigma(s,t) = 1$, if $v \in s,t$, $\sigma(s,t|v) = 0$.

- **Closeness centrality**: it detects nodes that are able to spread information efficiently through a subgraph and to easily communicate with the other nodes. It is mathematically defined

as the reciprocal of the sum of the shortest path distances from the node of interest u to all reachable $n-1$ nodes. Since the sum of distances depends on the number of nodes in the graph, an average is computed normalizing by the sum of minimum possible distances $n-1$:

$$C_C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{10}$$

where $d(v,u)$ is the shortest-path distance between v and u, n is the number of nodes that can reach u. *Wasserman* and *Faust* propose an improved formula for graphs with more than one connected component. The result is *"a ratio of the fraction of actors in the group who are reachable, to the average distance"* from the reachable actors. Letting N denote the number of nodes in the graph, the improved version of closeness is described by the formula:

$$C_{C-WF}(u) = \frac{n-1}{N-1} \cdot \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{11}$$

For single component graphs, it degenerates in the original formula.

- **Eigenvector centrality**: also called eigen-centrality or prestige score, it is a measure of the influence of a node in a network and quantifies the associated centrality looking at those of its neighbors. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. The eigenvector centrality for node u is the u-th element of the vector x defined by the equation:

$$Ax = \lambda x \tag{12}$$

where A is the adjacency matrix of the graph G with eigenvalue λ. By virtue of the *Perron–Frobenius theorem*, there is a unique solution x, all of whose entries are positive, if λ is the largest eigenvalue of the adjacency matrix A.

- **Harmonic centrality**: given a graph G with N nodes, the harmonic centrality of a node u is defined as the sum of the reciprocal of the shortest path distances from all other nodes to u:

$$C_H(u) = \sum_{v \neq u} \frac{1}{d(v,u)} \tag{13}$$

10.2 Communities' centrality measures

For each word belonging to the six communities considered, the scores of the seven centrality measures proposed in the previous section are evaluated. Therefore, graphic visualizations of the word network are obtained, placing an emphasis on the comparison between the main measures of centrality such as degree or pagerank and the remaining. These two centrality measures both provide information about cohesion and interpretability for the network, but pagerank has also the ability to describe closeness, similarity, friendship (with a direction) and dependence between the graph nodes more than degree. Consequently, in all the qualitative studies carried out for the communities, particular attention will be given to the pagerank measure. Before proceeding with the presentation of the main results for each community, to provide a reference of the order of magnitude for the various centrality measures, a statistical summary is proposed below which lists the average scores for each of them.

Community	Degree	Hub	Pagerank	Betweenness	Closeness	Eigenvector	Harmonic
Covid	0.0478	0.0043	0.0043	0.0098	0.2787	0.0263	75.4023
Climate	0.1209	0.0116	0.0116	0.0165	0.2988	0.0502	30.8780
Abuse	0.0417	0.0025	0.0025	0.0041	0.3609	0.0239	153.7889
Mental health	0.0284	0.0019	0.0019	0.0037	0.3372	0.0213	190.1219
Dating	0.0283	0.0015	0.0015	0.0024	0.3892	0.0189	281.1596
Market	0.0319	0.0031	0.0031	0.0056	0.2610	0.0271	93.4425

Table 4: Mean values for centrality measures with respect to the six communities

Once the centrality scores are obtained for each node in the network, it is possible to specifically analyze the behavior of each word in terms of importance in the community's tweets. First, a comparison is made between the various performance of the centrality measures and, secondly, a visualization for each community network is obtained by exploiting the information derived from the centrality's study.

10.2.1 Covid community

To directly compare the results of the centrality measures on the word community, a graphical representation is performed between pagerank and the remaining estimates through a scatter plot. The pagerank scores are shown on the x-axis while the other centrality estimates are considered on the y-axis, thus obtaining six comparative representations. To provide a clearer and more legible display, we have accompanied each of the analyzed centrality measures with a color scale. In this way, in addition to being able to read the numerical value directly on the ordinates, it is possible to qualitatively evaluate the centrality by observing the intensity of the point's color in the plane. Looking at the first of the following graphs, that one on the left, it is possible to notice how the degree and the pagerank centrality present an almost linear relationship: increasing values in the degree correspond to increasing values for the pagerank and vice versa. Also in the comparison between pagerank and hub centrality we can appreciate a hint of direct proportionality, even if compared to the previous case the growth of the hub score is slower.

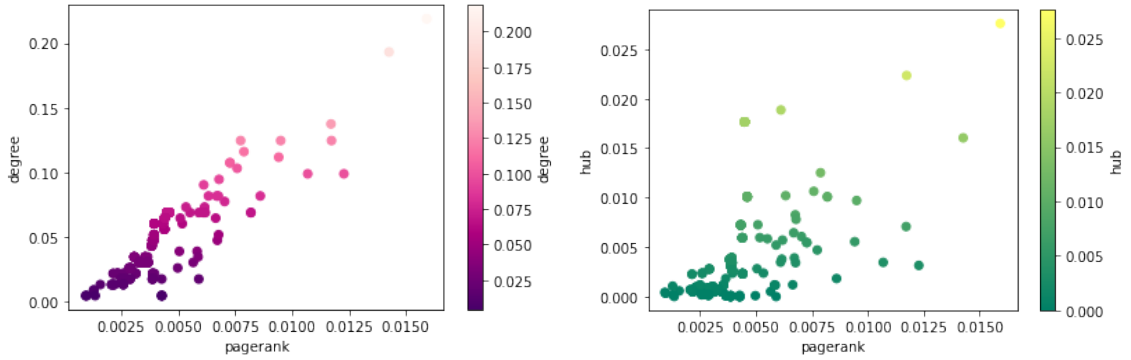


Figure 30: Pagerank comparison for covid community

For this community of words we find that a smaller number of nodes act as a bridge for other links even if characterized by a medium or high pagerank value. In particular only a little group of nodes, with a very high number of connections, present also a high betweenness. This result can be read in the light of the practical meaning of the network whose nodes actually correspond to words. A possible interpretation is that there are nodes that appear frequently with different combinations

of words but only a small part of these is recurrent in more tweets or rather in sentences with a wider volume of words. Furthermore we can suggest that for this network most of the shorter paths pass through nodes with a high pagerank. From the comparison between closeness and pagerank we can instead find that the network is mostly cohesive, with the exception of some small set of words that isolate themselves from the rest of the community. One might thus think that in the extracted tweets there are words that tend not to approach the others. We remember that closeness can be seen as a measure of center of gravity, therefore it is expected that the maximum communication of information is concentrated in a restricted area of the network. This type of result is best appreciated with another type of visualization.

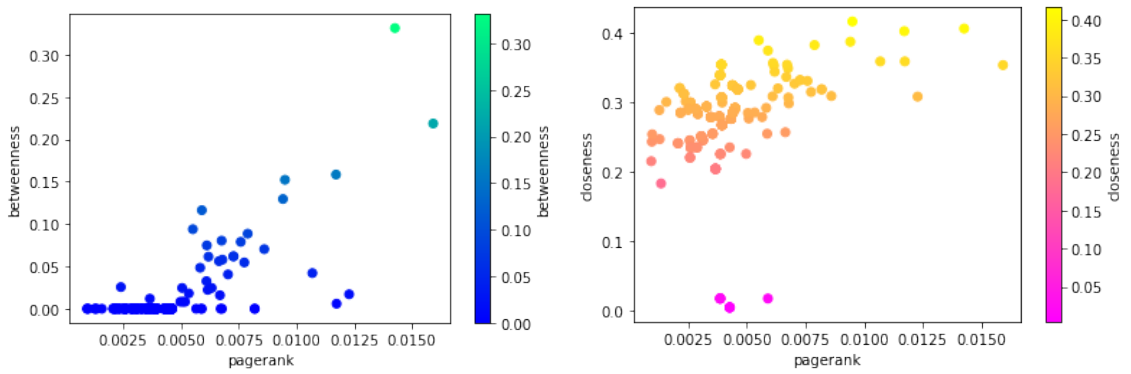


Figure 31: Pagerank comparison for covid community

Only a very small number of nodes has a high value of the eigenvector centrality, a fairly typical behavior given the characteristics of the network and the definition of this measure of centrality. In the case of harmonic centrality then we find a relationship similar to the one that existed with the classic measure of closeness.

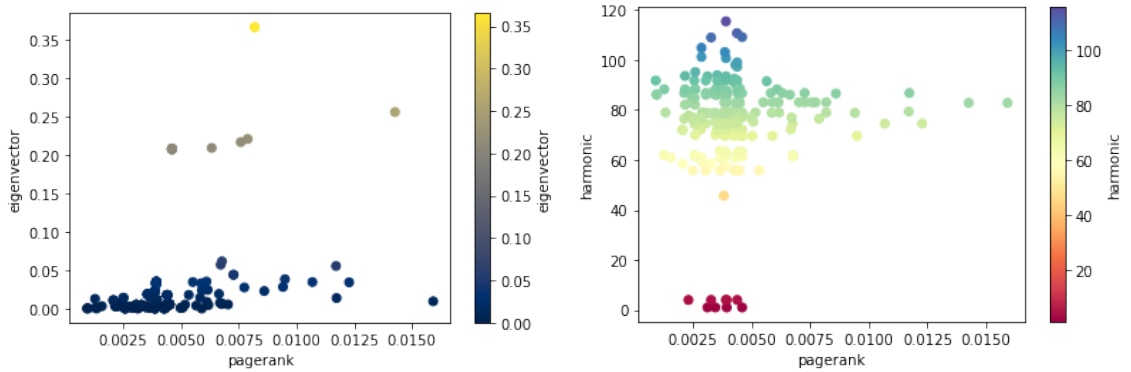


Figure 32: Pagerank comparison for covid community

At this point it is possible to propose a visualization of the community taking into consideration the information obtained with the study of centrality, using the various scores to distinguish and identify each node. In particular, the size and the color of the label indicate respectively the pagerank values and those associated to one of other six centrality measures described above. For this study, the color scale is obtained by considering the extremes of the interval for the selected centrality measure. Therefore, each representation is accompanied, on the immediate height, by a color scale which has to be read from left to right: the color on the far left indicates the minimum

value of the centrality score considered, while the right one indicates the maximum score for the specific community. The thickness of the gray lines connecting the words indicates the relative weight: if the line that connects two nodes is thin (in some cases it is almost not visible) it means that the weight of the link is low and therefore that the pair of words does not often appear in the same tweet. On the contrary, if the connection is thick it means that the two words tend to be used together in a speech.

Two centrality measures are compared on the left: the pagerank defines the size of the label while the color quantifies the degree. When a word is large, its pagerank value is high and therefore that word is linked to many other words in the various tweets. To read the color meaning you have to rely on a color scale: each degree value is converted to a color within a graduated palette where dark fuchsia is the minimum score and light pink is the maximum one. From the representation we can see what has already been discovered with the scatter plots: basically a word with a high pagerank also has a high degree (nodes with a larger label have a lighter pink tone).
On the right we can highlight the presence of hubs. The hubs in this network are few, observing the graph you immediately notice the presence of a super hub given by the words "like", in yellow. Knowing that the covid network is scale free, we expected the presence of large hubs. There are also other words with a more discrete score and which have a medium pagerank value.

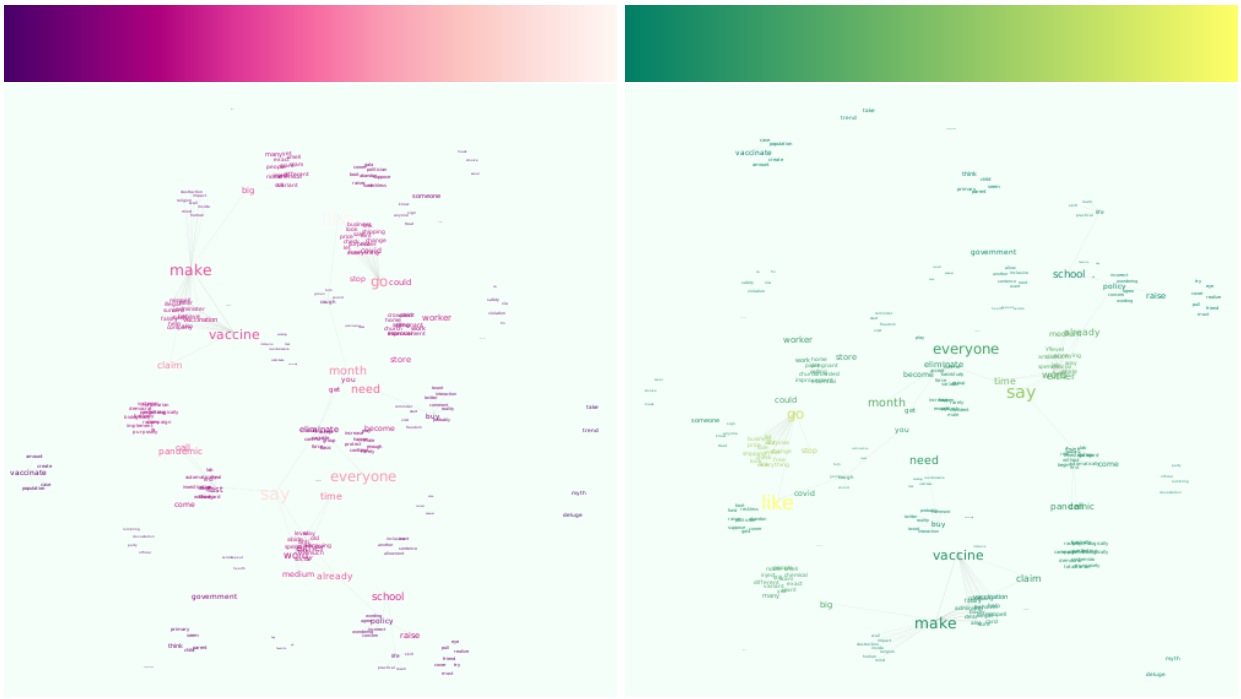


Figure 33: Pagerank (label size) with respect to degree and hub centrality (color scales) for covid community

On the left in the figure 34, we can appreciate the network representation considering the betweenness as centrality measure for color scale. As anticipated in the relative scatter plot, in this community there are very few nodes that act as a bridge for other words: "say", "like" and "everyone" are some of them. With the visualization on the right, the meaning of closeness becomes clearer and color is of great help here. The first thing that catches the eye is the presence of small groups of words identified by the fuchsia, indicating a low value for this measure of centrality. The rest of the network, on the other hand, has a color gradient from the center outwards which suggests

that there are some words that better spread information.

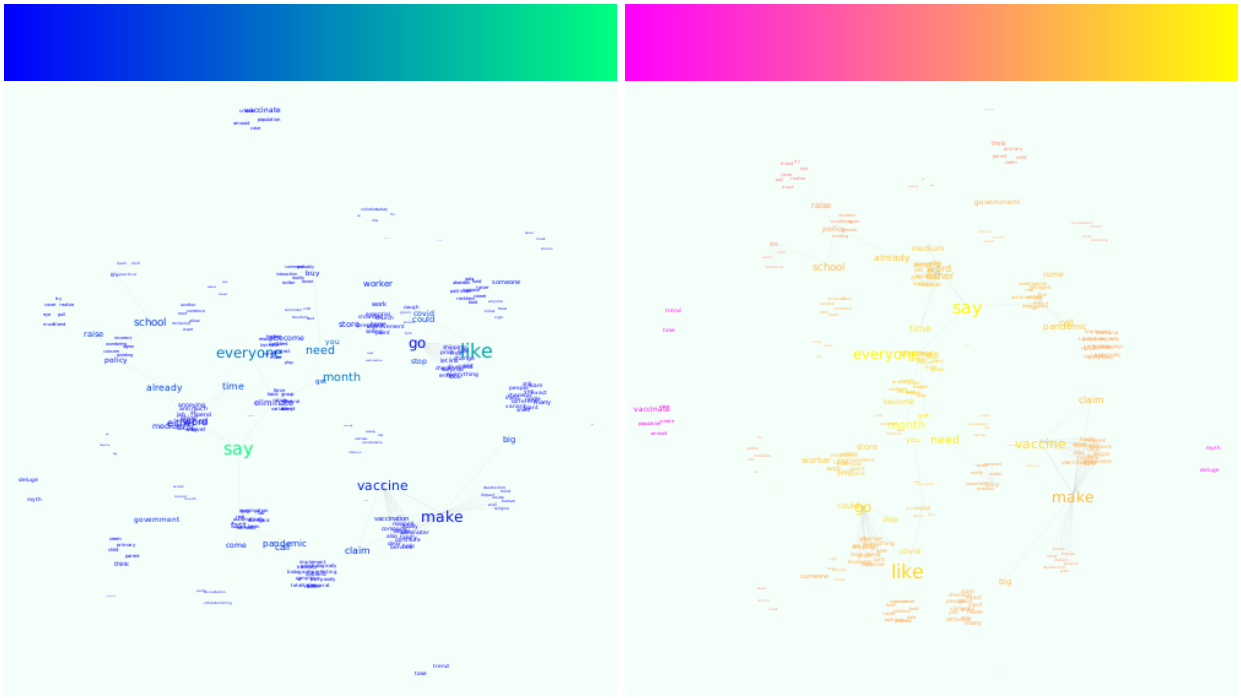


Figure 34: Pagerank (label size) with respect to betweenness and closeness centrality (color scales) for covid community

On the left in the image 35 the eigenvector centrality is compared to pagerank through a network visualization: as expected, only a small group of words has a high value of the eigenvector centrality, finding its center in the words “other” and “word”. For the last measure of centrality, harmonic one, we do not find a clear direct proportionality with the pagerank: the words with a higher value of the harmonic score have neither a high degree nor a high pagerank and are more located within the network. However, we can note that for small values of the pagerank there is a high spectrum of variability for the harmonic centrality, the values then stabilize when the importance of the node in terms of the first consolidates.

