Covid-19 related tweets inside and outside China

Alavi Seyedhamidreza Student nr. 2005952

Bigarella Chiara Student nr. 2004248

Cogato Matteo Student nr. 2026966

Gicquel Thomas Student nr. 2031779

Liang Yiling Student nr. 2008497

Liu Yichen Student nr. 2005920

Mellino Daniele Student 2013373

Poletti Silvia Student nr. 1239133

Rayri Akram Student nr. 1215988

Trolese Francesco Student nr. 2050691

Yang Jingwen Student nr. 2046837

Zangrando Emanuele Student nr. 2027817

Contents

1. Introduction

The Covid-19 pandemic, caused by the Severe acute respiratory syndrome Coronavirus-2 (SARS-COV-2), was identified for the first time in the Chinese city of Wuhan in December 2019 and soon developed into a global pandemic. The effects of the pandemic are not only observable by looking at the growth cases and the new barriers between countries, but also through particular behaviours on social media (i.e. negative sentiments and Asian hates). While they trigger massive media attention and are generally a front-page topic, media outlets keep those who may not be directly impacted up to date and aware of the current circumstances [1]. Social media, the platform that offers the opportunity to the public to participate and interact with others, discussing every social issue, and expressing the individual opinion, also it is an access that the mainstream media can lead, conducting the determined public opinion and government propaganda. So it is a vital resource for investigating the current circumstance.

The unique characteristics of social media, such as diversification of information, liberalisation of expression, and efficient transmission speed, had facilitated the propagation of a large number of rumours related to the spreading and blocking of the Covid-19 epidemic [2]. However, according to the geographical social distancing and barriers, the news channels and the official news account on social media (Twitter) have reported the pandemic situation in China with different points of view when the disease continues to spread in China and all over the world. It evokes much controversy, verbal abuse and marginalisation to Chinese and other Asian ethnicities. Beyond the agendas of the news media, we are witnessing a battle in which the current public health emergency is bringing latent political and economic conflicts to the surface on social media [3].

In the current research, we focus on Chinese official media (CGTN, China Xinhua News) and official media outside of China (BBC News World, CNN, The Associated Press, Reuters, Al Jazeera English) on Twitter, collecting and analysing periodic tweets about Covid-19 by two methods, **Community Detection and Sentiment Analysis**. Considering the distinction between the media inside and outside of China, we proposed the following hypothesis: the topic related to Covid-19 situation in China which is reported by the official news media inside and outside China are different and changes when the situation has changed; and also proposed the following research question: **Identifying the different approaches of official news on Twitter inside and outside of China; What topic, theme and keywords were related to the Covid-19 situation in China at the beginning and how the topics are changed when the situation has changed.**

1.1. Literature review

There are already some papers talking about Covid-related tweets related to China. Most of them focus on the discussion of Covid-19 by online social media users. They focus on different social media, most of them focus on the discussion on Twitter, but there is also a small amount of paper focused on Weibo (a platform like Twitter but used in China). Ammar, Budhwani & Sun and Rodrigues analysed tweets on Twitter. On the other hand, Da & Yang and Wang & Qian analyse tweets on Weibo.

Scholars who have studied Twitter have found that tweets about China and Covid-19 from Twitter users are relatively negative. They referred to "Covid-19" as "Chinese virus" and "China virus", primarily because of the tweets tweeted by US President Trump. This phenomenon has led to the stigmatisation of Covid-19 on Twitter and more negative sentiments from users. Furthermore, this has also led to increased prejudice and racism against Chinese people among some Twitter users.

Ammar analysed the public response that Trump referred Covid-19 to the Chinese virus. They coded 50 tweets and got eight themes, and they are pretty negative such as endangerment, stigmatisation, xenophobia. Most of them are related to China and Chinese people. They suggest that the prevalent theme addressed opposition to racism towards people of Chinese ethnicity [1].

Budhwani and Sun also focus on this aspect in their paper. Their paper tries to find out "if the prevalence and frequency of "Chinese virus" and "China virus" have increased after the US president mentioned these terms." They downloaded tweets from 50 states that contained the keywords and analysed them. However, they suggest that the Covid-19 stigma might continue on Twitter, which might hurt the Chinese communities in the US [4].

Rodrigues investigated the topics related to China with Covid-19 in Portuguese tweets. They retrieved 1.6 million tweets from March 19 to April 3. They found out that the most common themes are "Chinese virus" and "virus from China". They also suggest that this content shows the users' negative sentiments (like anger, sadness, and fear), showing the characteristics of political polarisation in Brazil and how users understand the pandemic. They suggest that Twitter may become the platform to show conflict, to spread scientific scepticism, stigma and racism against Chinese people [3].

However, scholars studying Weibo have focused on the emotions and behaviours of people affected by Covid-19. The severity of the Covid-19 condition influences people's emotions. They have also looked at the relationship between Covid-19 and the Echo Chamber Effect.

The article of Da & Yang analysed the tweets from Weibo and associated the emotions about Covid-19 expressed on social media are highly relevant with the severity

of the pandemic [5].

On the other hand, Wang & Qian, in their paper, analyse the echo chamber effect of users' response to rumours of Covid-19 on Weibo. They found out that the "retweeting system played an essential role in promoting polarisation and the commenting system played a role in consensus building. Furthermore, there might not be a significant echo chamber effect on community interaction." [6].

The above research provides many individual social media users' perceptions of Covid-19 and China. As seen from the above, individual users' views and discussions on this topic are primarily negative. However, they mainly focus on the tweets of personal users, and none of them studies how the official news account reports Covid-19 and China. So, there is a gap here for us to work on. Studying official news accounts may provide different results since they cannot express too many personal emotions. These official news accounts allow us to look at the other side of the discussion on Twitter about Covid-19 and China and study the tweets sent by Chinese media and those from outside China, which also show the same and different attitudes towards the same event in China and abroad.

1.2. Background

In this project we examine different approaches by several official News regarding coverage of the information about the pandemic and Covid-19 throughout different periods, their different impressions, views and also accusation points. We already know that some news and information have their origin from media supported by governments, it could be different governments such as China, Russia, USA etc.

There is a crisis going on in the public understanding, different social media platforms and western digital corporations such as Facebook, Twitter and Instagram along with their Chinese identical platforms such as WeChat are playing an important role in this regard and damage the public understanding by the false information about the Covid-19 within them. This pile of misleading information and false rumours could increase the public confusion along with promoting digital forms of racism and they bring people to this uncertainty about what is the right information and what is not, which source should they trust and which one not.

There was a popular conspiracy theory that the virus was developed as a tool to undertake a biological war against China, on the other hand inside China there were other rumours spreading the bioweapons research in a laboratory located in Wuhan which later on resulted in the genetic engineering of Covid-19 that was then released [7].

Generally, state-supported media used by a variety of governments such as the ones we mentioned already in order to influence foreign public opinion about a specific phenomena, states use it as a tool to have a control over the public. Other uses of state-supported media belong to counterbalancing dominant western media reports or carrying on top down propaganda operations. For example it was in 2005 that Russia released the English language Russia-Today (which later rebranded as RT) with a few operations. Over time this channel has developed an anti western narrative and step by step became a propaganda weapon of the Kremlin. At the same time RT has expanded its capacity aiming for publishing its content in other languages such as Arabic, French, German and Spanish. In fact through this channel the Russian government aimed at targeting other non-English speaking societies for different usages, either to attract more support for the state in places where anti-Western societies has strong resonance and influence or on the other for promoting a positive image of Russia in other states and regions for establishing stronger relationships [8]. This was only an example of Russia but in addition there are also a number of Chinese state-backed media targeting other societies throughout the world to support China's soft power influence over other states and its ambitions. As an example CGTN espanol channel, launched by China in order to target Spanish-speaking people [9].

This is the strategy in which governments try to impact the public by the media which they do support, and the reason for doing this project is regarding identifying different attitudes of official media and news belonging to China and U.S and throughout Twitter by their tweets and hashtags and to analyse our collected data from different period of the pandemic since the beginning of the Covid-19.

1.3. Methodology

"It's amazing that the amount of news that happens in the world every day always just exactly fits the newspaper." (Jerry Seinfeld, American actor and comedian, b. 1954) When news come to public, they have been chosen and produced by journalists, editors and the manager of the media and so on, based on their journalistic standards, positions and media interests. We all know that media reports are mirrors of the real world, but they deviate from the real world. The news media has an ability to influence the importance placed on the topics of the public agenda, which is called agenda-setting, driven by the media's preference and bias on things such as politics, economy and culture, etc.

Nowadays, media is no longer limited to newspapers, and the Internet has become the most efficient and convenient way of news dissemination. However, the "gatekeeping", which first instituted by Kurt Lewin (1890-1947, German-American psychologist) in 1943, is always taking its roles, no matter in traditional or internet media. During the global pandemic, we are full of curiosity about the world shaped by Chinese and foreign media: what are the differences between their reports, what impact did their reporting angles have on the public, and whether they intensified the

	Inside of China	Outside of China
TV Broadcast	CGTN *(13million+ followers)*	BBC News (World) *(34million+ followers)*, CNN *(56million+ followers)*, Al Jazeera English *(7million+ followers)*
News Agency	China Xinhua News *(12million+ followers)*	The Associated Press *(15million+ followers)*, Reuters *(24million+ followers)*

Table 1: Twitter accounts inside and outside China, with their respective number of followers.

contradiction between the East and the West?

In order to explore these questions, we choose the seven most influential news Twitter accounts inside and outside of China to take quantitative analysis. As listed in Table 1, these accounts are from TV Broadcast and News Agency, and the numbers of followers are as of February 15, 2022.

By collecting data on these accounts about keywords Wuhan, Coronavirus, Covid and vaccine, we want to compare two aspects between these inside and outside of China accounts : **what they choose to report** and **how they report the same things**.

In order to make the data accurate, we collected the data in three different time periods. The first period is from Jan 23, 2020 to Feb 23, 2020, during one month after Wuhan's lockdown, which is an important period at the Covid-19 beginning with a large number of reports from China and foreign. The second period is from Sep 20, 2020 to Oct 21, 2020, when the Chinese economy recovered, and the global epidemic worsened. The third one is from Mar 17, 2021 to Apr 17, 2021, when WHO released the traceability report and there was much news about it from different angles.

After collecting all the data, we compare and analyse their differences and connection by building semantic networks and community detection. According to the numbers of reposts and likes, we also know which perspectives of news report is more influential and acceptable.

1.3.1 Python libraries

In order to start collecting, processing and cleaning data towards building network nodes and edges, we needed a powerful programming language to handle and maintain the complexity and the enormous amount of the existing data. For this reason, we chose to write Python scripts on Google Colab/Jupyter Notebook using the following important libraries:

- The OS module: provides functions for interacting with the operating system.
- Pandas: is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series;
- Itertools: functions creating iterators for efficient looping;
- NetworkX: for studying graphs and networks;
- NumPy: for adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Matplotlib: a plotting library and numerical mathematics extension of NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.
- igraph: a library collection for creating and manipulating graphs and analyzing networks.
- Cairocffi: a set of Python bindings and object-oriented API for Cairo which is a 2D vector graphics library with support for multiple backends including image buffers, PNG, PDF, and SVG file output.

Moreover, and in order to better visualize and analyze networks, especially the large ones, according to different approaches, i.e, community detection, a specific layout algorithm, nodes and edges parameters together with the related numerical values, we used Gephi, the open-source network analysis and visualization software.

{'full_text': "German Chancellor Angela Merkel, 66, on Friday received her first dose of AstraZeneca COVID-19 vaccine, she announced via government spokesman. (People's Daily) https://t.co/ofSU559gVF",
'display_text_range': [0, 160], 'entities': {'hashtags': [], 'urls': [], 'user_mentions': [], 'symbols': [], 'media': [{'id': 1383128810334400516, 'id_str': '1383128810334400516', 'indices': [161, 184], 'media_url':
'http://pbs.twimg.com/media/EzHcUyqXMAQ8fBj.jpg', 'media_url_https': 'https://pbs.twimg.com/media/EzHcUyqXMAQ8fBj.jpg', 'url': 'https://t.co/ofSU559gVF', 'display_url': 'pic.twitter.com/ofSU559gVF',
'expanded_url': 'https://twitter.com/CGTNOfficial/status/1383128868874297344/photo/1', 'type': 'photo', 'sizes': {'medium': {'w': 1200, 'h': 675, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w':
1963, 'h': 1104, 'resize': 'fit'}, 'small': {'w': 680, 'h': 382, 'resize': 'fit'}}}]}, 'extended_entities': {'media': [{'id': 1383128810334400516, 'id_str': '1383128810334400516', 'indices': [161, 184], 'media_url':
'http://pbs.twimg.com/media/EzHcUyqXMAQ8fBj.jpg', 'media_url_https': 'https://pbs.twimg.com/media/EzHcUyqXMAQ8fBj.jpg', 'url': 'https://t.co/ofSU559gVF', 'display_url': 'pic.twitter.com/ofSU559gVF',
'expanded_url': 'https://twitter.com/CGTNOfficial/status/1383128868874297344/photo/1', 'type': 'photo', 'sizes': {'medium': {'w': 1200, 'h': 675, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w':
1963, 'h': 1104, 'resize': 'fit'}, 'small': {'w': 680, 'h': 382, 'resize': 'fit'}}}]}}

Figure 1: Example of the "extended tweet" field

2. Dataset

2.1. Data collection

As a first step, we had to decide from where we have to collect data, and we chose Twitter as data source, exploiting the options provided by the Twitter developer accounts, even though there were some limitations in terms of the maximum number of possible data that can be obtained monthly. Hence, divided data collection according to:

1. Definite crucial periods:
    - January-February 2020 : in this period, the world encounters the first case of Covid-19;
    - September-October 2020 : in this period, the number of mortality related to Covid-19 exceeded one million people worldwide;
    - March-April 2021 : in this period, the international concerns were about new variants of the pandemic, together with different types of vaccinations that can be taken in order to stop the spread;
2. Media Official Accounts Inside China: China Global Television Network (CGTNOfficial) and China Xinhua News (XHNews);
3. Media Official Accounts Outside China (mainly USA media): Al Jazeera (AJEnglish), The Associated Press (AP), BBC (BBCWorld), CNN (CNN), Reuters (Reuters).

Furthermore, data was collected according to some other criteria, i.e, English language and specific keywords that were : Wuhan, China, Coronavirus, Covid, Vaccine which are the most likely used ones to identify the tweets that will be interesting for building our networks. After the collection of the tweets for all the selected periods and Twitter accounts, we proceeded with the data aggregation, according to:

1. Media accounts and periods, so that we will be able to build networks in order to spot differences how media are spreading news inside a country, and also compare between media interests in the different two countries while fixing a period, i.e, we will have networks based on a given fixed period including tweets from different media accounts.
2. Countries, so that we will be able to build networks regardless the period, i.e, we will have networks based on a given country (either inside or outside China) that may contain tweets from different periods.
3. Countries and periods, so that we will be able to build networks that give us information about media interests in both groups (inside and outside China), i.e, we will have networks based on a fixed country and a fixed period or networks based on both countries and a fixed period.

We ended up with 3 source-based distinctions (inside China data, outside China data, and inside and outside China data grouped together) and 4 period-based distinctions (January-February 2020, September-October 2020, March-April 2021, and all the three periods grouped together) for a total of 12 datasets, whose dimension is reported in Table 2.

	USA	**China**	**All**
January-February 2020	7518	4875	12393
September-October 2020	8388	3097	11485
March-April 2021	3956	1375	5331
All periods	19862	9347	29209

Table 2: Number of tweets in each dataset.

2.2. Text extraction

Once we collected the tweet and saved them as `.csv` file, for each tweets we had not only its text but also many other information like time, account, id, etc. Since we needed only the text of the tweets, we inspected the "text" field of the data. Still a lot of tweets were truncated with ellipsis. This was because the full text of the tweets longer than 140 characters was reported in the "extended tweet" field (an example is reported in Figure 1). Unfortunately, in this field there are also many other information as can be seen figure. Since we started our work with `.csv` file we couldn't use the dictionary structure of the data like in the

`.pkl` files. Therefore in this case we had to use a regular expression to extract the correct text.
Still the text of the retweets was missing. This was because their text was in a different field, the "retweeted status", therefore once again to retrieve the correct text among other information we used a regular expression.
Finally, we put together all the extracted texts in a list to continue the analysis.

2.3. Cleaning

To continue with our research, it was necessary to carry out text cleaning to remove all the irrelevant parts of the tweets and focus the analysis only on the words that were important for our purpose. Our project involves the construction of a word-context network, where nodes are words and edges are links connecting two words that appear together in the same tweet. Before proceeding with the construction of the network, the following cleaning was performed on the text of the tweets:

- Removal of mentions (`@`) and hashtags (`#`);
- removal of http links, frequently found in tweets and linking to the websites of various news accounts;
- removal of punctuation;
- removal of stopwords;
- tokenization and POS tagging;
- lemmatisation.

The tokenization, POS tagging and lemmatization procedures were performed using NLTK (Natural Language Toolkit), a suite of libraries for natural language processing for English in Python. Only words belonging to the English language corpus and some proper names not present in the dictionary used were preserved. In fact, some proper nouns that frequently appeared in the collected tweets (e.g. Covid19, Coronavirus, Xinjiang, Xi Jinping) were discarded as they were not recognised by the NLTK dictionary as belonging to the English language, although they were very relevant for this research. After the text cleaning phase, it was possible to build the network from the cleaned tweets texts.

3. Exploratory Analysis

In order to try to get some insights about the collected data, we decided to perform an initial exploratory analysis. This section is entirely dedicated to the presentation of the methods we used in order to extract information from the "China" and "outside China" tweets.

In particular, we did some exploratory analysis on the network using the main centrality measures.

3.1. Network construction

After having cleaned all the tweets as explained in Section 2.3, we extracted a dictionary d containing all the words written in the tweets.
This procedure allowed us to represent tweet i with a feature vector $x^{(i)}$, for which:

$$x_j^{(i)} = \sum_{w \in \text{ tweet } i} \mathbb{1}_{\{w=j\}}$$

In other words, entry j of feature vector i is the frequency of word j in tweet i. Specularly, we have a feature representation of each word j in our diction given by the vector $(x_j^{(i)})_{i=1,..,n_{\text{tweets}}}$.
Now that we had a feature vector for each tweet, we decided to represent connections between words using co-occurrencies in tweets. By doing that, the adjacency matrix is represented as:

$$A_{ij} = \frac{1}{n_{\text{tweets}}} \sum_{t \in \text{ tweets}} \mathbb{1}_{\{i \in t, j \in t\}}$$

It worths noticing that by using a binary version of the features ($\{0,1\}$ valued, word either present in the tweet or not), the adjacency matrix can be represented as the scalar product between the feature representations:

$$A_{ij} = \frac{1}{n_{\text{tweets}}} \sum_{k=1}^{n_{\text{tweets}}} x_i^{(k)} x_j^{(k)} = \frac{1}{n} x_i^{(\bullet)} \cdot x_j^{(\bullet)}$$

This last formulation also reminds that we can use the created features to represent even higher order interactions and to keep an hypergraph structure instead of just the graph one (keeping higher order co-occurrencies), which can be an interesting future improvement.