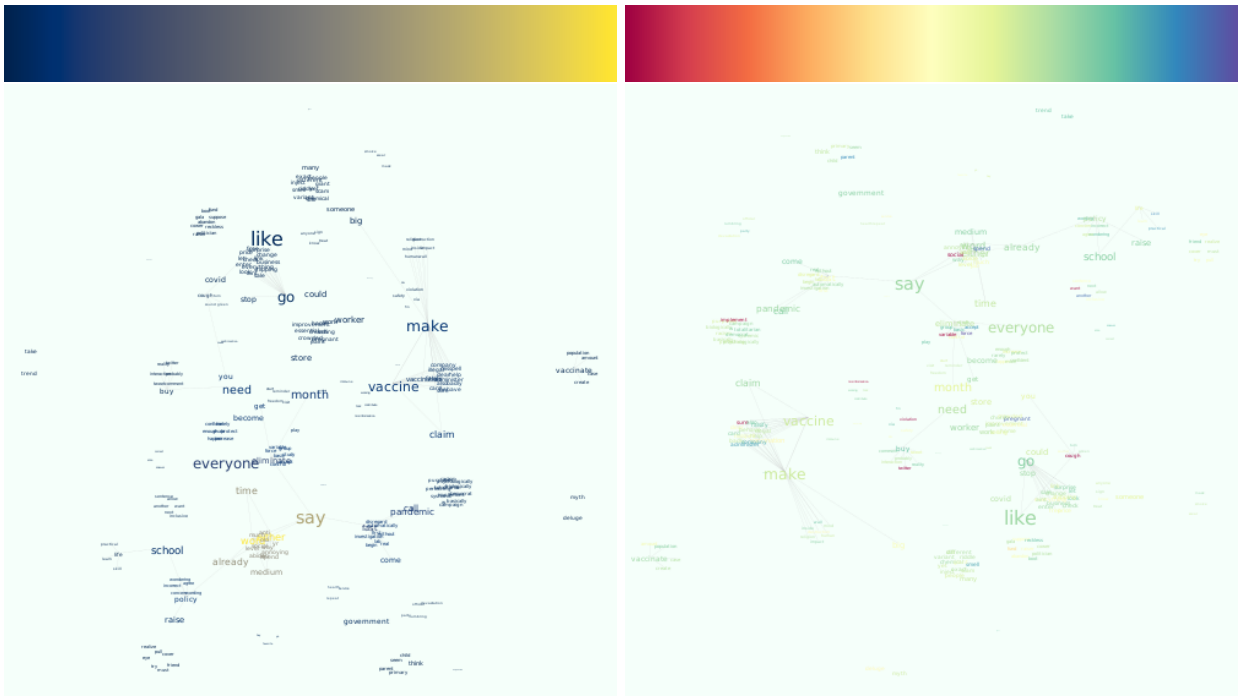
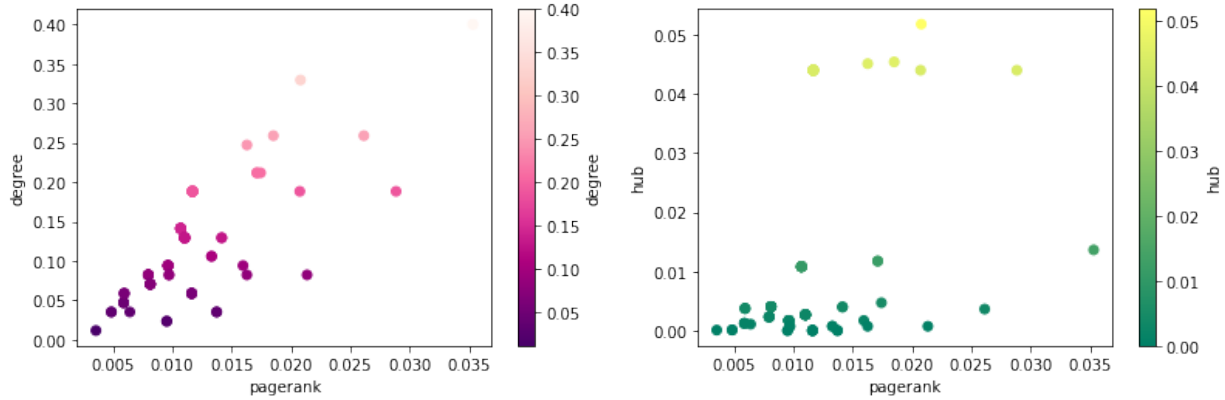


Figure 35: Pagerank (label size) with respect to eigenvector and harmonic centrality (color scales) for covid community

10.2.2 Climate

The procedure followed for the first community is the same considered for all the others, so beyond the natural differences related to the structure of the single word network there are no substantial changes. In particular, this new community associated with climate topics has a much lower number of nodes and connections than the previous one. Observing the scatter plots that compare pagerank to other centrality measures, we note that the behavior does not change with respect to what was found in the previous section. The pagerank and the degree maintain a linear proportionality even if for increasing values we notice a greater distribution of the points in the 2-dimensional plan. The climate community is also scale free, so the presence of large hubs is waited for. Expectations are satisfied, in fact by observing the graph on the immediate right (figure 36) you can see the presence of points with a high value of the ordinates that clearly differ from that of the other words.
The usual observations apply to all other centrality measures.



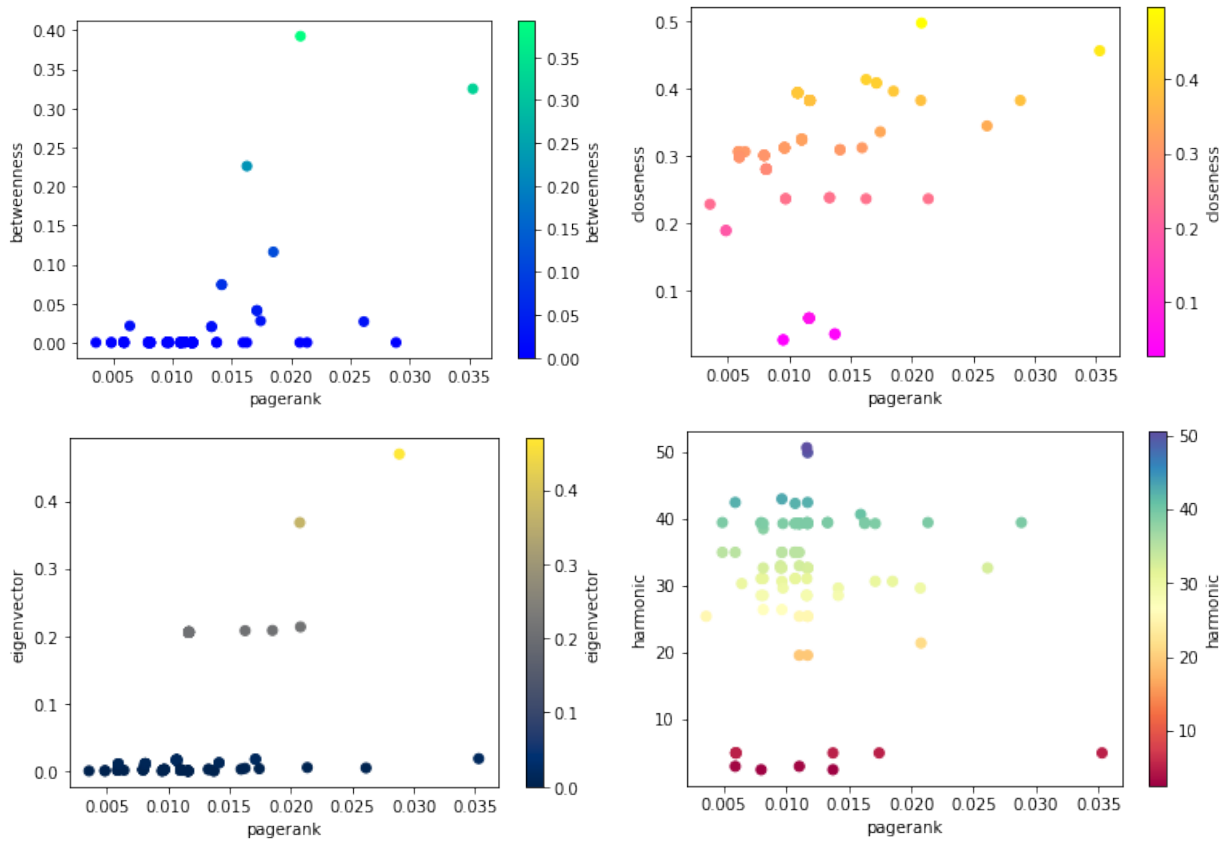


Figure 36: Pagerank comparison for climate community

Below we propose only two of the graphical representations of the network referring to betweenness and closeness. As usual, the size of the labels is given by the pagerank value for the single word and its color defines the score for the other measure of centrality, following the spectrum of the color scale. It is interesting to notice how words like “climate”, “change”, “water” present a high centrality score but only the first one acts as a bridge for the spread of information. To this are also added “product” and “plastic”, with a high level of betweenness. Also in this network there is an intensity gradient linked to the closeness scores for which there are small groups of isolated words that detach themselves from the main structure.

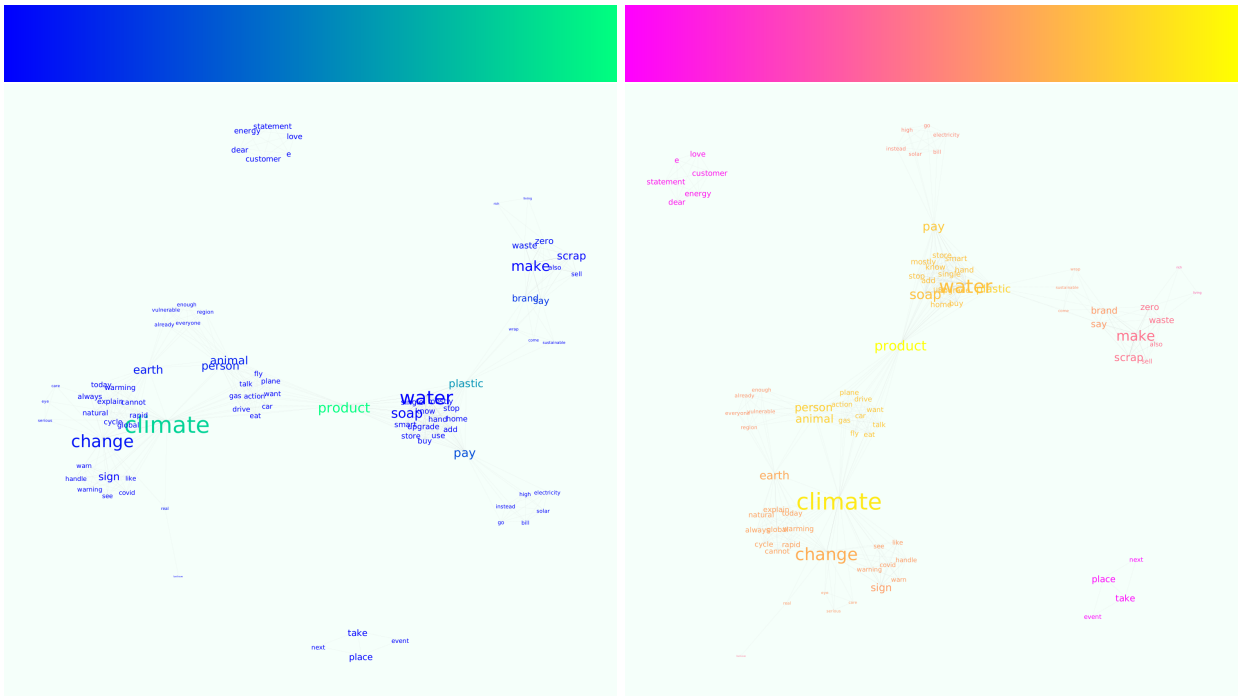
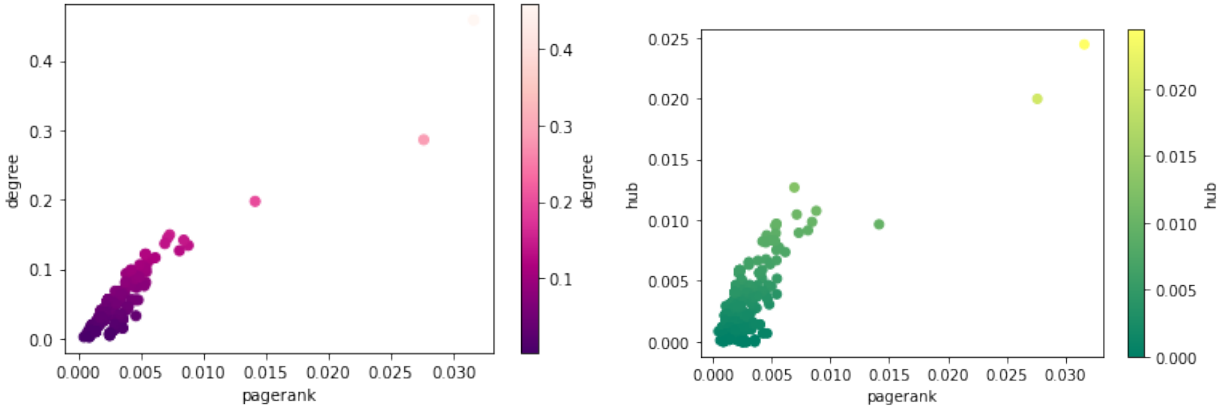


Figure 37: Pagerank (label size) with respect to betweenness and closeness centrality (color scales) for climate community

10.2.3 Abuse

The third community has a decidedly more consistent number of nodes and connections than the previous ones but what emerges immediately by comparing the values of the centrality measures, is that there is a couple of words - or a little more - with a very high importance. The presence of these super nodes, given by “law” and “gun”, can explain the behavior of the centrality scores for the remaining words of the network. Since these few nodes are extremely central, the scores associated with the others can only be significantly reduced to smaller values of importance, so much so that they tend to concentrate in a limited range centered at the beginning of the axes, as you can appreciate in the following scatter plots. We therefore note that for the various centrality measures, with the exception of closeness ones, an almost linear relationship with the pagerank is outlined.



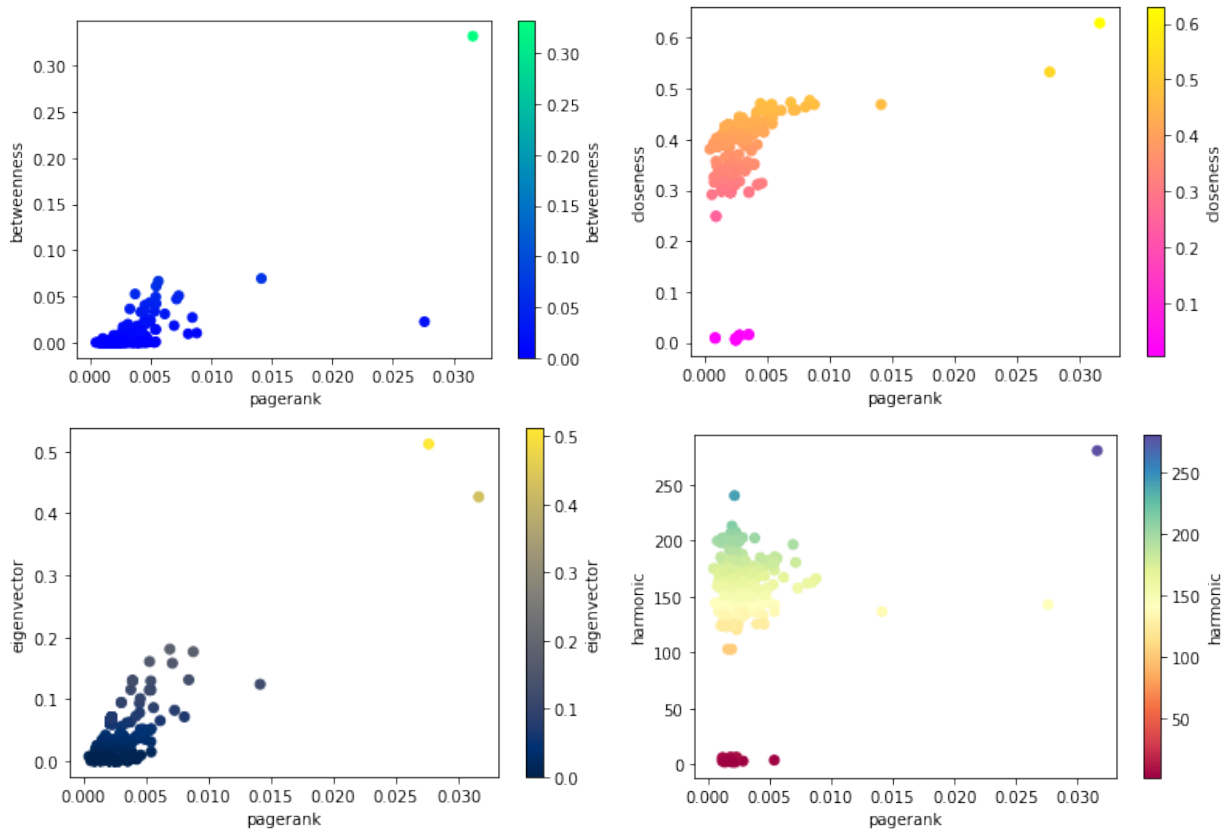


Figure 38: Pagerank comparison for abuse community

Before proceeding with the analysis of the graphical representation of the climate community shown right below, it is necessary to make a note on the chosen format. The high number of nodes does not allow a representation like those of previous networks, in which the presence of a edge between a pair of words was taken into consideration by trying to bring them close together in the available space. In the case of large networks, a representation of that type, although possible, is inadequate. So in subsequent views, a little information is sacrificed for the purpose of readability and comprehensibility. The words are therefore arranged randomly on the plane, always maintaining the previous parameters: thickness and color of the connection, size and color of the label keep the same meaning.

In the first graph of the figure 39 the pagerank is compared to the degree values, in the second one any hubs in the network are shown. We find that "gun" and "law" are the most central words (we give a little hint for visualization on the left, the word law is in light pink just above gun). This causes all the others to take on a secondary level. A possible reading of this result is that in the tweets associated with hashtags with the theme *abuse*, there are a few keywords that are juxtaposed numerous times to numerous words. In particular, the word "law" is central in the net, which indicates its frequent recurrence in tweets concerning abuse issues. It not only has a high centrality value but represents the point with the highest value of betweenness, closeness and harmonic centrality (see the scatter plots in the figure 38). Wanting to give a heuristic interpretation to these results, one could say that society is calling out to the legal system to deal with situations of abuse. We could also ask ourselves a question: could it be that people do not feel sufficiently protected and warded by the law? A rough answer to this question could be provided by analyzing the sentiment associated with this word. The centrality of the node "gun" may be due to the

particular choice of hashtags as a filter for the community. In this topic we also consider those tweets concerning the use of weapons with particular attention to guns.