3.2. Keywords importance

The first thing we decided to do after cleaning the tweets was to check some different measures of "importance", in order to rank the words in the extracted dictionary. We did this in some different ways, the majority of them by exploiting the graphical structure created in Section 3.1.

3.2.1 Frequency

The simplest way to evaluate some kind of importance in the words dictionary is to look at the marginal frequencies

$$f_i = \sum_{k=1}^{n_{\text{tweets}}} x_i^{(k)}$$

The problem of this kind of measure is that it overfocuses on really common words by not relevant in this particular context.
In Figure 2 we plotted the words with the highest frequency in the tweets from and outside China. Apart from the difference in the most frequent word, in the China plot we can notice a shorter tail: this indicates a bigger variety of common words outside China (even though some of them are especially related to USA, like president and Trump).

3.2.2 TF-IDF

In order to solve the problem of common use words, we tried TF-IDF as a measure of importance.
The main two ingredients to construct it are the **term frequency**(TF) and the **inverse document frequency**(IDF):

- the TF can be constructed in different ways, but the standard one is to use the estimated probability that the word i appears in tweet t, as shown in the following

$$\text{tf}(i|t) = \widehat{P}(i|t) = \frac{x_i^{(t)}}{\sum_i x_i^{(t)}}$$

- the IDF is instead the negative log-likelihood relative to the frequency of a term in the whole set of collected tweets T, as shown in the following:

$$\text{idf}(i|T) = -\log\Big(\widehat{P}(i|T)\Big) = -\log\Big(\frac{|\{t \in T | i \in t\}|}{|T|}\Big)$$

The final TF-IDF measure is obtained as the product of the two:

$$\text{tfidf}(i|t,T) = \text{tf}(i|t)\text{idf}(i|T)$$

In particular, $\text{tfidf}(i|t,T)$ is big if $\widehat{P}(i|T)$ is small and if $\widehat{P}(i|t)$ is big. The rationale behind this measure is that we want to penalize words that are common among all the collected tweets, trying in this way to filter out common and out of context terms.
In Figure 3 we plotted the word with the highest TF-IDF value over all tweets in the China and outside China samples. The initial decreasing rate of the two curves look similar, and some common terms had been filtered out. As a future improvement it would be interesting to use also other version of TF-IDF to try to filter out other common use words.

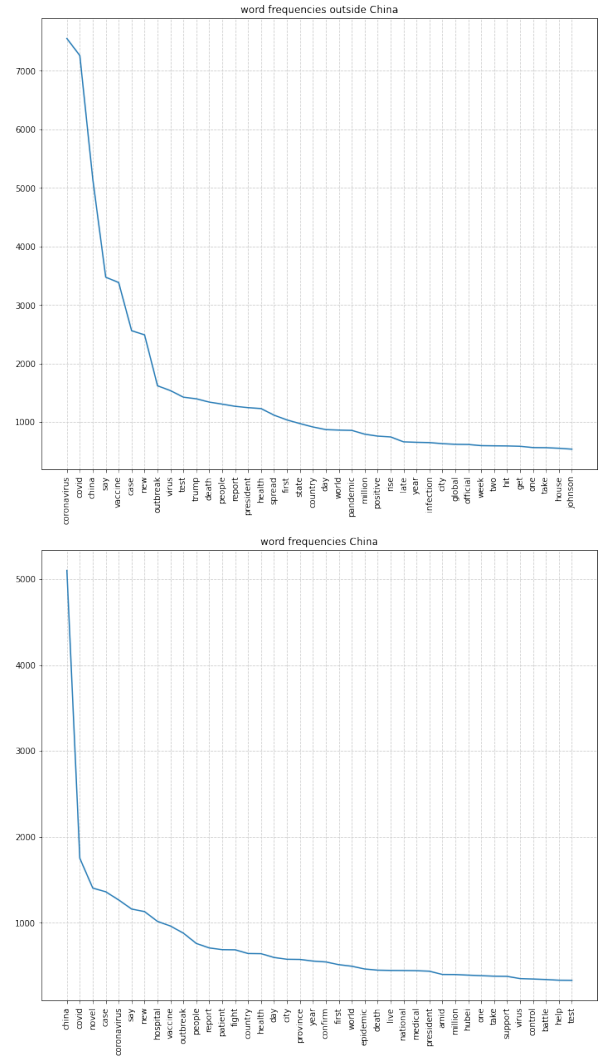


Figure 2: Top frequency values for words inside and outside China.

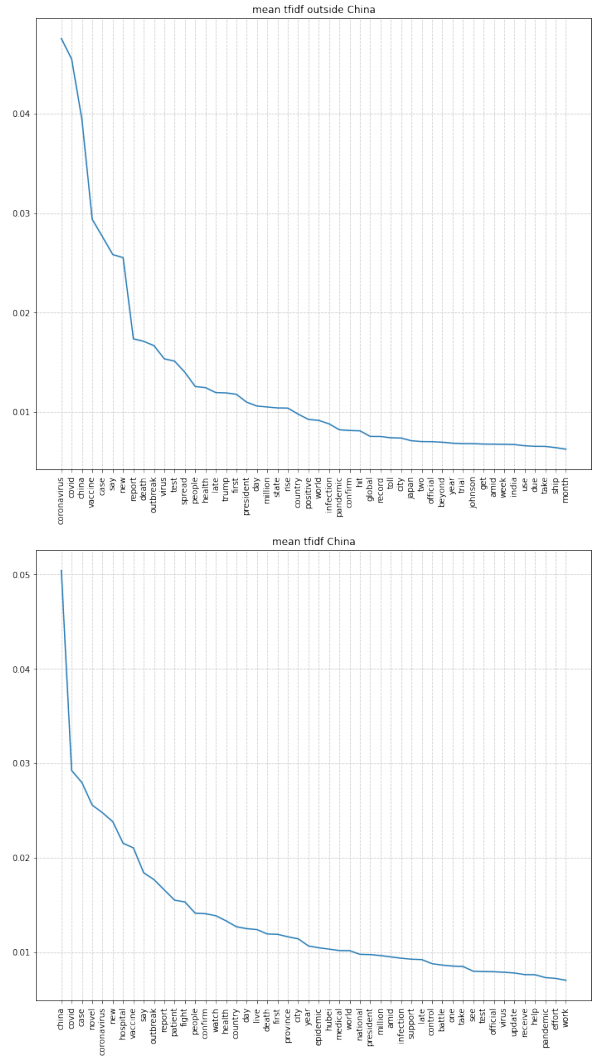


Figure 3: Top TF-IDF values (averaged over documents) of words in China and outside China.

3.3. Centrality measures

After having focused on "non graphical" importance measures as a preliminary, we decided to try some different centralities.
In particular, we decided to use:

- Degree, in order to capture how much a word is "local" in the graph;
- Pagerank;
- Betweenness, to look for "bridge words";
- Closeness

Before visualizating the centralities and the most important nodes according to them in 3.4, we decided to briefly remind about each one of them a bit more in detail.

3.3.1 Degree

The (out) degree of a node i in a graph with (possibly weighted) adjacency matrix A is defined as:

$$d_i = \sum_j A_{ij}$$

Since our graphs are all undirected, d_i is the number of edges connected to node i.
In figure 4 we can observe the log-log plot of the degree distribution of the two networks: the behaviour is almost equal in the two graphs, except for a small shift to the left of the China mean degree. It worths also noticing that for big enough degrees the two distributions follow a power law pattern.

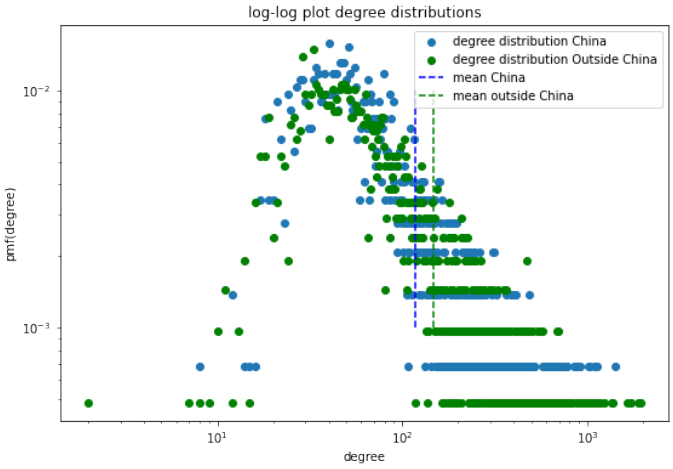


Figure 4: Log-log plots of the word degree distributions inside and outside China.