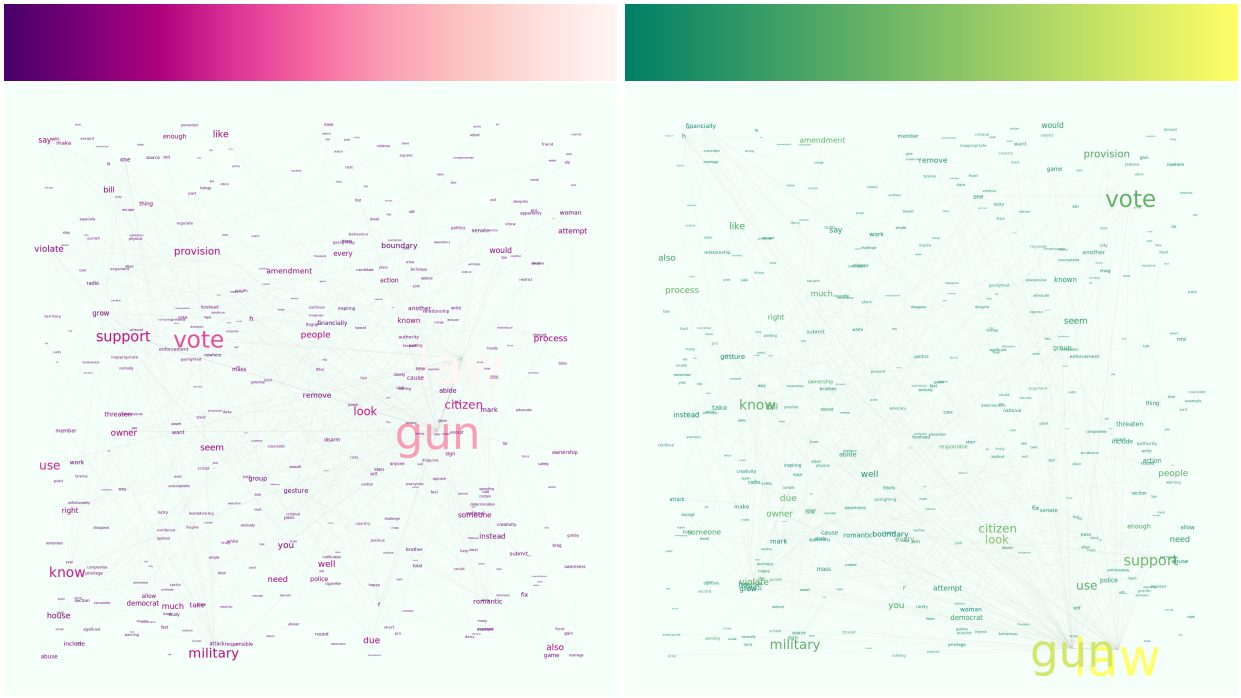
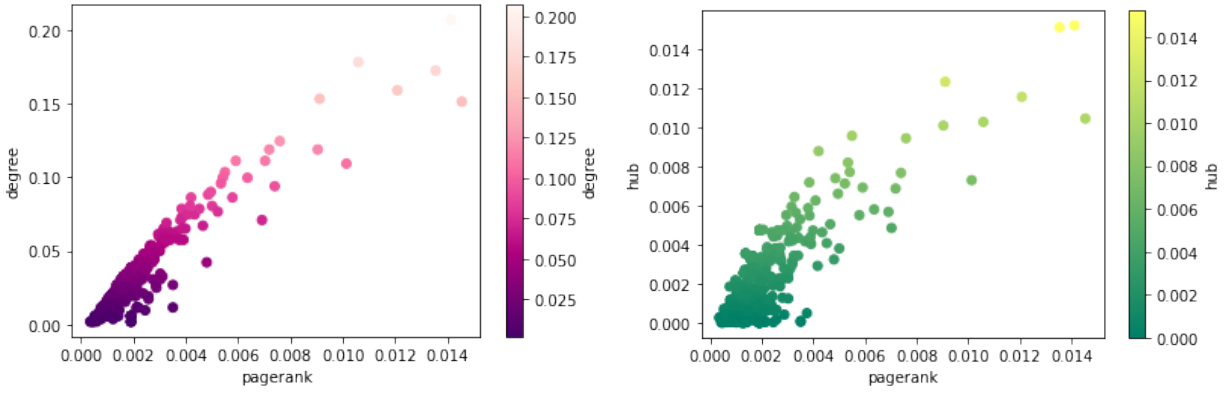


Figure 39: Pagerank (label size) with respect to degree and hub centrality (color scale) for climate community

10.2.4 Mental health

For tweets with a *mental health* theme, the words have a more distributed behavior: they assume values in the ranges identified by minimum and maximum scores with a more gradual variability. Observing the scatter plots right below, referred to the pagerank-degree and pagerank-hub comparison, it can be seen how the points tend to be distributed more homogeneously in the plan, following a more evident linear trend.



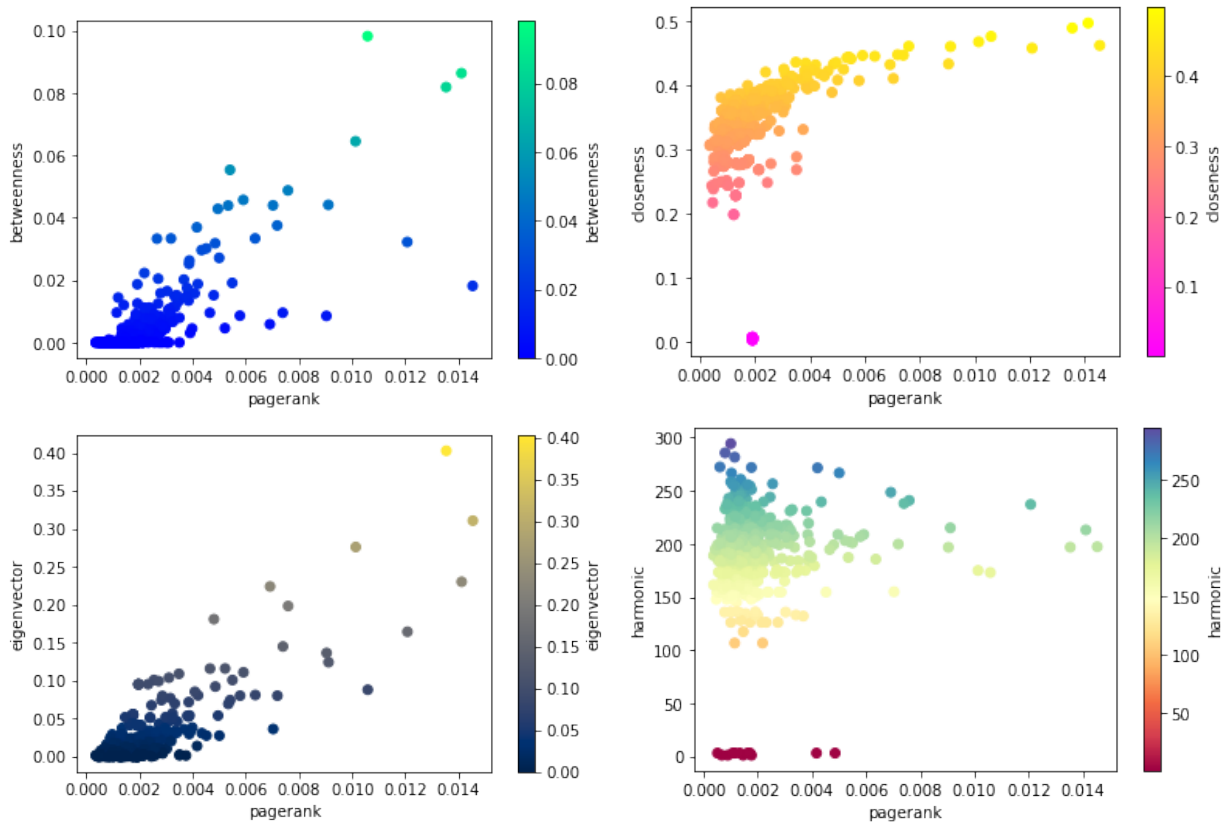


Figure 40: Pagerank comparison for mental health community

In general we can say that, for the sake of brevity, the qualitative study of these comparison graphs between the centralities produce results similar to those of the previous community but there is an important substantial difference. Even if there are more central nodes than others, this disparity is not as pronounced. It can be hypothesized that although there are main central words, the focus is more on a recurring register or lexicon in the tweets than on the single term.

More significant is the graphical representation of figure 43, once again obtained by randomly arranging the nodes in space. Unlike the community with the abuse topic in which a few words appeared recurrently, in the tweets with hashtags related to the argument of mental health it seems that there is a certain specific language, more than keywords we can talk about a small basic vocabulary.

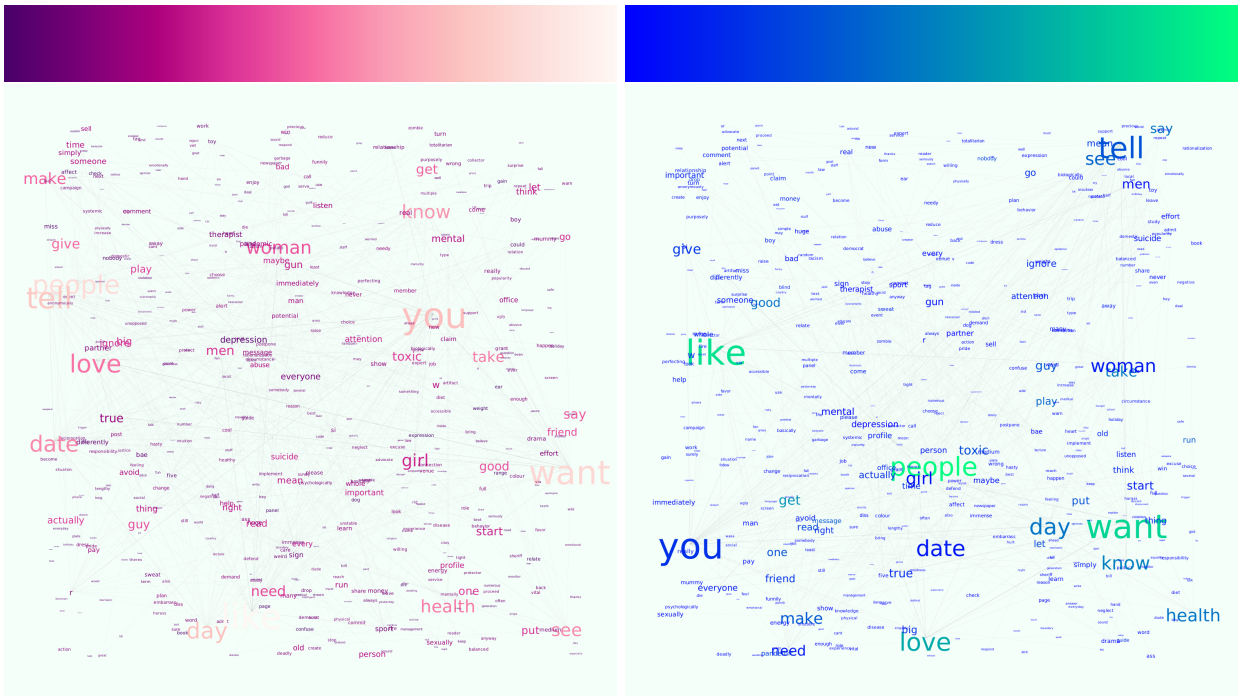
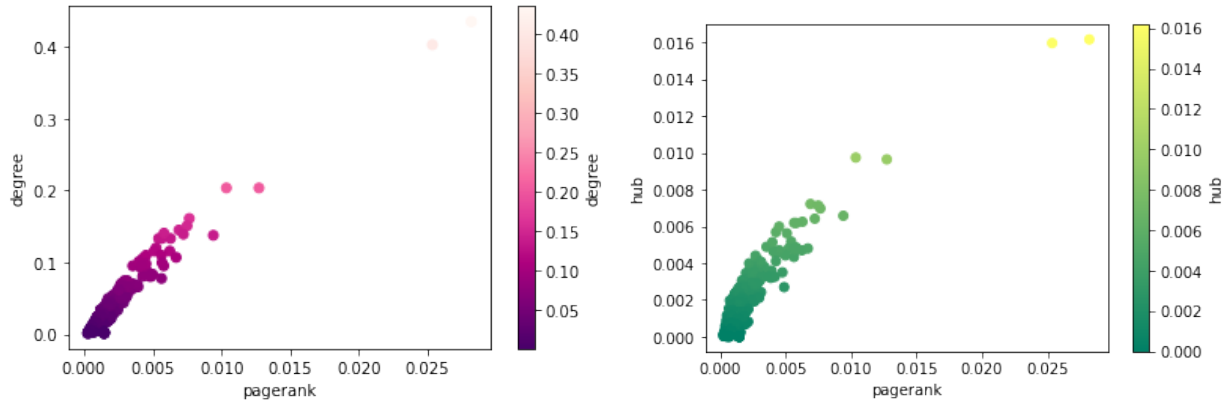


Figure 41: Pagerank (label size) with respect to degree and betweenness centrality (color scales) for mental health community

10.2.5 Dating

For the community with the "dating" theme, if we stay on a range of low values, we find that between the pagerank and the other centrality measures a clear linear relationship appears, with the exception of proximity measures such as closeness and harmonic centrality. Once again we note the presence of a couple of super nodes, given by the words "date" and "woman", which not only have high centrality values but also take on a main and exclusive role in the diffusion of information (see betweenness and closeness measures in figure 42). With the exception of these two words we can say that in general centrality is democratically distributed within the network.



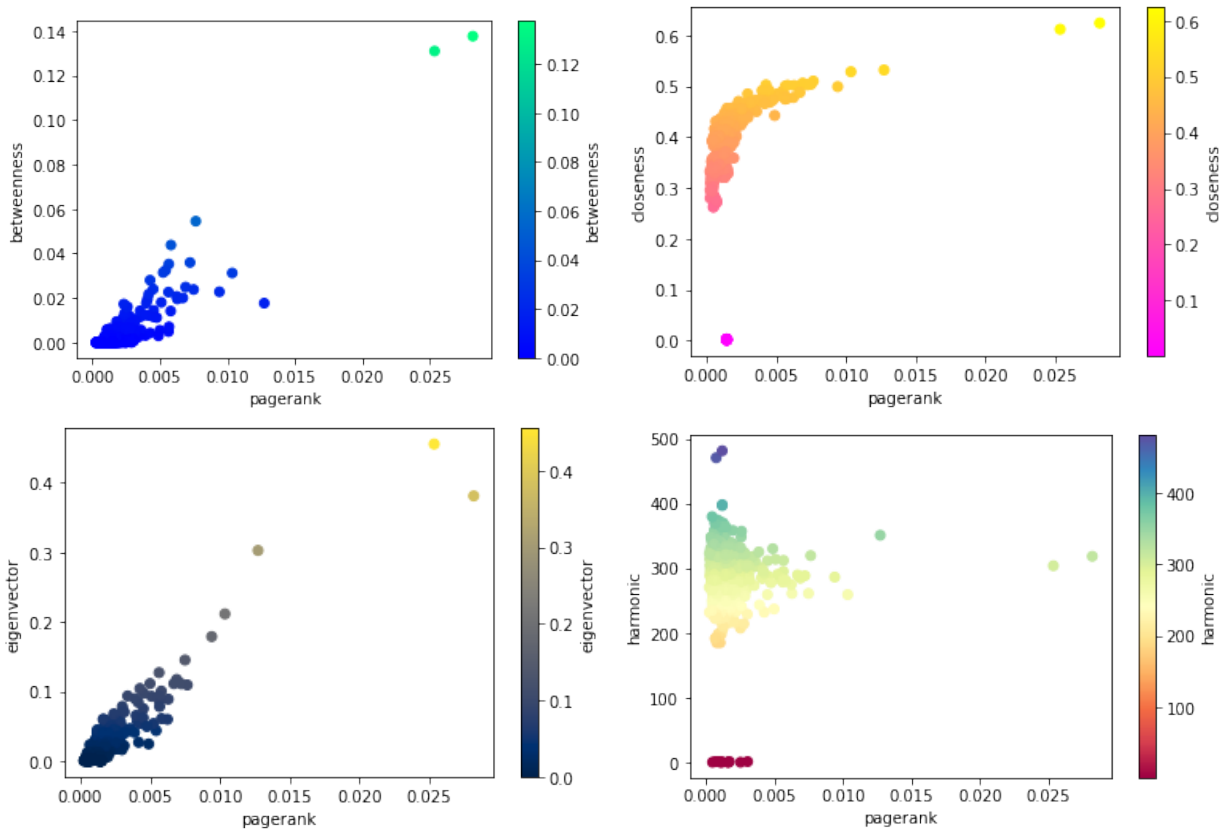


Figure 42: Pagerank comparison for dating community

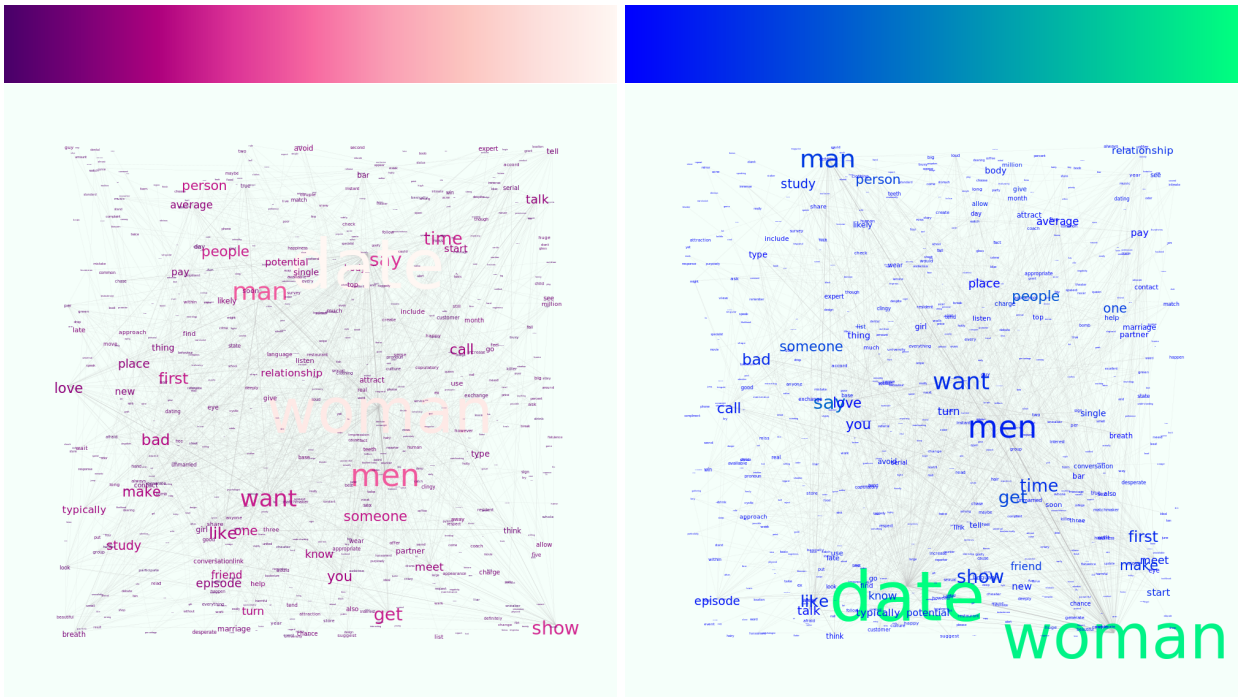


Figure 43: Pagerank (label size) with respect to betweenness and closeness centrality (color scales) for dating community

10.2.6 Market

To conclude without burdening the discussion, we renew the comments already made for the other communities. A linear trend between pagerank and degree characterizes the *market* network. There are several nodes with high centrality values but only a couple have high betweenness and closeness scores. The other words of the community do not differ too much in terms of centrality and influence. Below we propose the usual graphic interpretations, to be read with the criteria already set out in the previous sections.