3.3.2 Pagerank

Pagerank is one of the simplest examples of eigenvector centrality.
The main idea is to consider a Markovian jump-diffusive walk $(X_t)_t$ on the graph, with transition matrix:

$$P_{ij} := P(X_{t+1} = j | X_t = i) = \alpha \frac{A_{ij}}{\sum_j A_{ij}} + (1 - \alpha) q_j$$

where q_j is the probability of jumping to node j.
The Pagerank vector is defined as the stationary distribution π of this chain (often the jump part is required to ensure existence and uniqueness of it), so as a solution of the eigenvalue equation:

$$P^T \pi = \pi$$

A high Pagerank value for a node means that in the long time regime, it is more probable to observe the chain transition there.
In figure 5 we plotted the Pagerank distributions in the two graphs. As before the behaviour is similar, except for a longer tail in the outside China graph. In 3.4 we also included the values of the highest pagerank words.

3.3.3 Betweenness

Betweenness centrality exploits the idea of geodesics in the graph to construct a measure of importance. Intuitively, a node has a high betweenness if it's often present when walking the shortest path between two nodes.
More formally, the betweenness can be defined as:

$$\text{bet}_i = \frac{1}{(N-1)(N-2)} \sum_{v \neq w \neq i} P(i \in \Gamma_{vw})$$

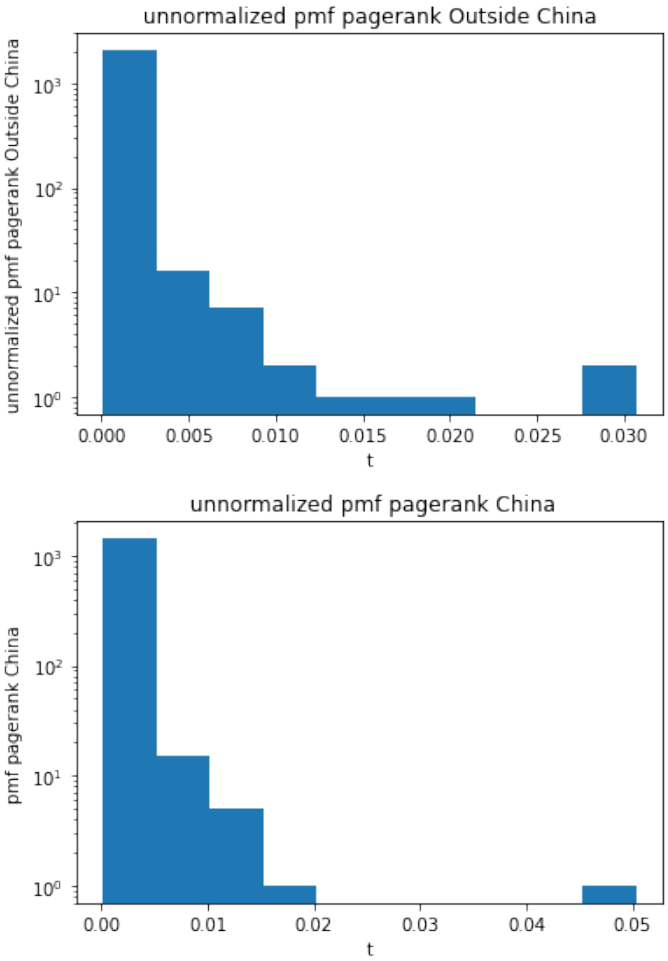


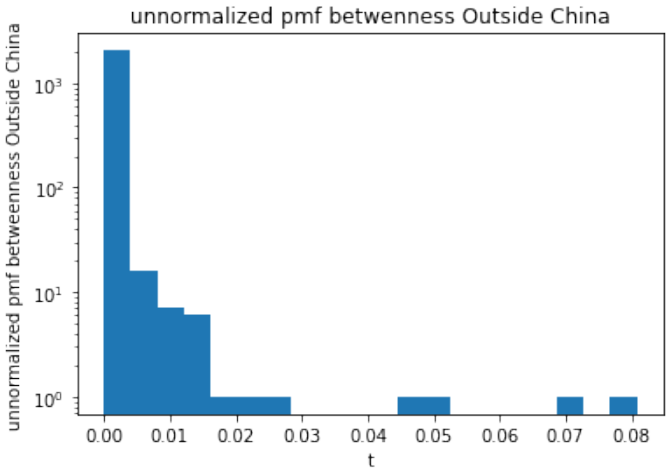
Figure 5: Pagerank distributions inside and outside China,log scale.

where Γ_{vw} is the set of geodesics connecting nodes v and w in the graph. This can be interpreted as the probability that i belongs to a geodesic passing through two nodes, averaged over all the couples of nodes.
In figure 6 we plotted the betweenness distributions in the two graphs. As in the pagerank case, the behaviour in the two graphs is similar . In 3.4 we also included the values of the highest betweenness words.

3.3.4 Closeness

As betweenness centrality, also closeness centrality exploits the idea of graph geodesics to construct a nodal importance measure. The geodesic metric d in a graph with adjacency matrix A is defined as:

$$d(x, y) = \inf\{l \geq 0 | A^l_{xy} > 0\}$$



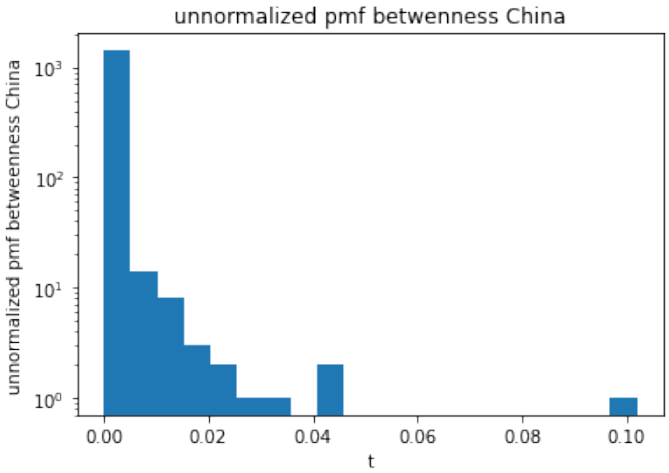
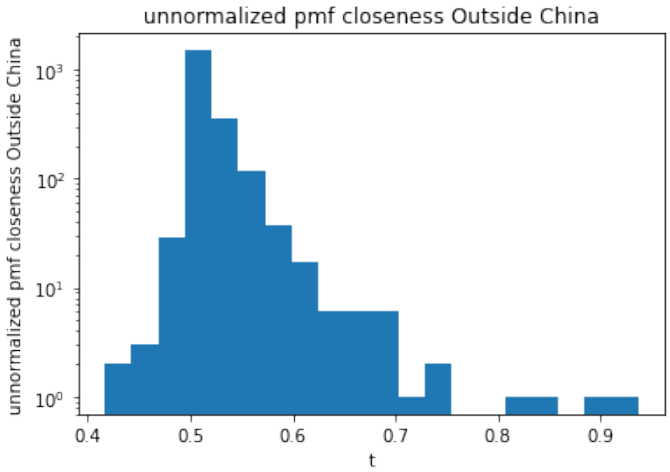


Figure 6: Betweenness distributions inside and outside China, log scale.



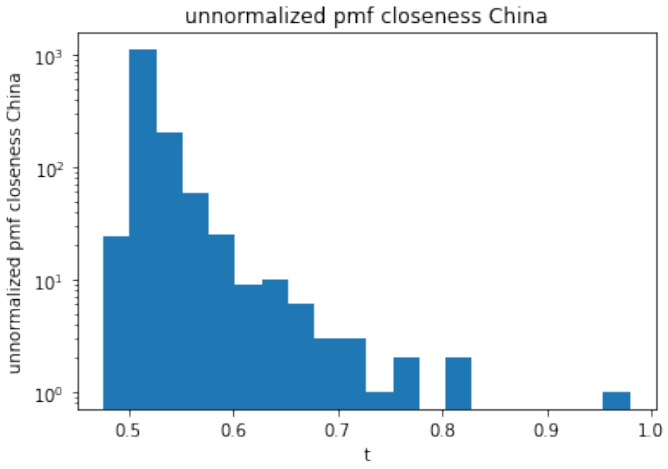


Figure 7: Closeness distributions inside and outside China, log scale.

This last definition of metric is measuring the smallest number of steps needed in order to reach y from x moving through edges of the graph.
The closeness is defined as the inverse of the average distance between a node and all the others:

$$\mathrm{cl}_i = \frac{1}{\frac{1}{N}\sum_j d(i,j)} = \frac{1}{\mathbb{E}_{j\sim \mathrm{Unif}(G)}[d(i,j)]}$$

An high centrality value means that the node in question is on average near to all other nodes in the graph.

In figure 7 we plotted the closeness distributions in the two graphs. The behaviour of the distribution here is a bit different in the two cases: the China distribution is a bit more shifted to the right and shows less of a "gaussian behaviour" in log scale. This can be read by saying that by comparing the two graphs, in the outside China one the most common closeness value is represented by the mean better than in the other one. In 3.4 we also included the values of the highest closeness words.