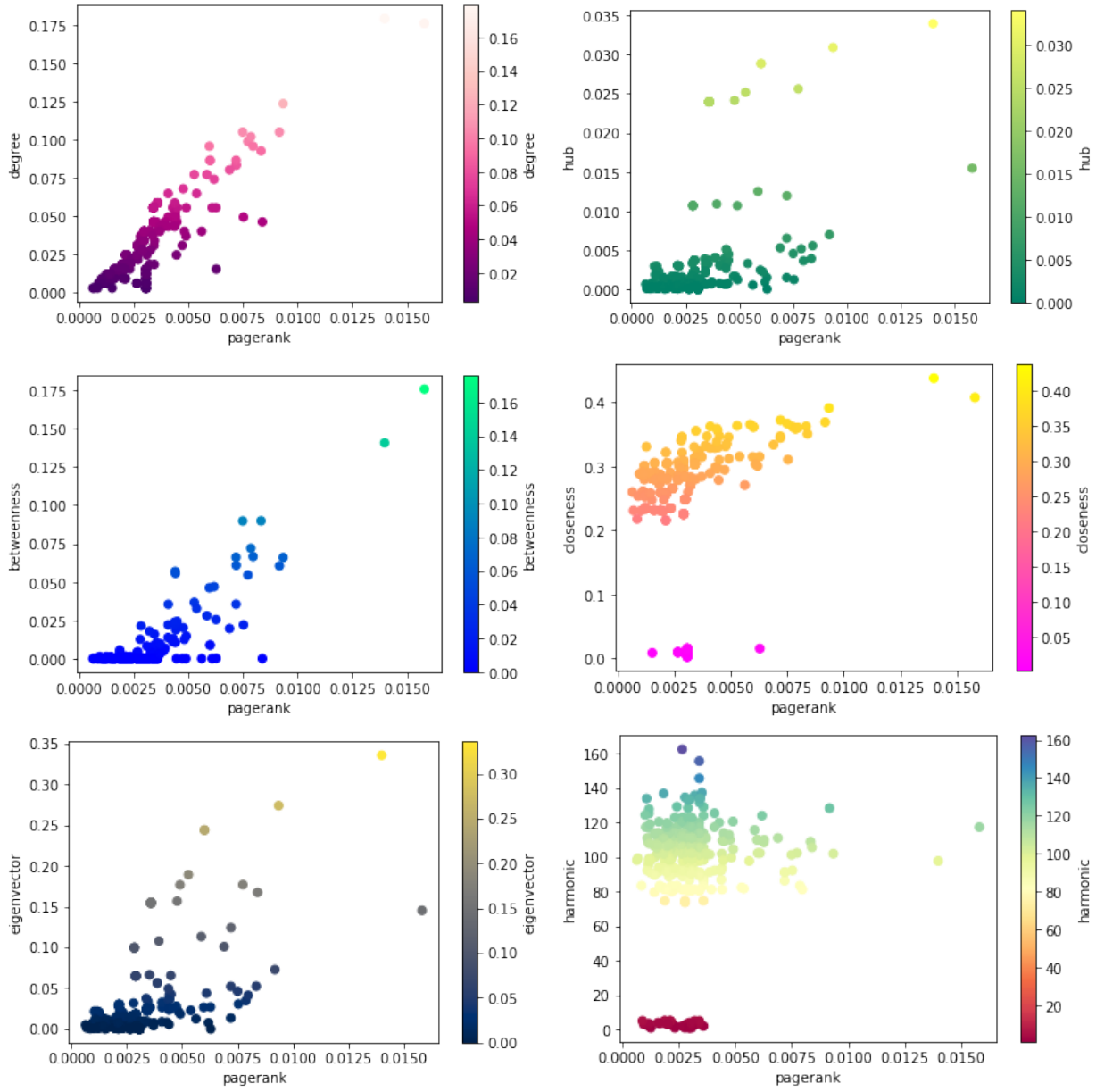


Figure 44: Pagerank comparison for market community

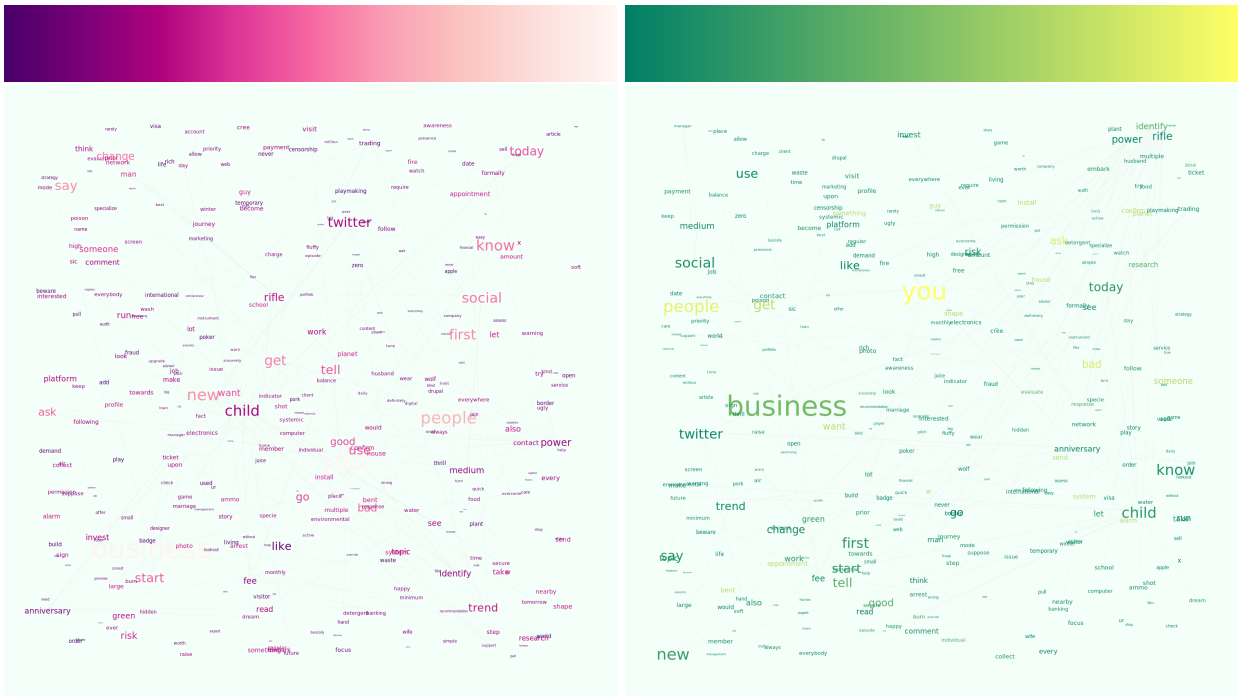


Figure 45: Pagerank (label size) with respect to degree and hub centrality (color scales) for market community

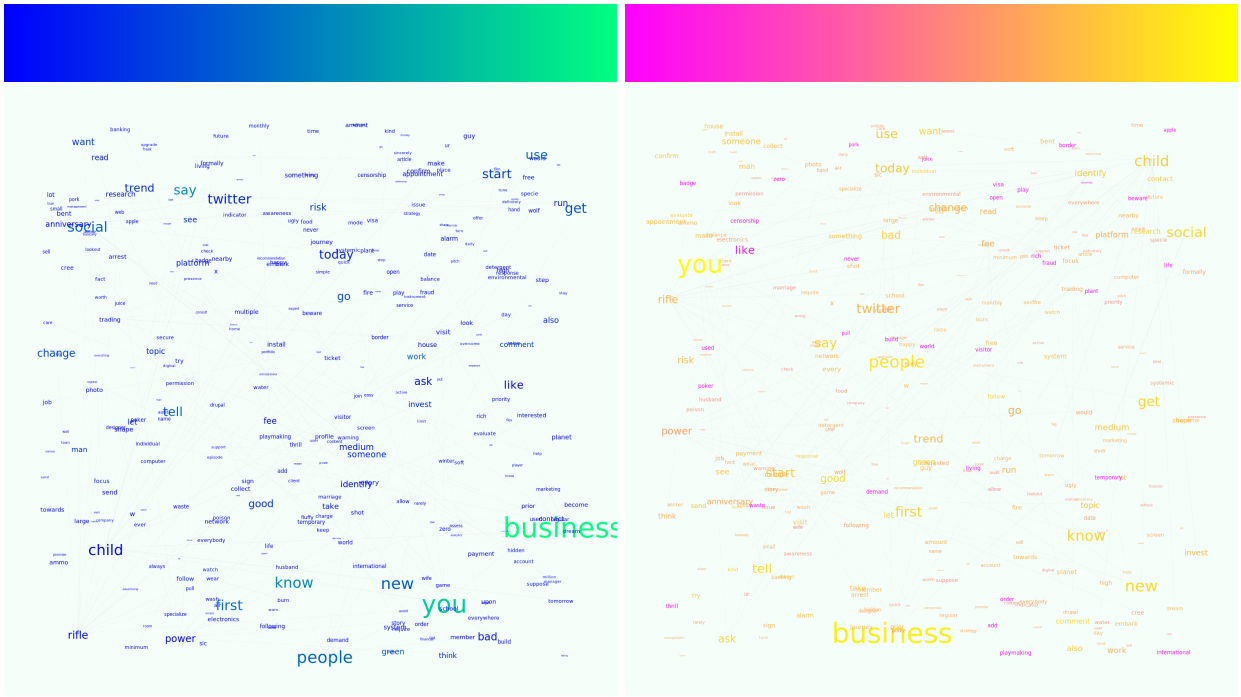


Figure 46: Pagerank (label size) with respect to betweenness and closeness centrality (color scales) for market community

10.3 Clustering coefficient

Camilla Quaglia

In this section both the local and global clustering coefficients are examined. We do this through the use of the *Networkx* library in *Python* [25] [26]. The networks considered are the words networks for the six communities above.

The local clustering coefficient is a number in [0,1] associated to each node in a graph that quantifies how close its neighbors are to being a clique. Therefore this measure can determine whether a graph is a small-world network, a property linked to the scale-free regime.

Here since we work with weighted graph, the local clustering coefficient c_k for the node k is defined as follows:

$$c_k = \frac{1}{deg(k)(deg(k)-1)} \sum_{v,u} (\hat{w}_{k,v}\hat{w}_{k,u}\hat{w}_{v,u})^{1/3} \tag{14}$$

where the edge weights $\hat{w}_{k,v}$ are normalized by the maximum weight in the network $\hat{w}_{k,v}/max(w)$. If $deg(k) < 2$, the value of c_k is assigned to zero. To specify, the summation in the formula is over the pairs of neighbors which form a triadic closure with node k.

Local clustering coefficient equal to zero means that the neighborhood of the inspected node is not connected, conversely if it is equal to 1 indicates strongly connected neighborhood. In the graphs in figures 47 and 48 the six communities are represented and the colors of the nodes indicate the value of the clustering coefficient, according to the legend. Note that only the words corresponding to nodes with clustering coefficient equal to 1 are displayed.

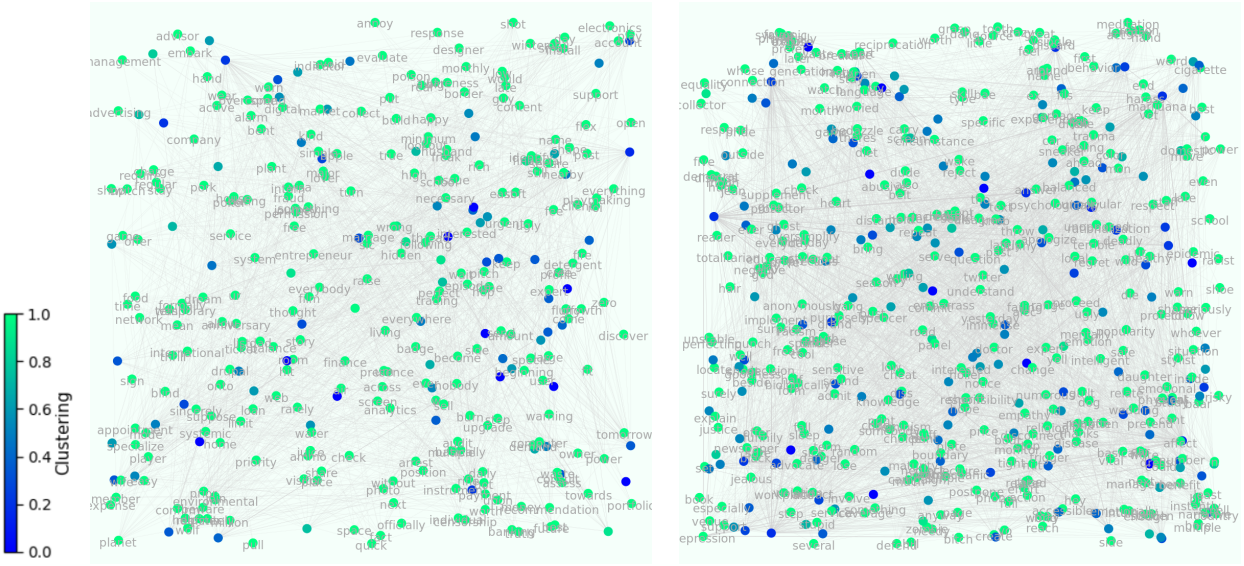
(a) Network of words, *Abuse* community

(b) Network of words, *Climate* community

(c) Network of words, *Covid* community

(d) Network of words, *Dating* community

Figure 47: Local clustering coefficients



(a) Network of words, *Market* community (b) Network of words, *Mental health* community

Figure 48: Local clustering coefficients in each community. Nore that the layout of the graph is "random", that is to say the real position of the nodes in the network are not preserved, to allow a more clear visualization of words.

In the following the average clustering coefficient for each community is computed, following the definition

$$C = \frac{1}{n} \sum_{k \in G} c_k \tag{15}$$

where c_k is the local clustering coefficient of node k and n is the number of nodes in the graph G. As one could guess looking at the graph, the results are high. The interpretation of high average clustering coefficients lies in the fact that our communities can be considered *small-world* networks, and this is connected with the fact they are all scale-free.

Community	Average clustering coefficient
Covid	0.89
Climate	0.90
Abuse	0.90
Mental health	0.83
Dating	0.81
Market	0.87

Table 5: Average clustering coefficient

A further insight is provided by the plot in figure 49. On the x axis there are nodes degrees, divided per "degree class" while on the y axis there is the average local clustering coefficient associated to that degree. A similar trend is shared among all the communities: hubs i.e. words with high degree (many connections) have a lower clustering coefficient, that is to be expected. In fact when a word appears in many tweets (high degree) its neighbors (the words in the same tweets) have, probably, less edges among them. One can expect the aforementioned trend in a social network analysis: the connections of an influencer are less likely to connect each other.

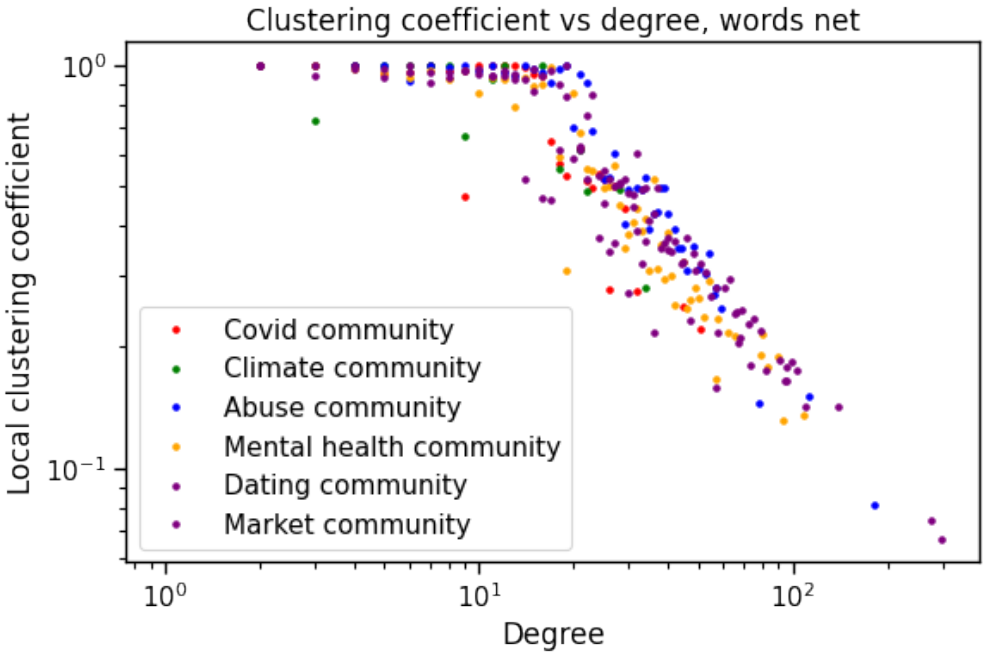


Figure 49: Average local clustering coefficient vs degree. The plot is in log-log scale.

11 Sentiment Analysis

Sentiment analysis aims at gaining some insights about the general feelings inside a text, in our case inside the tweets referring to a specific topic.