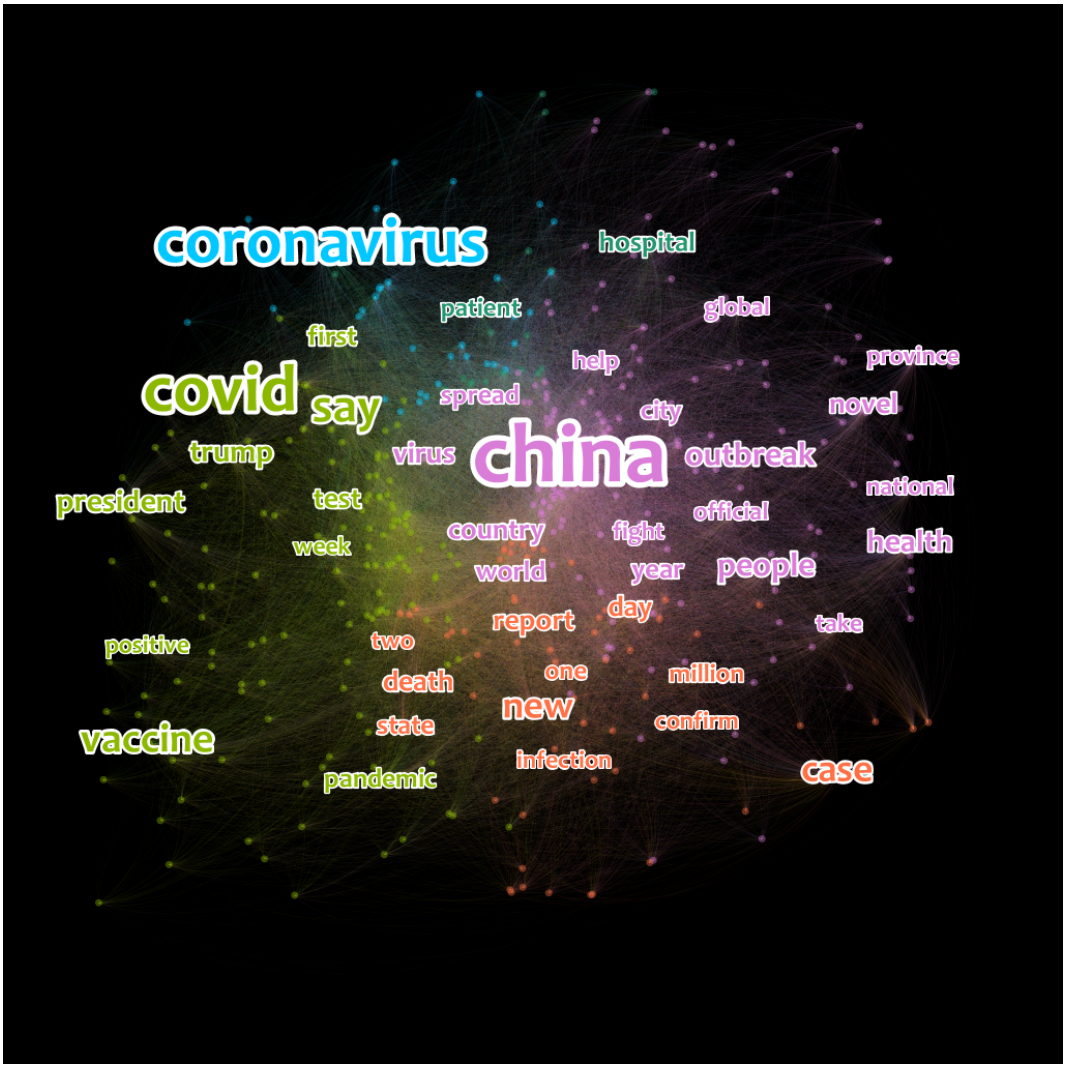


Figure 8: The whole network obtained from China and outside China tweets.

3.4. Network visualization

We started the network visualization task plotting the full network: the result, obtained using the Gephi tool, can be seen in Figure 8. This graph was made by analyzing all the more than 29000 tweets we have collected (therefore for this network we considered together the China and outside China tweets).
In the first place, in order to build the network, we extracted and cleaned the text of the tweets, as explained in the previous sections. Then we created a dictionary with all the different words present in the tweets, counting how many each one was repeated. This dictionary will represent the node list of the network. Then in order to build the edges, we considered connected two words if they appeared in the same tweet. The more frequently two words appeared together, the stronger we considered their link in the graph. Therefore we have built an undirected and weighted graph. Since we collected a huge amount of data we decided to discard words that were repeated less than ten times. Still we were able to have a graph with 2717 nodes (the words) and 60760 edges (the links between the words).
Once we imported the lists of nodes and edges in gephi, we ran the Fruchterman and Reingold algorithm for a better visualization. Then we ran PageRank and community detection algorithms. The higher is the PageRank value the bigger is the word. Observing Figure 8, we are already able to see that there are few communities, to be precise five, and three big hubs: “Coronavirus”, “Covid’” and “China”.
In order to get more insight on the network, we analyzed separately the China and outside China data and then compared them using some centrality measures.

3.4.1 PageRank and communities

We started studying the importance of keywords in the China network, therefore using the 9350 tweets collected from the inside China accounts. The network was build in the same way as is Section 3.4. Then we imported the network on gephi and we ran the PageRank algorithm, with dumping factor equal to 0.85, and the community detection algorithm. Finally, for a better visualization we reduce the number of nodes from 1423 to 500, discarding the ones with lower PageRank values. We can see the result in Figure 9.

We can see there is one big hub, "China", and many smaller ones. We can find four different communties:

- the purple community with words like: "China", "country", "Beijing", "Xinhua", "Jinping". Probably more related to what happens inside China.
- the blue community with words like: "case", "report", "death", "asymptomatic", "pneumonia". Probably more related to Covid-19 reports and symptoms.
- the green community with words like: "covid", "vaccine", "health", "dose", "develop", "vaccinate", "johnson". Probably more related to the vaccines.
- the black community with words like: "hospital", "patient", "medical", "doctor", "medicine". Probably more related to the patient care and the hospitals' situation.

We repeated the same analysis for the outside China network, which is bigger then the China network since was built from 19865 tweets. Using these tweets and discarding the words repeated less then ten times we get a network of 2044 nodes and 40570 edges. Still for a better visualization we reduced the number of nodes to approximately 500. We can see the result in Figure 10.

We can see different big hubs: "covid", "coronavirus", "China", "vaccine". Differently from the China network, the words "china" and "coronavirus" are now more important (in terms of Pagerank values) and strictly related. As in the China network, we can find four different communities:

- the green community with words like: "China", "outbreak", "virus", "country", "rise", "global", "epicenter", "Hubei", "globally", "market", "million". Probably more related to the spreading of the Covid-19 initially in China and then in the whole world.
- the blue community with words like: "Trump", "president", "white" and "house", "election", "debate", "congress","republican", "democratic". Probably more related to the Covid-19 debate in the USA.
- the purple community with words like: "covid", "vaccine", "dose", "research", "trial", "vaccination","johnson", "develop". Probably more related to the vaccines and the scientific research to cure the coronavirus disease.
- the black community with words like: "quarantine", "cruise", "japanese", "stay" and "home", "airport. Probably more related to the some mixed news from all over the world.

Therefore, we can see how each network has one community focusing on the China or USA situation, then the other communities are more or less similar and focus on what happens around the world related to the vaccine or the hospital's care or the spreading of the virus.
In particular we can see how in the outside China network, there is a dominant position of the USA internal debate that involves the president and the two main parties.
Finally, in Table 3 and in Table 4, we report the first ten higher values in terms of PageRank for each network. We can see how they are pretty similar, but once again the political aspect is more relevant in the outside China network.

Word	PageRank
china	0.101007
covid	0.024933
say	0.020023
novel	0.019632
case	0.018584
coronavirus	0.015258
new	0.014843
hospital	0.013752
vaccine	0.012462
outbreak	0.011297

Table 3: Ten highest PageRank values in the China network.

Word	PageRank
coronavirus	0.049211
covid	0.047017
china	0.031058
say	0.025504
vaccine	0.020716
new	0.014366
case	0.01369
trump	0.011573
president	0.009626
outbreak	0.0096

Table 4: Ten highest PageRank values in the outside China network.

Figure 9: Community detection and PageRank on the China network.