11.1 Vader Sentiment

Dacia Braca

This library uses a list of lexical features - words contained in the sentence which are labeled as positive or negative according to their semantic orientation - to calculate the sentiment of a text [29] [30]. The algorithm is applied on an entire original tweet - without any kind of cleaning or preprocessing - and it returns a dictionary with four scores of sentiment: *negative*, *neutral*, *positive* and *compound* (i.e. aggregate score). Only the last one varies in the interval $[-1, 1]$ - differently from the others which are something like probabilities defined in the range $[0, 1]$ - and it resumes the contribute of each sentiment features in one single value. If compound is near to -1 it means that the relative sentence tends to be more negative, on the contrary it tends to be more positive.

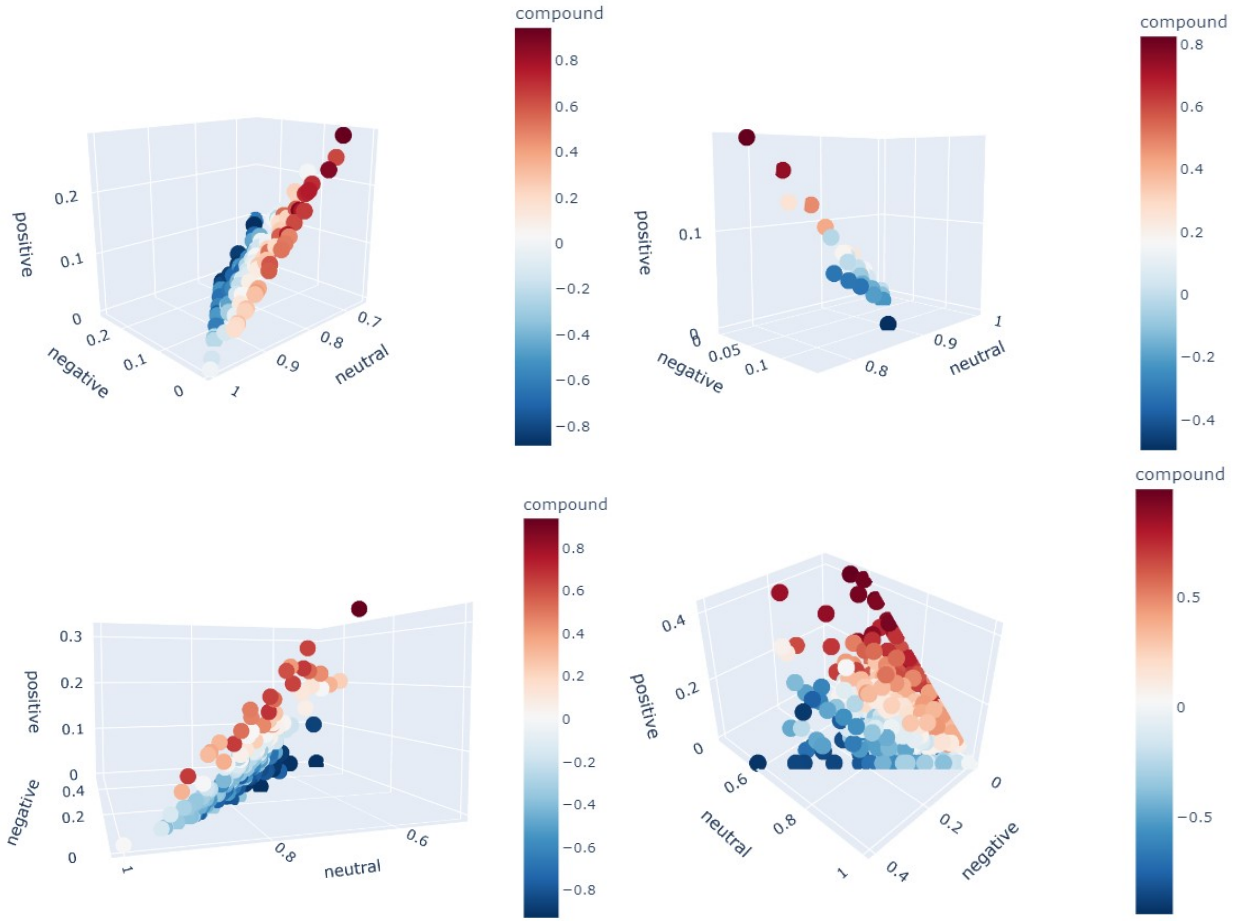
Since the scores are assigned to the sentence and not to the single word, we have developed a procedure that allows us to extrapolate a feeling for each of the nodes forming the community networks. Given the list of tweets belonging to one of the six communities considered - using the presence of the hashtags associated with a specific topic as a filter - and given its nodelist, we checked in which tweets the various nodes appeared. The score for the single word is therefore obtained by calculating an average of the scores associated with the tweets in which the word itself appears. A formalization of the procedure is the following: let W be one of the words in the nodelist of the community C and let $T = \{T_1, T_2, \ldots, T_w\}$ be the set of w tweets in which the word appears. If the scores for each tweet T_i is given by S_i - where the score is one of the available four listed

before - the final sentiment for word W is calculated as

$$S_W = \frac{1}{w} \sum_{i=1}^{w} S_i \tag{16}$$

where w is the total number of tweets to which the word W belongs.

A still image of a three-dimensional display of the four sentiment scores for each of the communities considered is proposed below. The values of negative, positive and neutral sentiment have been positioned on the three axes while the compound is coded through a color scale in which blue represents the minimum score and red the maximum one (referred to the community sentiment results). Since this originally was an iterative representation, different viewing angles have been reported in the various communities since the behavior of their scores is similar. This choice allows to acquire a better perception of three-dimensionality. What you notice is that the points tend to arrange themselves neatly along a plane. It is possible that a word with a non-zero positive score also has an associated negative score but the probability of obtaining higher values for one is inversely proportional to the other one.



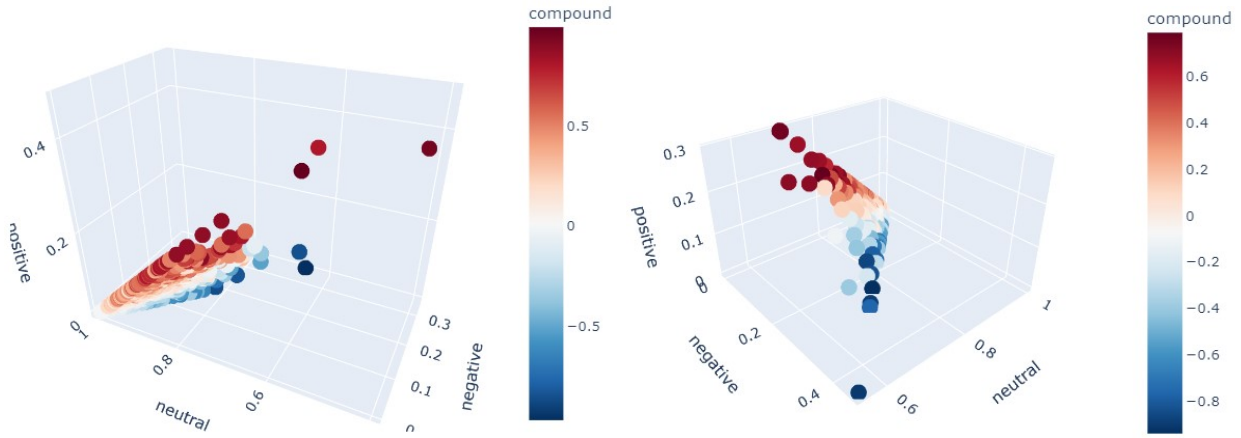


Figure 50: 3D comparison between the Vader Sentiment scores for each community

At this point it is legitimate to ask how sentiment acts on words in terms of centrality, for example by considering *pagerank* again as the main measure. Within a specific context, such as the topics of the six communities can be, one wonders what is the predominant sentiment of a word with a high importance in the network. Indeed, comparing centrality with sentiment can provide important information about the social and human nature of the community. For this reason, six views of the network are proposed, one for each community, in which the pagerank determines the size of the label as usual, while the compound - a score that summarizes the overall sentiment of the word in a thoughtful way - is represented by the color scale which has to be read appropriately. Since sentiment scores vary in a fixed range, we need to set the extremes of the color map on the reference values. In the case of compound, dark blue represents the minimum possible score -1, while the red indicates the maximum one, that is $+1$. We note that more than a possible correlation between centrality (in this case pagerank) and sentiment, what guides the emotion of a word is the context from which it is taken or the argument of speaking (the community topic).
For the first two communities (figure 51 right above), there is a juxtaposition of feelings: some words are extremely positive, others extremely negative. Not surprisingly, areas of feeling are created, whereby words with a sentiment tend to bond with related ones. For the *covid* community, is interesting to analyze the word "vaccine" which has an average negative sentiment. Could they be tweets from "no vax" people? For the *climate* related community the main nodes, i.e. words like "climate" and "change", take on a neutral sentiment.

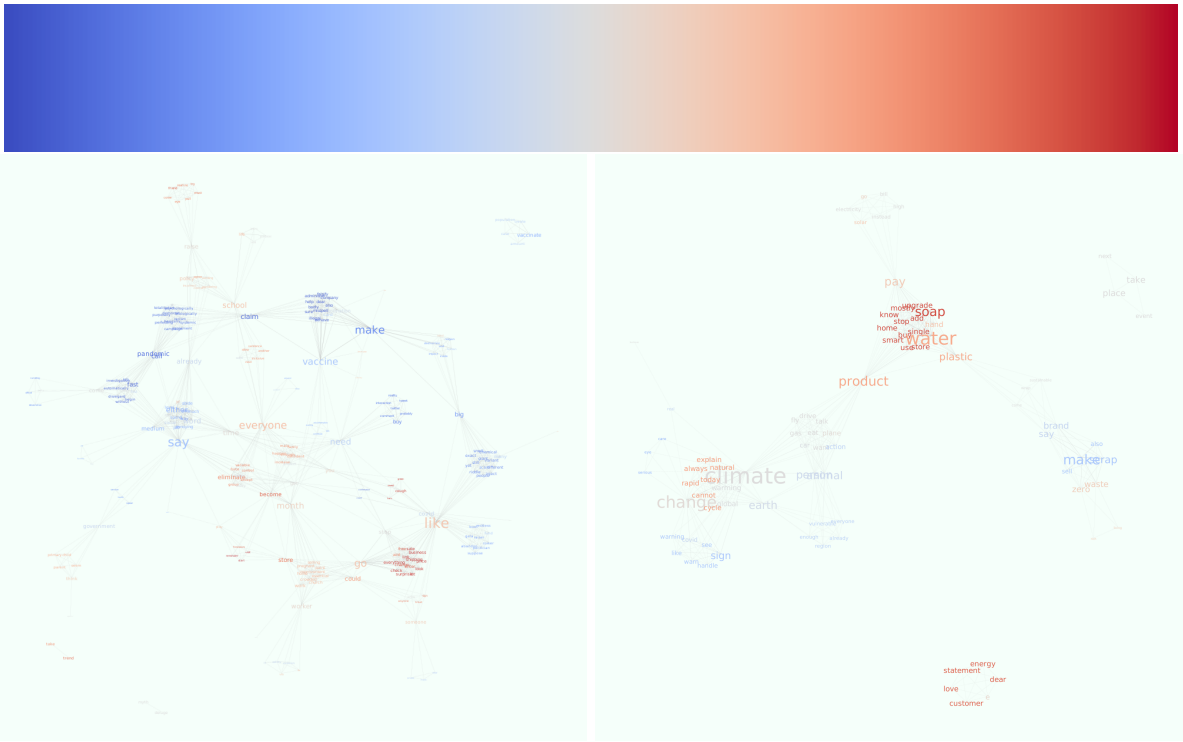


Figure 51: Pagerank (label size) and compound score (color scale) for covid and climate community

Passing to figure 52, the most recurring sentiment for tweets dealing with *abuse* episodes, of course, could only be negative. The words with a greater centrality are all negative; some words with a positive feeling are present but unfortunately they do not assume a role of significant importance in the network. An interesting insight concerns the sentiment of the first ones. It seems that in abuse-themed tweets the word "law" appears frequently with a negative connotation and the same fate touches the word "gun". This result has different interpretations: it could be assumed that people do not feel protected by the law or even that the relationship between the law and the use of weapons is not well seen.

For *mental health* themed tweets, the situation is reversed as the sentiment is on average positive or neutral if we take the centrality given by pagerank as the discriminating factor. This may reflect the social intention to emphasize the importance of the issue and encourage constructive debate. Could it be an attempt to normalize or even exorcise a topic which up to now days has been considered controversial and shameful? However, consider that words like "like" or "love" can affect the estimate of the score for the sentence, making it not perfectly compatible with the overall intent. Further analysis of the text could clarify any doubts. Moreover, we remind that among all the communities considered, this is the one that has the largest number of filtering hashtags, some of which intuitively have a positive connotation and others a negative one. The set of information is therefore composite in sentiment and the consequence of this is that the average sentiment could in some way be conditioned by heterogeneity. Some tweets can therefore misleadingly affect the estimated sentiment for a word.

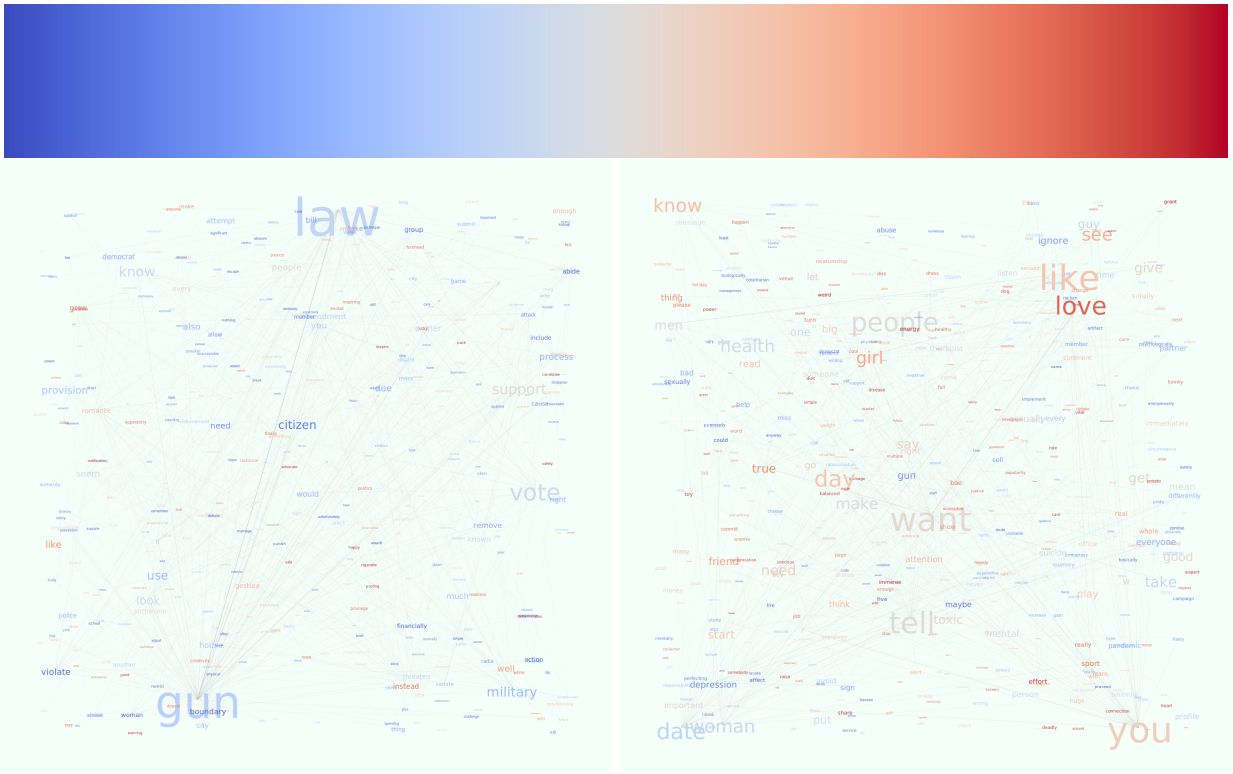


Figure 52: Pagerank (label size) and compound score (color scale) for abuse and mental health community

The sentiment associated with the *dating* community is globally positive - this aspect may depend on the particular selection of hashtags to create the community - while that of the *market* one is composite.

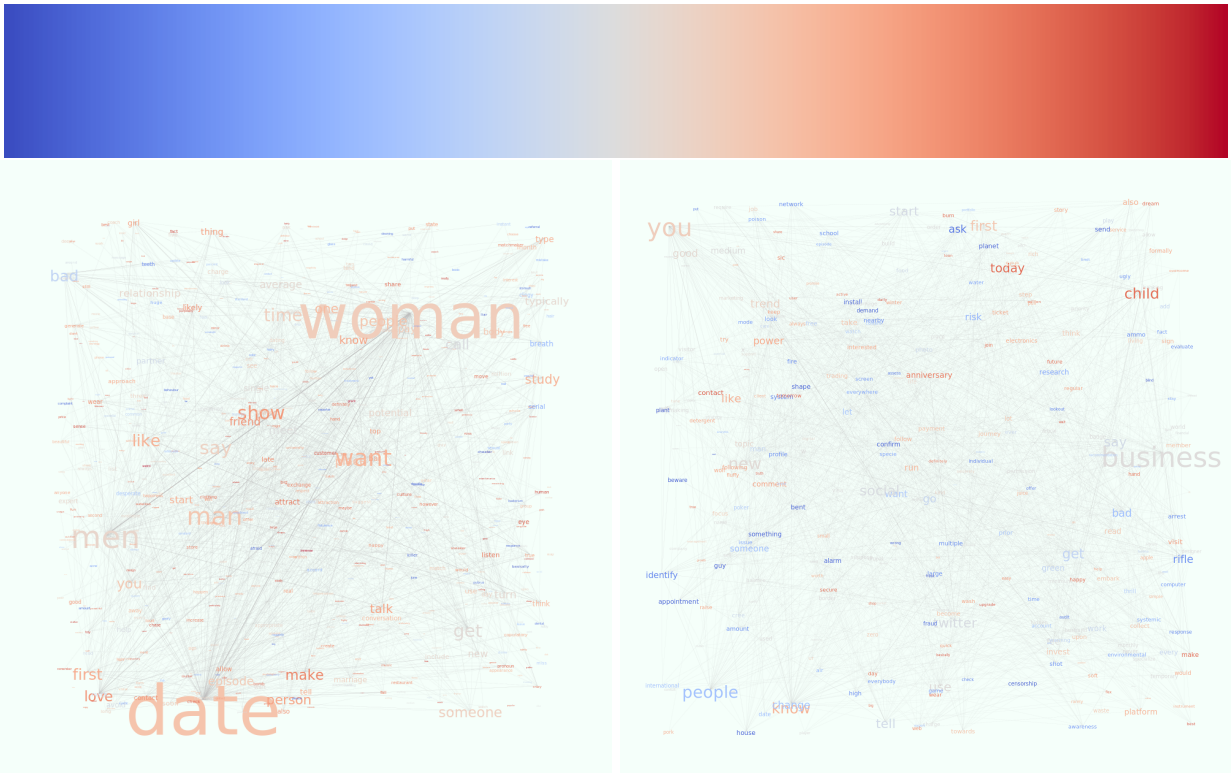


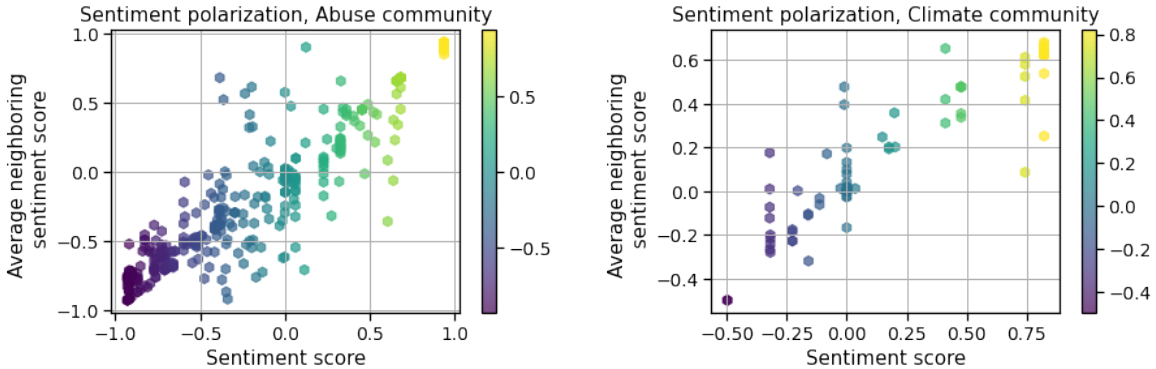
Figure 53: Pagerank (label size) and compound score (color scale) for dating and market community

11.2 Correlation with neighboring sentiments

Camilla Quaglia

In this section we exploit the results of *Vader Sentiment* algorithm to answer the following question: does the sentiment of a node correlates with the sentiments of its neighborhood?

As discussed in the section above, *Vader Sentiment* assigns to each node/word a *compound* value, based on the context/Tweet this word is in. This compound is a number in the range [-1,1], where values near 1 indicate that word is *more* positive, conversely values close to -1 indicate the opposite. In the graph in figure 54, there are the sentiments, on x-axis, of each word, while on the y-axis the average sentiment of the nodes neighbor.



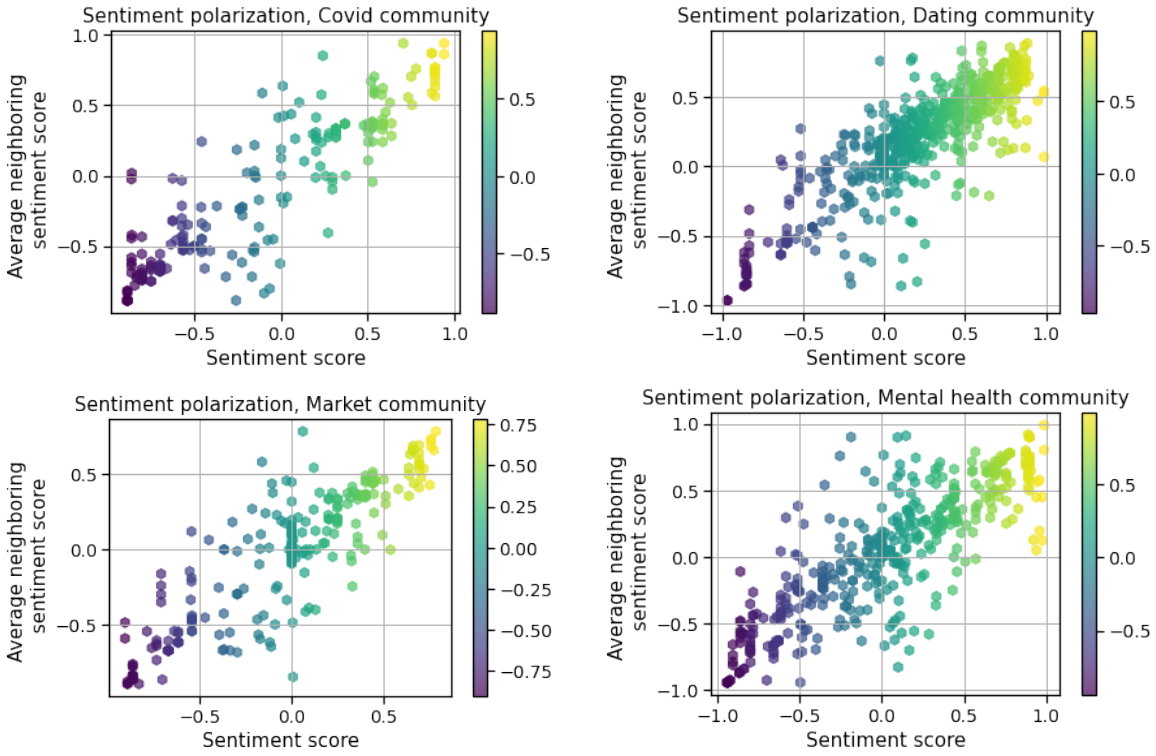


Figure 54: Average neighborhood sentiment vs each node sentiment.

From the figure 54 we can make some conclusions. All the communities present a positive correlation between the *compound* (sentiment) score of a node and the score of its neighborhood. This means that positive words are connected, therefore present in the same tweets, with positive words, as well for negative and neural ones. Also the *Abuse* community presents an high density of negative sentiments, in accordance with previous analysis. One can be a little surprised by the *Mental health* community, where the sentiments are also very positive. This can be a sign of a social debate on this topic.

11.3 Text Blob

Dacia Braca

TextBlob is a Python library for processing textual data. It provides a simple Application Programming Interface for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. For lexicon-based approaches, a sentiment is defined by its semantic orientation and the intensity of each word in the sentence. This requires a pre-defined dictionary classifying negative and positive words. Generally, a text message will be represented by bag of words [29] [30]. After assigning individual scores to all the words, final sentiment is calculated by some pooling operation like taking an average of all the sentiments.

As in the previous case, the algorithm is applied to the entire sentence and the output is given by a pair of scores, respectively called *polarity* and *subjectivity*, which summarize the sentiment conveyed by the sentence. The first is a float between $[-1, 1]$, where -1 indicates negative sentiment and $+1$ a positive one; the second is still a float but in the range $[0, 1]$ and generally refers to opinion or judgment. In particular, subjectivity quantifies the amount of personal opinion and factual

information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. Also in this case the score is associated with the tweet so to obtain the sentiment on the single word it is necessary to make an average calculation considering the appearances of the word in the various tweets, following the procedure described in the section 11.1.

For this second sentiment analyzer, let us consider the views of the network for each community with the usual convention to obtain a comparison with centrality measure: the size of the label is given by the pagerank score while the color is encoded in a sentiment score. However, it is necessary to make an important note about the visualization and the assignment of the color to the word according to the associated score. As anticipated in the previous section 11.1, the color map for the study of sentiment must coincide with the range of definition. In the case of polarity, dark purple indicates the minimum value fixed at -1, while yellow indicates the maximum $+1$. For the subjectivity the same rule applies but the lower bound corresponding to the dark blue is mapped to the value 0. For each community, the display pairs associated with the sentiment scores for this second library are reported. What we can immediately notice by quickly consulting the representations with polarity, is that for all communities *TextBlob* sentiment analyzer returns an almost neutral score (color of nodes has a range focused on the map's center). Therefore more interesting is the analysis of the subjectivity score, compared to indicators of *Vader*, provides some more information.

The results for the *covid* community are shown below. Here we find a good variability for this score even if the most central words in the network tend to be inserted within tweets with a medium-high factual component (darker colors). There are groups of nodes with a high sentiment of opinion but these do not have a main character of influence.

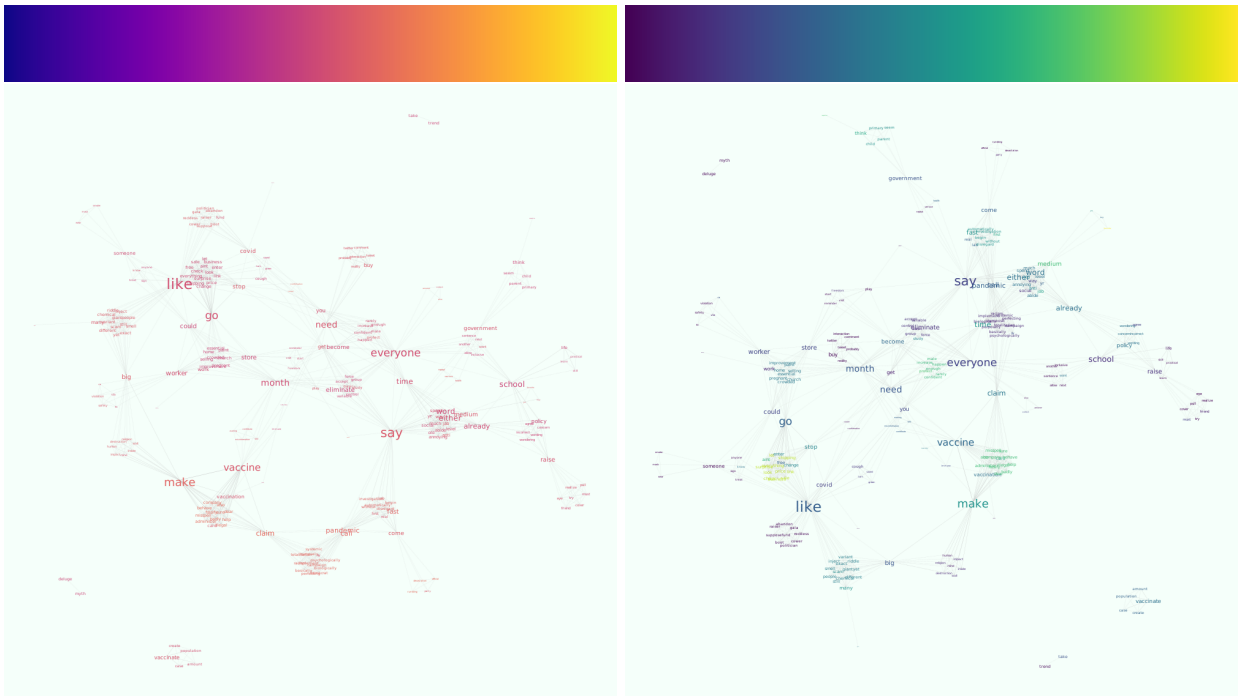


Figure 55: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for covid community

Similar considerations can be made for *climate* community. In general, however, words with

a high emotionality do not appear. It seems that the topic is approached with a good level of objectivity.

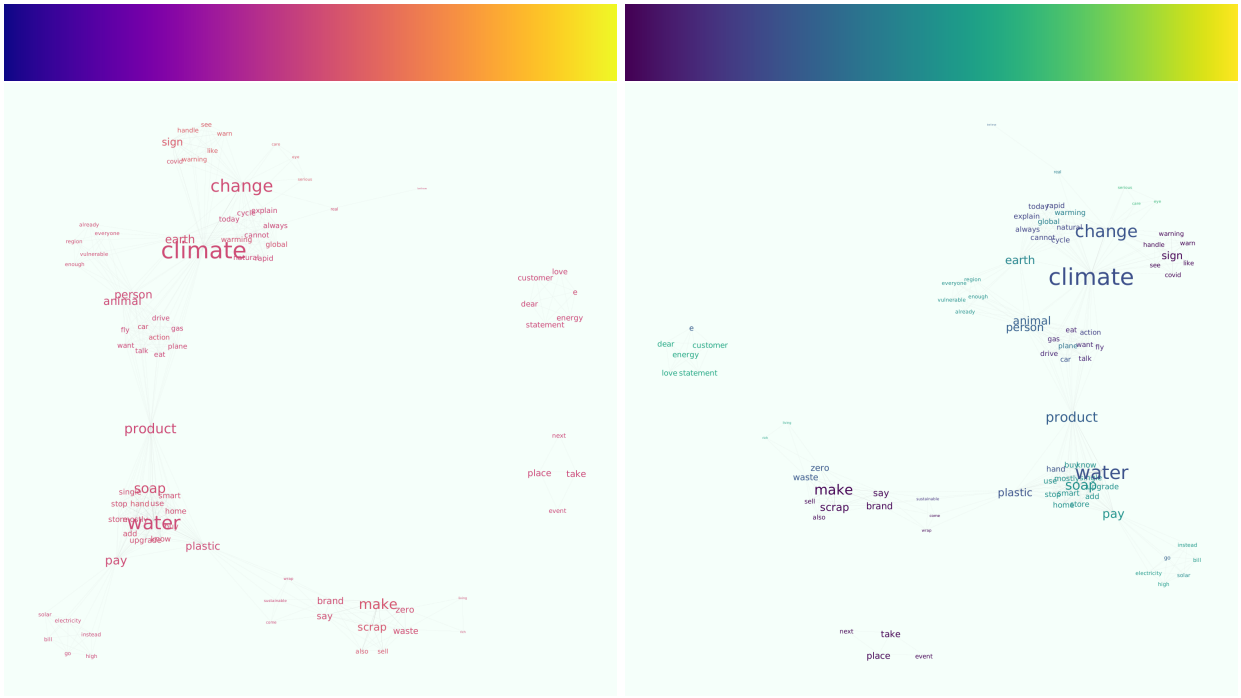


Figure 56: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for climate community

Even in the case of the community regarding *abuse* it is noted that the words with high centrality present at the same time a good objectivity. Words with a subjective character appear but this does not assume a marked centrality.

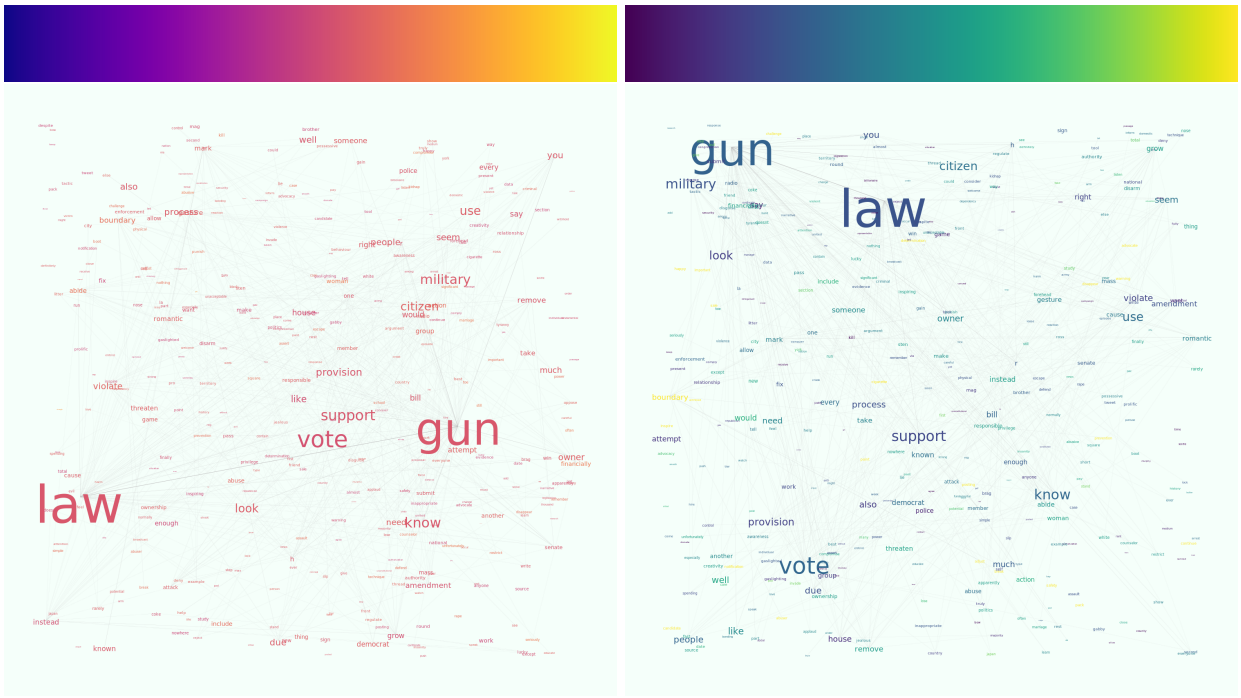


Figure 57: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for abuse community

For the *mental health* topic, average behavior tends to be subjectivized. People are likely to use a higher level of emotionality and non-factuality when tweeting about social and mental arguments.