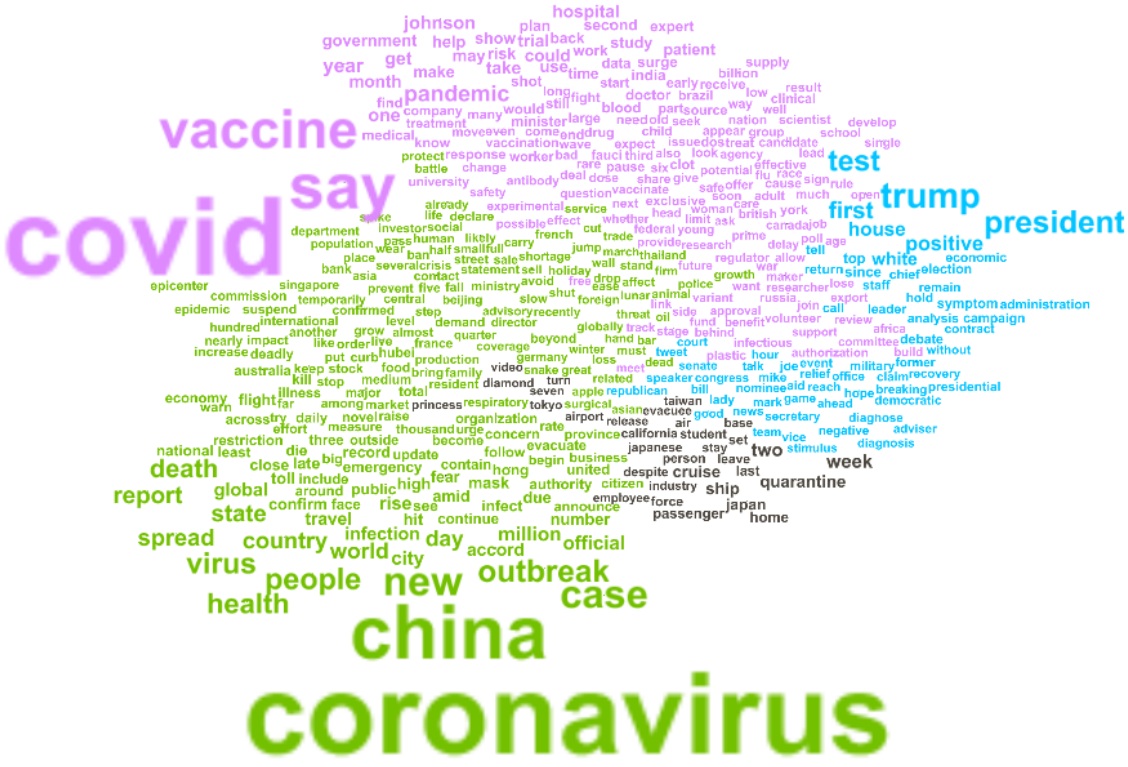


Figure 10: Community detection and PageRank on the outside China network.

3.4.2 Betweenness centrality

Considering the same networks as in Section 3.4.1, we calculated the betweenness centrality. The betweenness centrality measures for each nodes how many shortest path pass through this node. We can see the result in Figure 11. The greener is a word the higher is its betweenness value.
As one may expect, both in the China and outside China network the nodes with highest betweenness values are the big hubs, since they directly connect a huge quantity of nodes.
Finally, in Table 5 and 6, we report the first ten higher values in terms of betweenness centrality for each network.
We can see how approximately the nodes with the highest betweenness values are also the ones with highest PageRank values.

Word	Betweenness centrality
china	579256.389738
covid	79457.631877
say	52440.549191
novel	32944.175106
vaccine	22751.325828
coronavirus	19597.301179
new	18657.494121
people	17947.631131
case	15803.571954
outbreak	14470.042173

Table 5: Ten highest betweenness centrality values in the China network.

Word	Betweenness centrality
coronavirus	482181.309425
covid	449335.27516
china	223353.918017
say	165026.420144
vaccine	97556.094534
new	51689.27506
trump	43865.897278
people	30030.633603
president	29517.486102
case	28583.67608

Table 6: Ten highest betweenness centrality values in the outside China network.

3.4.3 Closeness centrality

Considering the same networks as in Section 3.4.1, we calculated the closeness centrality. We used the closeness centrality, and not the harmonic centrality, since the graph is strongly connected.
Closeness measures how far is one node to the others through the formula:

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

where $d(y,x)$ is the distance between vertices x and y.
We can see the result in Figure 12. The more red is a word the higher is its closeness value.
Finally, in Table 7 and 8, we report the first ten higher values in terms of closeness centrality for each network. Once again we find that approximately the nodes with the highest closeness values are also the ones with highest PageRank values.

Word	Closness centrality
china	0.923977
covid	0.649315
say	0.636811
novel	0.615584
coronavirus	0.596977
new	0.583504
people	0.583026
outbreak	0.579226
case	0.571314
vaccine	0.566083

Table 7: Ten highest closeness centrality values in the China network.

Word	Closness centrality
coronavirus	0.814784
covid	0.795828
china	0.70255
say	0.702301
vaccine	0.641034
new	0.615003
trump	0.596154
case	0.592947
virus	0.589074
president	0.5882

Table 8: Ten highest closeness centrality values in the outside China network.

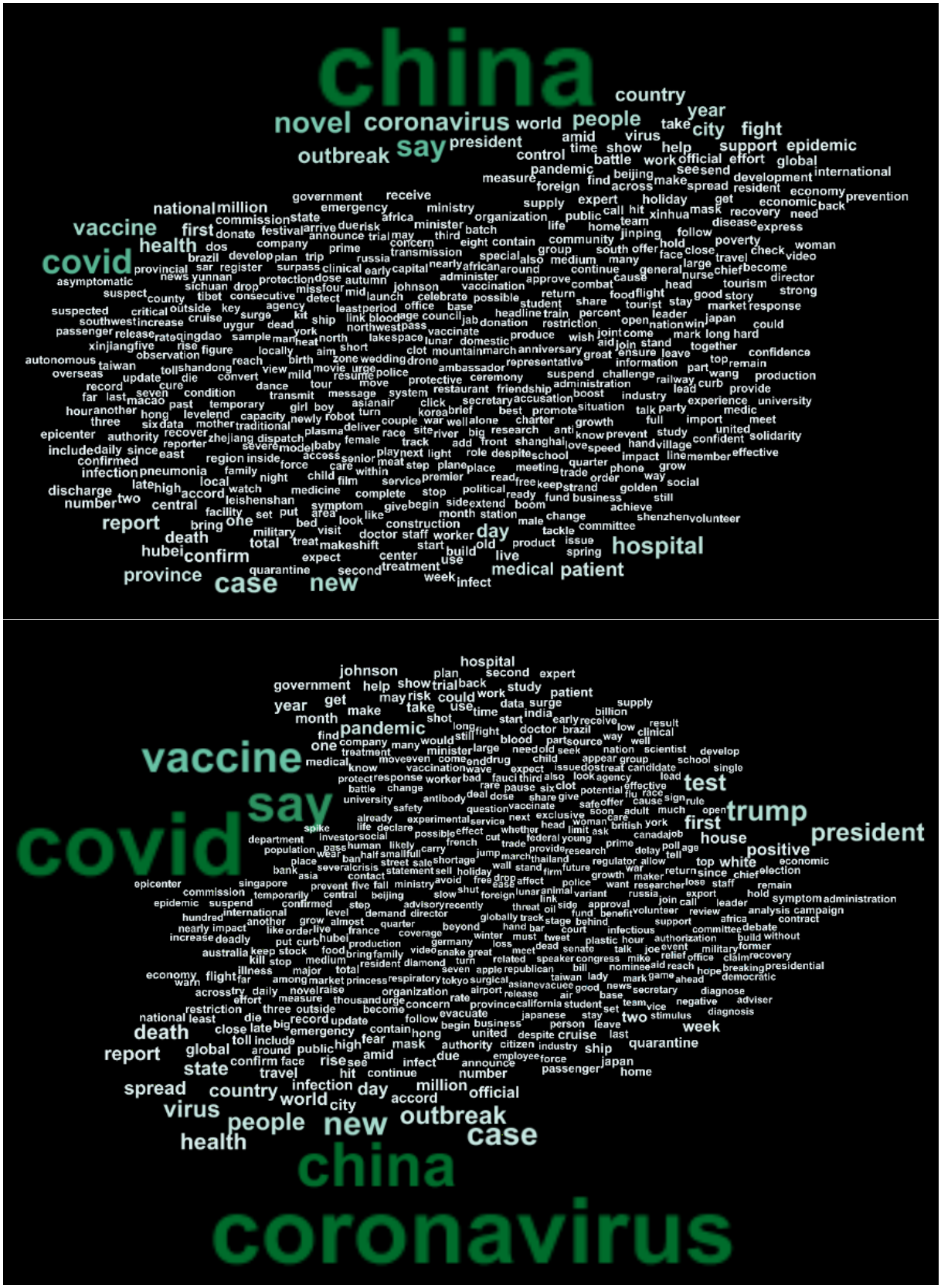


Figure 11: Betweenness centrality visualization. On top the China network, on the bottom the outside China one.