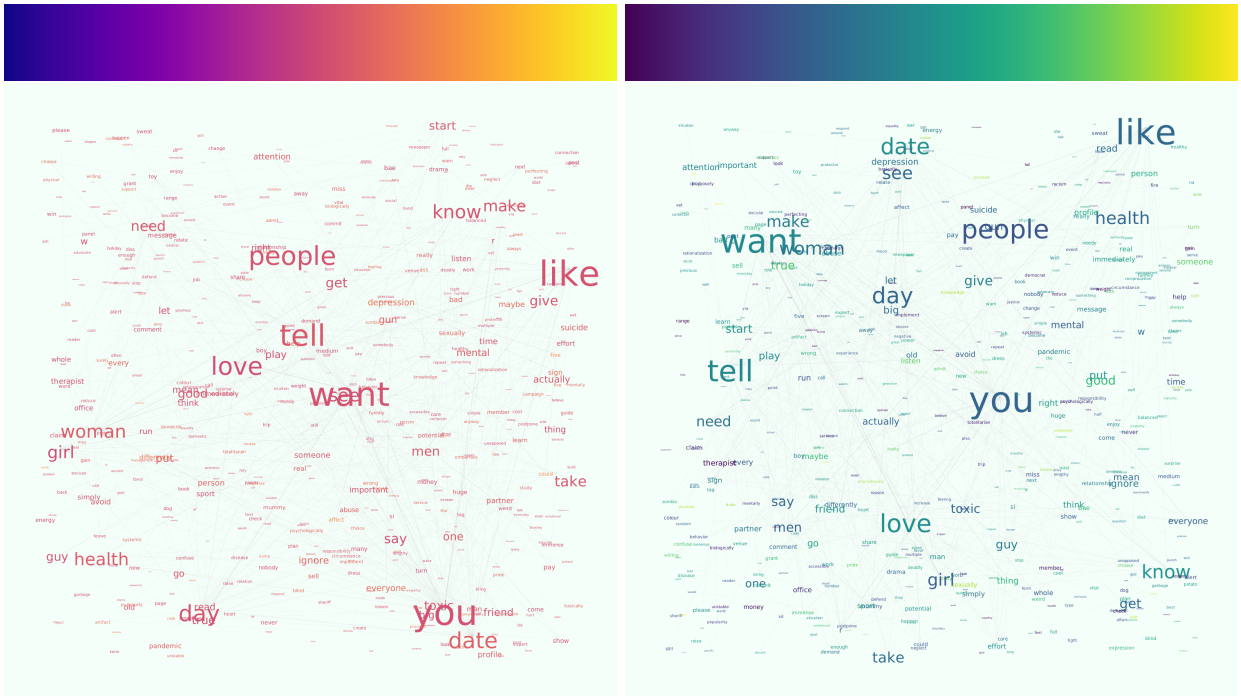


Figure 58: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for mental health community

The average subjectivity score becomes even higher in the case of *dating* community. According to the representation, people tend to talk about social relationships with an increasing level of emotion. In fact, even the most central nodes such as "woman" and "date" have a good level of subjectivity.

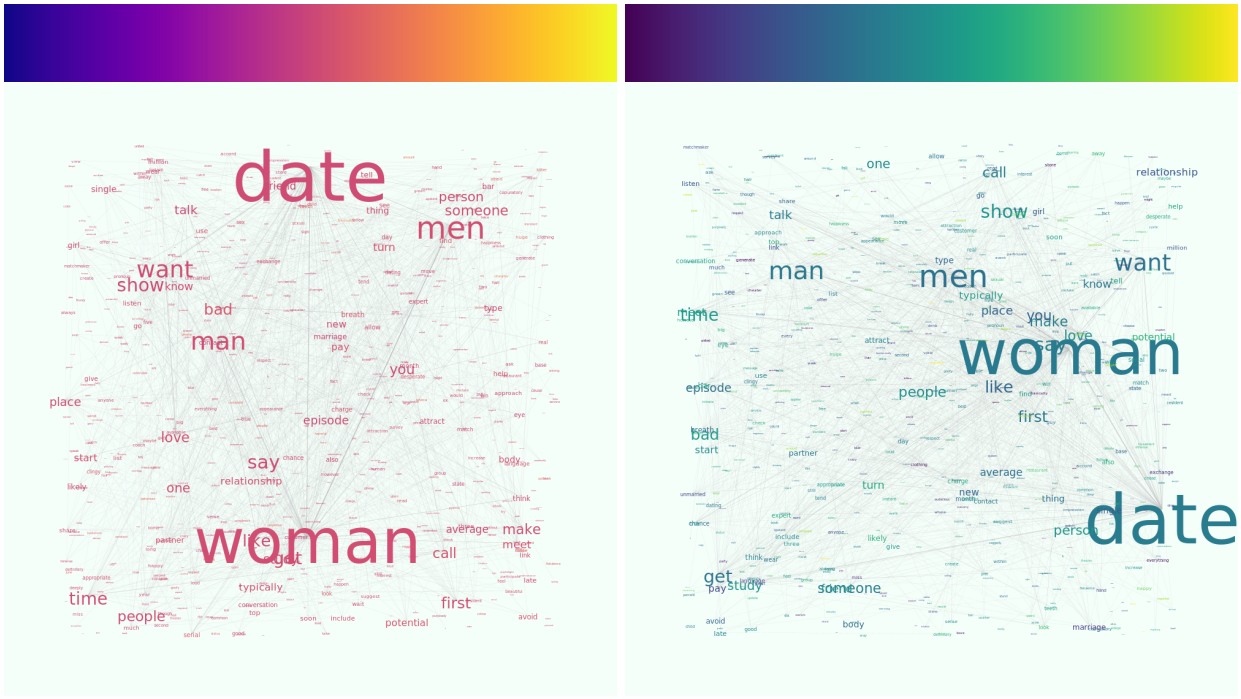


Figure 59: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for dating community

The latest community on *market* issues has a mixed behavior. The most central words generally have greater objectivity than the others.

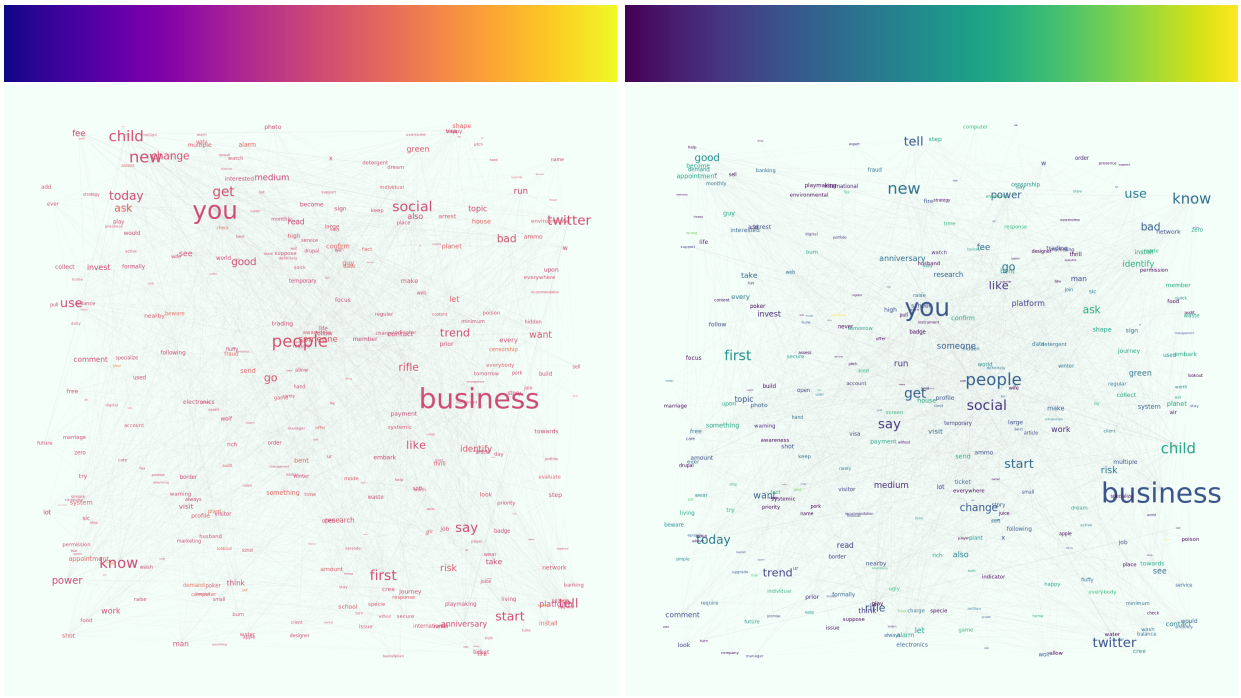


Figure 60: Pagerank (label size) with respect to polarity and subjectivity score (color scales) for market community

11.4 Linguistic Inquiry and Word Count (LIWC)

Michele M. Crudele

In this section related to sentiment analysis, we make use of the Linguistic Inquiry and Word Count (LIWC) program [16]. Basically, it reads a given text (in our case a tweet) and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Because LIWC was developed by researchers with interests in social, clinical, health and cognitive psychology, the language categories were created to capture people's social and psychological states.

In order to make it work, we can use the original tweets, from which we only remove emoji and change all the *u* characters in *you*, as explained in section 4.2. Since our goal is to compare the feelings among the different topics we are analyzing, we run it on subsets of the original dataset, each one containing only the tweets related to one of the communities found in section 5.2. The output of LIWC consists in a .csv file with different sentiment scores associated to each single tweet. From it, we can obtain the table in Figure 61 by computing their average values, that we can consider as the sentiment scores of the community they refer to.

	MENTAL_HEALTH	COVID	CLIMATE	ABUSE	DATING	MARKET
Analytic	63.72	68.25	70.57	73.93	71.77	73.28
Authentic	20.68	16.91	27.06	6.62	22.08	19.21
Tone	44.29	47.67	44.52	34.65	48.64	47.43
i	1.40	1.95	0.58	0.40	1.39	1.22
we	0.41	1.45	0.51	0.41	0.40	0.98
you	2.46	1.44	1.19	2.16	2.27	1.89
they	0.95	0.73	0.61	0.74	0.47	0.40
negate	1.55	2.25	1.84	1.11	1.51	1.63
verb	10.60	11.00	12.06	10.38	9.50	9.73
adj	4.02	2.86	2.65	3.28	3.91	4.12
interrog	1.46	1.23	1.89	1.04	1.35	1.37
affect	6.19	3.48	4.44	4.68	4.29	3.66
posemo	3.59	2.24	3.06	2.04	2.97	2.65
negemo	2.48	1.24	1.38	2.65	1.25	1.02
anx	0.26	0.17	0.29	0.27	0.20	0.21
anger	0.66	0.48	0.00	1.56	0.31	0.31
sad	0.76	0.16	0.29	0.20	0.26	0.09
social	11.17	8.31	5.41	8.20	14.80	11.00

	MENTAL_HEALTH	COVID	CLIMATE	ABUSE	DATING	MARKET
female	1.48	0.15	0.00	0.33	1.65	0.23
male	1.26	0.24	0.00	1.31	1.55	0.67
insight	1.69	1.47	1.20	1.30	1.14	1.24
feel	0.17	0.00	0.43	0.22	0.16	0.10
health	2.24	1.36	0.56	0.28	0.47	0.59
sexual	0.11	0.11	0.00	0.16	0.26	0.05
risk	0.51	0.50	1.03	0.66	0.37	0.57
focuspast	0.83	1.48	1.34	1.88	1.02	1.01
focuspresent	8.52	8.24	9.24	6.85	7.62	7.53
focusfuture	0.51	0.42	1.15	0.72	0.58	0.50
money	0.60	1.49	2.76	0.53	0.54	6.17
relig	0.25	0.17	0.00	0.00	0.09	0.00
informal	4.52	3.02	3.60	3.34	3.36	5.26
swear	0.58	0.17	0.00	0.08	0.15	0.10
netspeak	3.83	2.51	2.80	2.75	3.09	4.98
assent	0.19	0.10	1.32	0.14	0.36	0.15
QMark	0.89	0.71	0.86	1.09	1.06	0.48
Exclam	0.95	1.47	0.73	0.52	0.55	1.23

Figure 61: LIWC results for the different communities of topics. By comparing the value of a categeory for the different topics we can gain interesting insights about the feelings that characterizes them.

Looking at it we can gain some interesting insights about the communities. The market topic is, of course, the one with the highest money score. Abuse is the topic with the highest score of *anger*, *negemo* (negative emotions) and *focuspast*, while it has the lowest value in the *health* category. Climate community is the one that focuses more on the future but also on the present and has the highest score of anxiety (*anx*) and risk. Mental health shows the highest values of positive emotions, affective processes and health, while having the lowest value of analytical thinking, due to the presence of many encouraging tweets. Dating community is characterized by the highest scores of sexual and social categories. Finally, the Covid-19 community is the one with the highest score of exclamations and negation, probably due to tweets written by deniers and no-vaxes.

All these results are really sensible and also consistent with the ones obtained in the following section 11.5.

11.5 Positive & Negative Nodes Removal

Michele M. Crudele

In this section we test the robustness of the six communities of topics to the removal of different kind of sentiments: positivity, neutrality, negativity and subjectivity. We do this using the results obtained in sections 11.1 and 11.3, that associate different scores of different sentiments to each word contained in the communities we analyze. What we do here is to exploit those results to remove from the semantic networks different fractions of all the words related to a specific feeling and we see how this deletion impact the size of the semantic networks: the more fragile a network is to the removal of a specific sentiment, the more it is related to that feeling.

To reach this goal, we first have to decide a rationale for determining whether a word is to be considered as positive, negative or neutral. Indeed, the *compound* variable in the results of

section 11.1 assumes continuous values between -1 (negative word) and 1 (positive word). As a consequence, we decide after some trials to consider as neutral all the words whose compound is between -0.3 and 0.3; if the compound is greater than 0.3, we consider it as positive; if the compound is less than -0.3, we consider it as negative. In the same way we use the variable *subjectivity* from section 11.3: we consider a word as subjective if its subjectivity value is greater than 0.3.

Just like in section 9, on the y-axis we plot the relative size of the giant component. In the following plots we can see how it evolves if we remove positive, negative, neutral or subjective words from them. In particular, the plots in Figure 62 shows the effect of the removal of different feelings for each community, while the plots in Figure 63, 64, 65, 66 show the robustness of all the communities to the removal of the same kind of feeling.

From the plot in Figure 62 related to the abuse topic, we can see how that community is much fragile to the removal of words associated to a negative mood: if we remove all of them, the giant component loses more than the 50% of its nodes. This means that the tweets related to abuse are very characterized by negative feelings, as expected. Looking at Figure 64 then, we can also notice that the abuse topic is the most related to negative sentiments among all the topics we are analyzing. On the contrary, the most positive-feeling-oriented topic is *dating*: this is due to the fact that many times people write about past dates using the hashtag #RedFlag in a funny way, laughing about situations or people they went on a date with. From Figure 62 we can see how climate and market topics are instead mostly related to neutral feelings; this could indicate a high presence of facts rather than opinions, and indeed this is what Figure 66 confirms.

The plot about subjectivity in Figure 66 shows clearly that tweets about dating are the most subjective ones together with those about mental health, while those related to business and climate contain less opinions and more factual information; abuse topic is more or less in the middle. From that plot we also notice that the Covid-19 community is one of those with the lowest presence of personal opinions. This is a very interesting insight about that community, especially if merged with the information about it that we gain from the table in Figure 61 (that assign a high score of negation and exclamation to the Covid-19 community): deniers and no-vaxes are able to present their theories making them seem like facts rather than opinions.

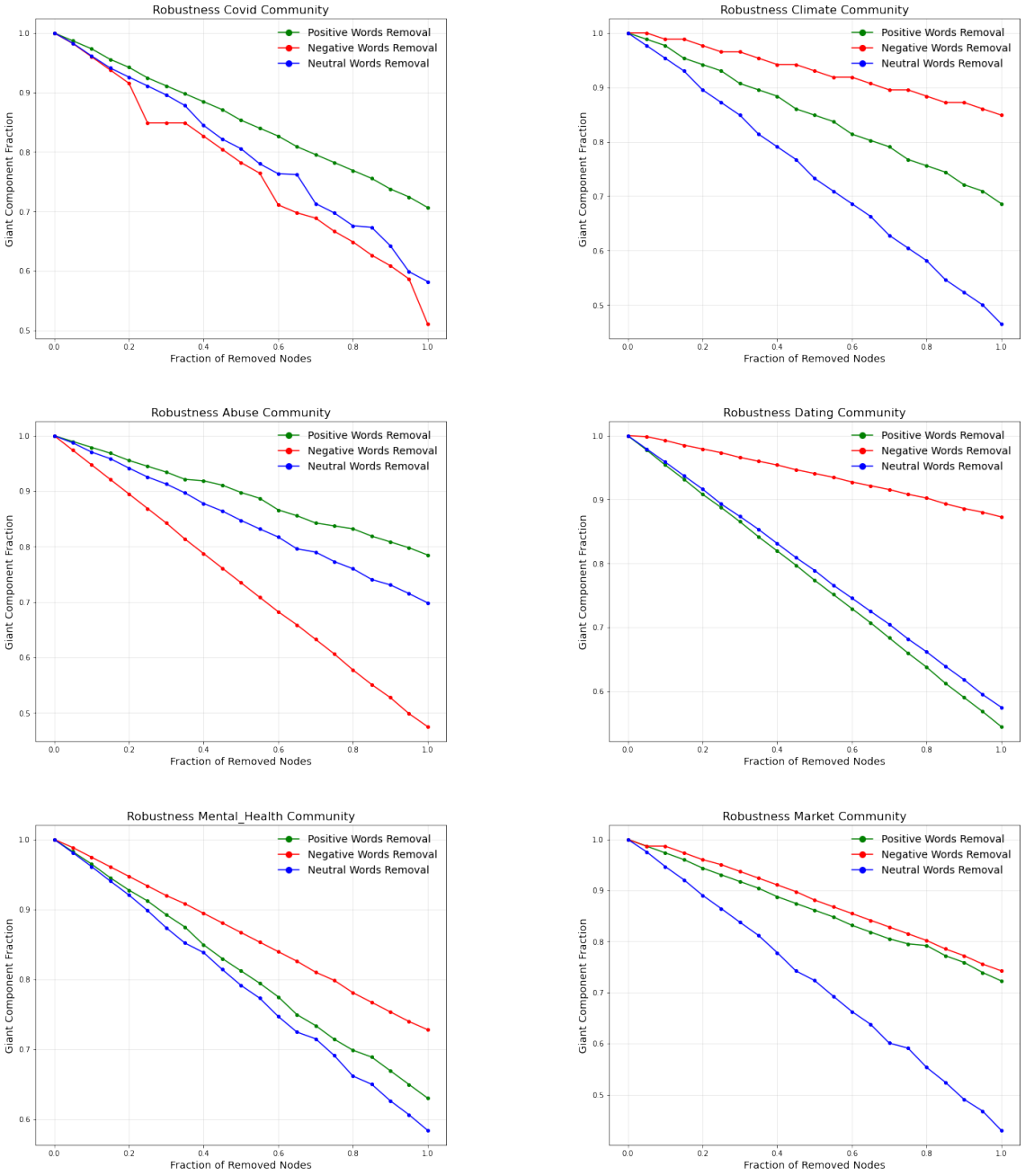


Figure 62: Robustness of the six communities we consider to the removal of different kind of sentiments. Negative feelings characterizes the Covid-19 and especially the abuse topic; dating and mental health communities are instead related to positive feelings, while climate and market topics are neutral, as confirmed also by the plot in Figure 66.