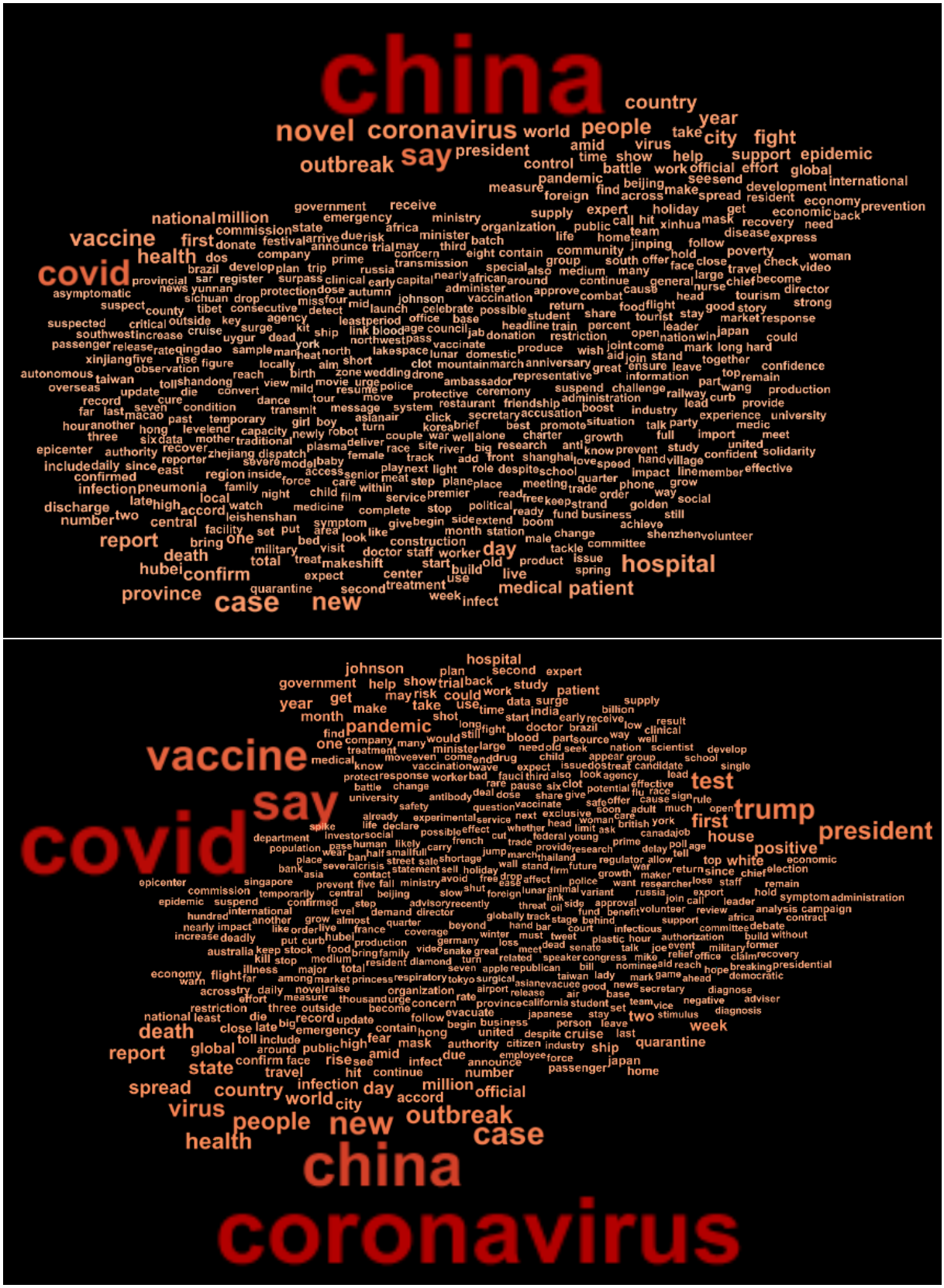


Figure 12: Closeness centrality visualization. On top the China network, on the bottom the outside China one.

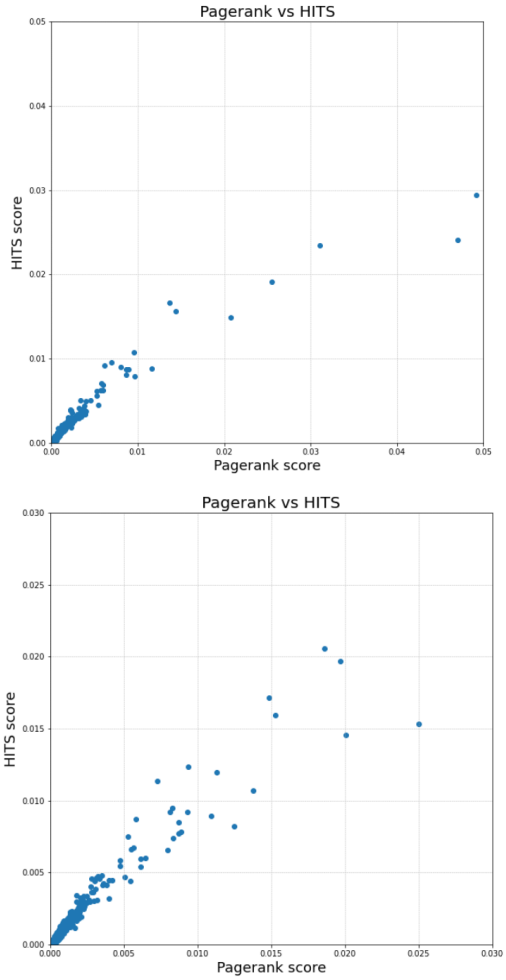


Figure 13: On top the China network and on the bottom the outside China one.

3.4.4 PageRank vs HITS and other correlations

Finally, in order to have further insight on the networks, we compared the HITS and PageRank scores. Both these algorithms are used to identify the most important nodes. In particular HITS distinguishes between hubs (node with high number of outgoing links) and authorities (nodes with high number of incoming links). Anyway, since we have undirected graphs, we have no differences between hubs and authorities. We can see the result in Figure 13.
Comparing the scores of HITS and PageRank we found that they are quite similar, since both these algorithms aim to find the more important nodes. In particular, we can see that, in some cases, especially in the outside China network, the PageRank values are higher than the HITS.
As a matter of curiosity, we also plotted other centralities versus the PageRank, as shown in Figure 14, 15 and 16.

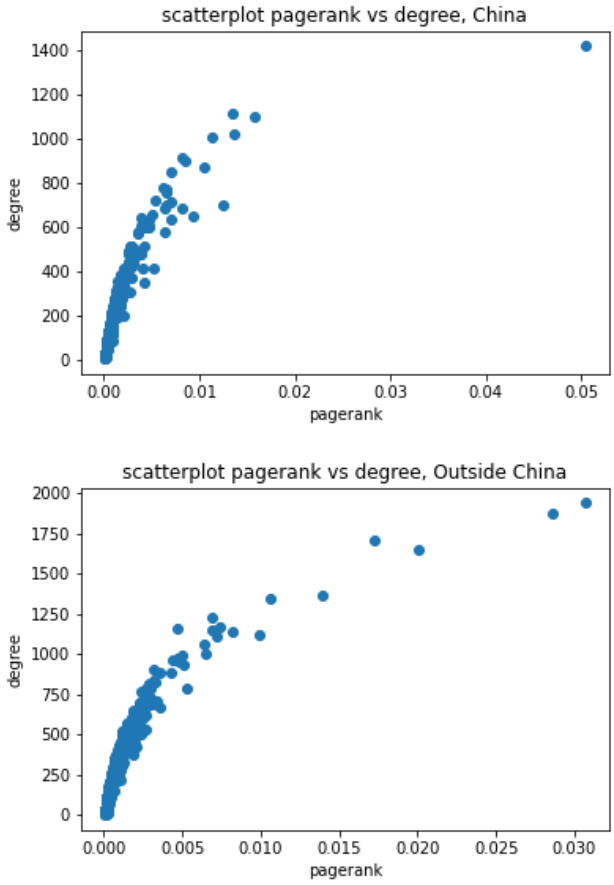


Figure 14: Scatter plot to show the correlation of PageRank and degree values in the two graphs.

3.4.5 Remarks on the exploratory analysis

Analyzing the China and outside China network with some centrality measures we have found that, even thought the networks have many nodes (more than one thousand each), only a few of them, the big hubs, determines the features of the networks. This can be seen also from the probability mass function of the different measures we analyzed: PageRank, betweenness, closeness (Figure 5, 6, 7). The majority of the nodes have small values and very few nodes with high values on the different centrality measures make the networks very small. This can be also seen by the diameter and average path length values in Table 9.

Network	Average path length	Diameter	Average degree
China	2.0749	4	25,140
Outside China	2.1036	4	39,697

Table 9: Average path length, diameter and average degree on the China and outside China network.

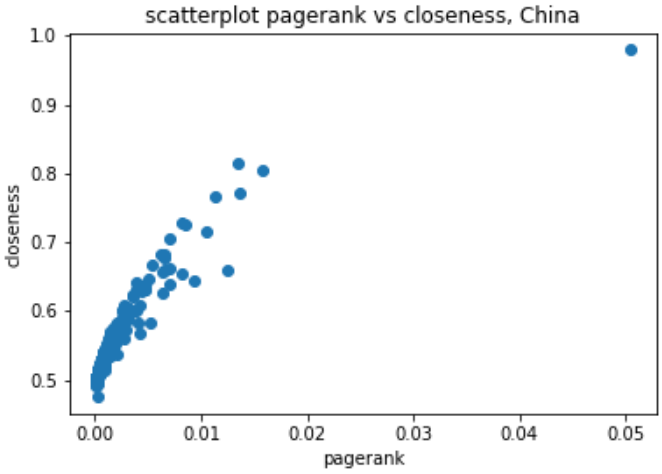


Figure 15: Scatter plot to show the correlation of PageRank and closeness values in the two graphs.