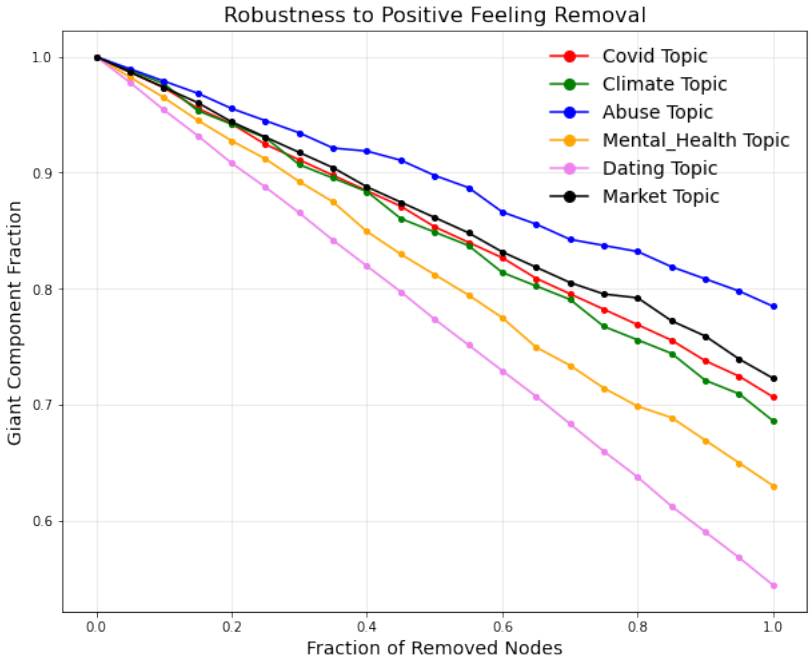


Figure 63: Robustness of the different communities to positive nodes removal. The dating topic is the one with the highest presence of positive words, due to the great amount of tweets talking about past dates in a funny way. On the contrary, abuse community is the one with the lowest presence of positive feelings, as expected.

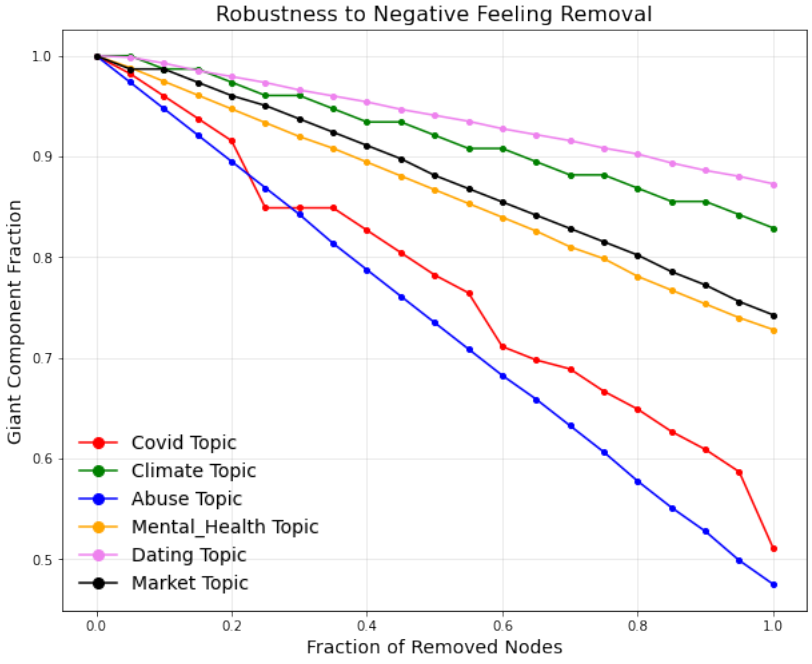


Figure 64: Robustness of the different communities to negative nodes removal. These results confirm those obtained from Figure 63.

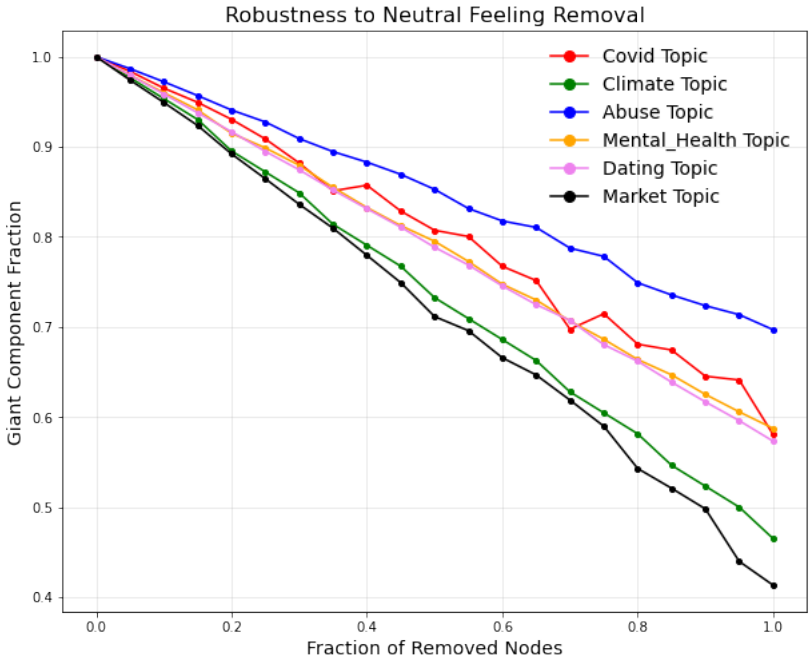


Figure 65: Robustness of the different communities to neutral nodes removal. The market and climate topics are the ones with the highest presence of neutral words, while abuse community is more polarized.

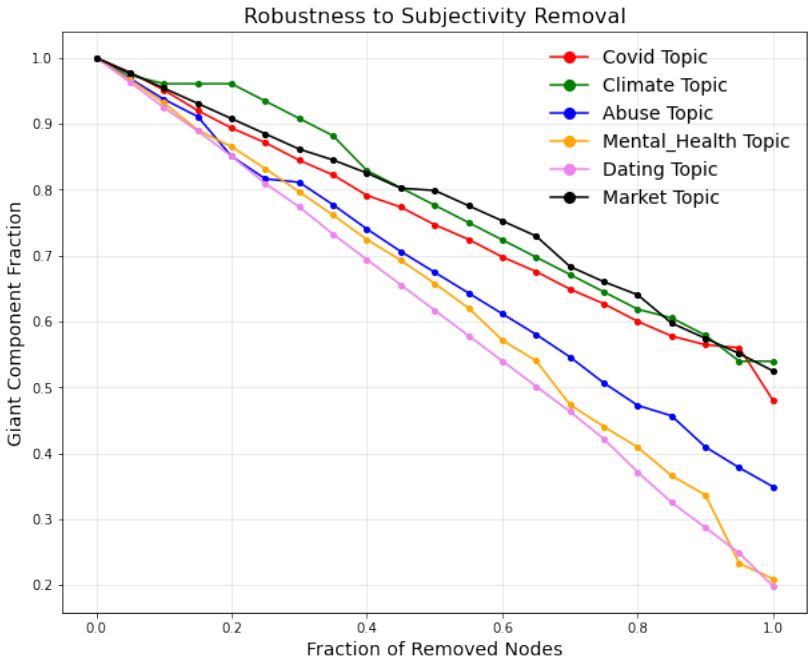


Figure 66: Robustness of the different communities to subjective nodes removal. Dating, mental health and abuse topics contain more personal opinions w.r.t. covid, climate and market topics, more related to facts.

12 Summary of the Results

We have analyzed the topics related to the RedFlag trend on Twitter through the network of hashtags linked to *#RedFlag*. The more interesting topics we have found are about Covid-19, climate change, abuse, mental health, dating and market. For each of those topics we have built a semantic network of words and we have analyzed each one of them using different Network Science techniques and tools. The analysis of the degree distribution, centrality measures and robustness has revealed their scale-free nature. Then, different tools have been exploited to perform a sentiment analysis of the semantic networks and they have all produced sensible and consistent results: dating topics are characterized by positive feelings, while abuse is the topic with the highest score of negativity; market and climate change topics are instead mostly related to factual information, with a low score of subjectivity in their tweets. A set of tweets not containing the hashtag #RedFlag has been downloaded to check that the results are in fact influenced by the presence of that hashtag in the tweets.

13 Discussion

Gala Pradillo Díaz & Jacqueline Pamela Padilla Torres

Based on the results obtained in the analysis and extraction of data related to the hashtag red flag on Twitter, six communities have been determined. More detailed research has been carried out, in order to obtain more sensitive results of the social spheres in which this trend has influence.

These communities show us a reflection of the basic field of a society such as public health, climate, interpersonal relationships and economy. As mentioned earlier in the report, the practice of targeting and exposing someone or something because one disagrees with their beliefs or practices has been done for a long time in a wide variety of contexts, both in the personal place and in the non-profitable space, also specifically in politics. But today the effect this can have on a person's life can be devastating and can have consequences that would not have been possible without the power of social media. (Forbes, 2017) Minor issues can morph into major issues, and major issues can become huge problems once someone hits "send" as stories, photos and commentary spiral out of control. This tendency has served as a social method of exclusion and even retribution for individuals in both the personal and private spheres, although it is less common but certainly still exists in the case of the private sector, referring to companies.

The report hypothesizes that if red flags are seen as social standards (norms), the cancel culture is the reaction upon acting in concordance within those red flags.

In the semantic network that refers mostly to words about the Covid. It tends to be related to the polarization in society nowadays. Such as the vaccination debate or the reaction of society to the government's measurements. This is common into the hegemonic culture since it appears a counter-discourse, reflected here as opposite opinions. In the case of the pandemic, the information was changing fast and it was necessary to avoid spreading false information in social media through personal opinions. There were a lot of public figures giving their opinion about the topic. An example of this is the actress Vanessa Hudgens. In March 2020, through the social network Instagram, she made a statement which promoted disinformation to her followers, inciting them to not follow the quarantine measurements. This behavior was directly related to red flags according to the results. As a consequence of this act, people started talking about the influence she has with a specific public, important aspect that led her to being socially "canceled". (Cooper, 2021)

However, this phenomenon of the "cancel culture" can also have a negative as well as positive associations, such as in the case of the semantic network that refers to topics about the climate. A red

flag can be related to the consumption habits of an individual, as we see reflected in words such as "plastic, waste, animal products", however it is feared that the so-called "cancel culture" is related to the fact that young people stop raising their voices against the climate crisis because they get carried away by the idea of "fanaticism". This community is rank in the LWIC with a high number related to the focus in the present, when for a positive impact it should be focused more in the future. Taking into consideration the topics related to consequences of climate change. However, if this issue is inclined to the cancellation of companies or individuals for promoting products and practices that are harmful to the environment for profit, it can have a positive impact that can be transferred from the virtual world to the day-to-day life of a community. In the LWIC we can see that the rank of verbs here is high, which can be interpreted that actions are being proposed around the main words. The community related to "abuse", corresponds to social issues that have been of great relevance nowadays in the politics and social arena. It refers to issues around gun laws in the US, elections and voters. As Habermas refers, citizens in the public sphere discuss among themselves topics that concern the general interest, and even issues regarding the regulation of civil society.

In the field of social interrelations, there's also the community around the topic of dating, which is focused on topics that have a higher numbers in the LWIC indicators such as social, female, male and feelings in comparison with the rest of the communities, which shows that this kind of red flags are specific to the behaviors for interactions with high feelings. In this community a clear evaluation of the behaviors is done, in the moment that somebody talks about aspects that correspond into the private sphere, such as the intimacy details of a date, this becomes part of the public sphere. The consequences of open commenting about personal experiences, can expose a person's reputation in an exponential dimension. An example of cancel culture for this case could be the one of Justin Timberlake, since after the release of the documentary of former partner Britney Spears, people start talking about the way he demonized her and her career after they broke up, showing public harmful behaviors that condones misogyny. Even though these actions happened more than 10 years ago, when it came back to the spotlight, had the consequence to lead Timberlake into the cancelled public personalities list. Since the ways of communication became more accessible to everybody through the internet, a new paradigm of communication opened a wider public sphere. The red flags and its consequences as part of the cancel culture are also related to fields, such as mental health. This community has a negative sentiment, here it is clear that, in order for someone to be cancelled, one has to decide to share openly about a past experience and the red flags that are involved. It can be related to topics of depression, abuse, toxic behaviors in a relationship and even related to self esteem. It can be understood in this community, the movement of the private sphere reflected in personal and hurtful experiences being transformed into a public sphere at the moment of openly sharing.

For example, the case of Ellen Degeneres. She was accused of creating a toxic environment for her set personal, once one started talking about uncomfortable experiences, a chain of opinions from others were also shared and they all agreed on the patter of toxic behaviors. Causing people to speak up and change the reputation she had in the public arena. The pattern of re-emergence after being "canceled" is not uncommon for people with certain levels of privilege. But for private citizens, a mistake caught on social media can follow them for years to come.

To conclude, we can observe in the report that actually the red flag hashtags are part of what people discuss in the public sphere. In fact, we can relate it with Habermas point of view, as by this trend people make public things about their private lives, but which concern somehow the general interest (like the dating advice). Also, they help to produce morals, explanations, perceptions, values or customs of the dominant culture, for example encouraging people to vaccinate or to act actively against a common problem. Anyway, red flags also show alternative ideas for example against the

vaccine, therefore they also produce counter public spheres.
On the other hand, the cancel culture works as an immediate reaction of the society to punish individuals that act against the morals, perceptions and values already determined. This can be possible even without a facible proof of the act, just from the trust in the statements shared in the public sphere through the social media. Here it is important to emphasize the power that the internet has given to the words of an individual. In a matter of a short time the reputation visualization of somebody can be completely altered.

14 Limits & future directions

After analyzing the data obtained from the research of the red flag hashtag, it is clear that the phenomena focus on specific behaviors, most of the time, giving warning signs about the social relationships in different fields. While the Cancel culture focuses on acts that involve the unacceptable behaviors in a society and have concordance with the red flags but are not necessarily connected to each other or in a codependent relation.

The obtention of data from specific communities can be useful for further analysis. In this case it was used the command *#redflag*, but users in this trend also tend to use the emoji of the flag instead of using the word. This means that in further research, it would be convenient to select another type of keywords and to do the investigation from another point of view, taking into consideration the trend of using emojis instead of hashtags. The migration of the trend from one platform to another also changes the way it is perceived, since the form of expression and the age of the public using that social media differs. This means that the meaning of the "red flag" trend has changed over time and also influences how people use it.

In the case of Twitter, while in some topics, red flags were still used to express an opinion, in some others they were used for advertisement reasons or even turned to be used as a "meme", joking about the people who used it in a sarcastic way. In this case, the hashtag changed the way it was perceived and changed the meaning that society used to give it. Making it a less serious trend.

Regarding how the trend is part of the creation of values of the dominant culture, and how Twitter is structured in terms of power, it should be sensible to investigate which users, including public figures and companies, have more interactions and therefore more visibility. In other words, it should be taken into account how Twitter, as a business looking forwards economic benefits, stratifies the visibility of specific messages, profiles and trends, in favor of the dominant culture and the ones that rule it.

References

[1] Alexander, J. (2020). K-pop stans overwhelm app after Dallas police ask for video of protesters. The Verge, VOXMEDIA. Available at: https://www.theverge.com/2020/6/1/21277423/k-pop-dallas-pd-iwatch-app-flood-review-bomb-surveillance-protests-george-floyd

[2] Castrelo, V. (2018). The habermasian public sphere. Its obsolescence in times of new digital platforms- InMediaciones de la comunicación, vol. 13 nº 1, pp 71-87 Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=6698225

[3] Cooper, M. (2021) Stars Who Drew Backlash for Their Controversial Views on Coronavirus: Vanessa Hudgens, Evangeline Lilly and More. Available at https://www.usmagazine.com/celebrity-news/pictures/stars-who-drew-backlash-for-their-controversial-views-on-coronavirus/

[4] Dictionary.com (2020). Cancel culture. Pop culture dictionary. Available at: https://www.dictionary.com/e/pop-culture/cancel-culture/

[5] Echo Chamber (media) (2022). At Wikipedia. Available at: https://en.wikipedia.org/w/index.php?title=Echo_chamber_(media)&oldid=1068449862

[6] Filter bubble (2022). At Wikipedia. Available at: https://en.wikipedia.org/w/index.php?title=Filter_bubble&oldid=1065563874

[7] Forbes (2021). Be Careful: Cancel Culture Is Here To Stay. Available at: https://www.forbes.com/sites/theyec/2021/10/05/be-careful-cancel-culture-is-here-to-stay/?sh=30005fc62ac7

[8] Habermas, J.; Lennox, S. & Lennox, F. (1974). The public sphere. An encyclopedian article. New German Critique, 3, 49-55. Retrieved: 12/12/2017. Available at: http://www.jstor.org/stable/487737

[9] Hailstone, J. (2021). Cancel Culture Fears Stopping Young People Speaking Out On Climate Crisis, Study Finds At Forbes Available at https://www.forbes.com/sites/jamiehailstone/2021/12/20/cancel-culture-fears-stopping-young-people-speaking-out-on-climate-crisis-study-finds/?sh=54e07eb5283a

[10] Hegemonía Cultural (2021). At Wikipedia. Available at: https://es.wikipedia.org/w/index.php?title=Hegemon%C3%ADa_cultural&oldid=140068635

[11] Peña, G. (2018). La construcción de la hegemonía en redes y prensa: la disputa del relato 15M. Universidad Cumplutense de Madrid, Facultad de Ciencias de la Información. Available at: https://eprints.ucm.es/id/eprint/47725/

[12] Wimmer J. (2005) Counter-Public Spheres and the Revival of the European Public Sphere. Javnos The Public, 12:2, pp. 93-109. Available at https://www.tandfonline.com/doi/abs/10.1080/13183222.2005.11008890

[13] Williams J. (2021) Justin Timberlake Apologizes to Britney Spears and Janeth Jackson for "Not speaking up". Available at https://www.newsweek.com/justin-timberlake-britney-spears-apology-1568982

[14] https://www.dictionary.com

[15] https://github.com/JustAnotherArchivist/snscrape

[16] https://liwc.wpengine.com

[17] https://docs.python.org/3/library/re.html

[18] https://www.nltk.org

[19] https://gephi.org

[20] https://en.wikipedia.org/wiki/PageRank

[21] https://en.wikipedia.org/wiki/Louvain_method

[22] http://networksciencebook.com/chapter/8

[23] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679

[24] https://networkx.org

[25] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html

[26] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html

[27] https://networkx.org/documentation/stable/reference/algorithms/centrality.html

[28] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.hits_alg.hits.html

[29] https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/

[30] https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664