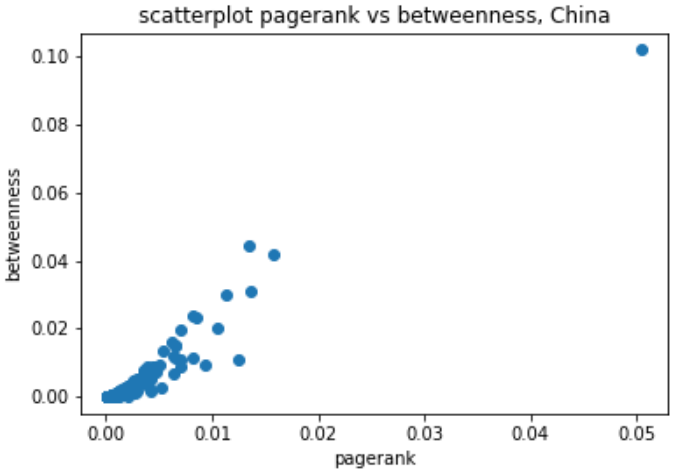
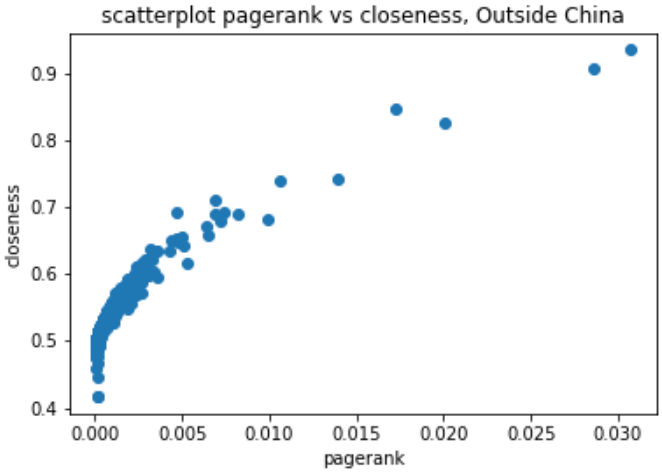
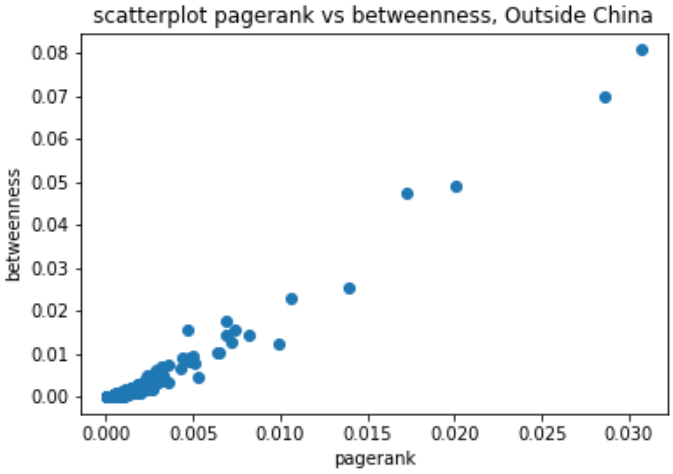


Figure 16: Scatter plot to show the correlation of PageRank and betweenness values in the two graphs.





3.5. Tweets ranking

Some analysis on the tweets based on the number of retweets and likes have been made before the construction of the networks. First we sort out the most retweeted/liked tweets by location so we can compare the top tweets. (Table 10 and 11).The results we get were quite off board compared to what we were excepting. Moreover, the topics in the tweet are too specific for the comparison to be relevant. Then to get a more general idea about topics that were the more likely to be shared we computed the average number of retweet/like group by keywords used for the collection of data (Table 12, 13, 14,15,16 and 17). The results are easier to compared but really limited by the few words we use to collect the data.

Tweet topic	retweet
New hospital for coronavirus patients in Wuhan	7994
Residents chant to hold on despite adversity.	2387
Thanks to health professionals	2115

Table 10: Topics of the 3 most retweeted tweets in China.

Tweet topic	retweet
Journal of Medicine has condemned the Trump for its response to the pandemic	19472
Trump instructed aides to stop negotiating on a coronavirus aid plan until after the election	10044
Teacher has died after testing positive	8138

Table 11: Topics of the 3 most retweeted tweets outside China.

Word	Average retweet
wuhan	96.5
coronavirus	59.6
china	21.9

Table 12: Most Retweeted keyword in China for January-February 2020.

Word	Average retweet
wuhan	15.4
china	12.4
vaccine	9.6
coronavirus	8.6
covid	7.5

Table 13: Most Retweeted keyword in China for September-October 2020.

Word	Average retweet
vaccine	12.5
who	10.1
delta	9.0

Table 14: Most Retweeted keyword in China for March-April 2021.

Word	Average retweet
wuhan	170.6
covid	152.5
coronavirus	102.0
china	84.0

Table 15: Most Retweeted keyword outside of China for January-February 2020.

Word	Average retweet
covid	96.2
coronavirus	68.0
vaccine	68.0
china	67.0
wuhan	49.0

Table 16: Most Retweeted keyword outside China for September-October 2020.

Word	Average retweet
covid	70.6
vaccine	63.2
who	58.3
delta	38.0

Table 17: Most Retweeted keyword outside of China for March-April 2021.

3.6. Robustness

In order to better understand the structure of the co-occurrency network, we investigate his robustness capabilities. To do that we look at the behaviour of the number of connected components and of the size of the giant component under different nodes removal strategy (attacks). In particular we consider different type of attacks :

- **Random failure**: we remove random nodes;
- **Hubs removal**: we remove the words with an higher degree first;
- **Closeness removal**: we remove the nodes with higher closeness centrality first;
- **Betweenness removal**: we remove the nodes with higher betweenness centrality first.

Starting from the original network, we select the nodes to eliminate according to the removal procedure chosen. Afterwards we compute the giant component size and the number of component afterwards we remove the nodes and repeat the procedures. In the case of Closeness and Betweenness removal each step of the algorithm we select a batch of nodes, since they are more computationally demanding. The results are shown in Figure 17 and 18 respectively for inside and outside China.

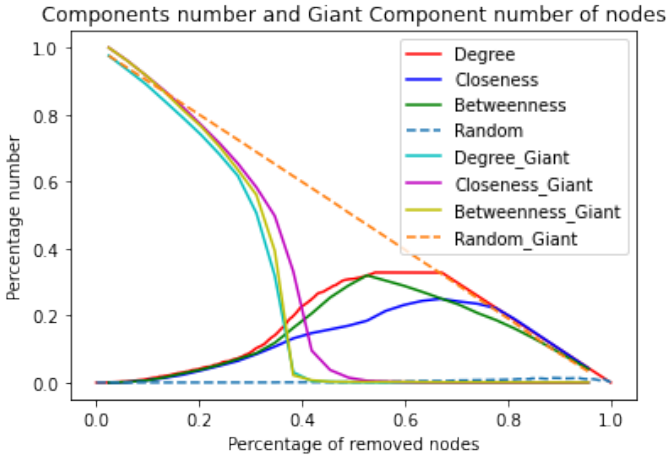


Figure 17: Gran component and components number robustness Inside China (batch = 150).

As we can see, the networks behave in a similar way. Deleting the 40% of the number of nodes is enough to disrupt both networks, either if we use Degree or Closeness removal strategy. It indicates that the network has many hubs. This also explain the behaviour under random attack which is similar to the one of a scale free network.

After having done that, we evaluated how some targeted attacks would affect the conductance of the co-occurrency

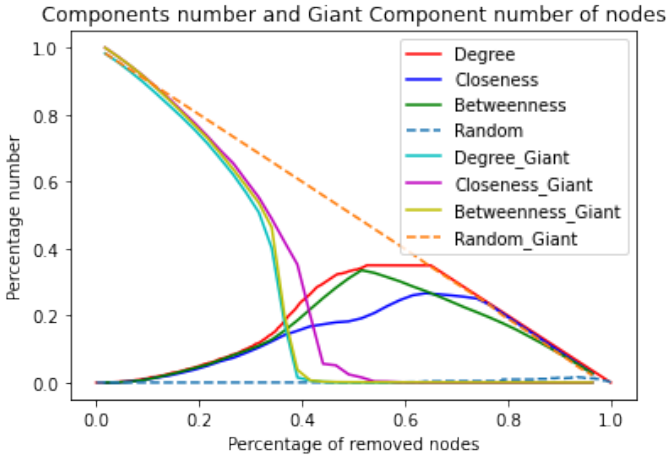


Figure 18: Gran component and components number robustness Outside China (batch = 150).

network.
The conductance of a network is defined as:

$$\phi(G) = \min_{S \subset V} \left\{ \frac{\sum_{i \in S, j \in S^c} A_{ij}}{\min(d(S), d(S^c))} \right\}$$

where $d(S) = \sum_{i \in S} \sum_{j \in V} A_{ij} = \mathbb{1}_S A \mathbb{1}$ (the sum of the degrees of the nodes in the set S plays the role of a sort of volume in the graph). Using the definition of $d(S)$, with a bit of rearrangement this last definition can be interpreted in a probabilistic way: the conductance is the minimal probability over all subsets of nodes S of the maximum between the probabilities of:

- transitioning from S to S^c in one step (using the standard transition matrix $P_{ij} = \frac{A_{ij}}{\sum_k A_{ik}}$);
- transitioning from S^c to S in one step;

In practice, $\phi(G)$ tells us how hard is to leave a subset of nodes in the graph in the worst case scenario (the optimal S).
In order to produce Figure 19 we incrementally removed targeted nodes (as explained before, ranking them in a decreasing order according to their centrality value) and we approximated the optimal S and S^c by using the Kernighan-Lin bisection algorithm. In Figure 19 we plotted the relative estimated conductance $p \mapsto \frac{\phi(G_p)}{\phi(G_0)}$ as a function of the percentage of removed nodes p, both inside and outside China. As noticed before, this approach lead us to similar conclusions in both networks, with also a similar linearly decreasing behaviour for all the three kind of attacks. In both networks a change of slope can be noticed around 20% of removed nodes, indicating a slower destroying rate after a certain threshold has been overcome. This first 20% of targeted

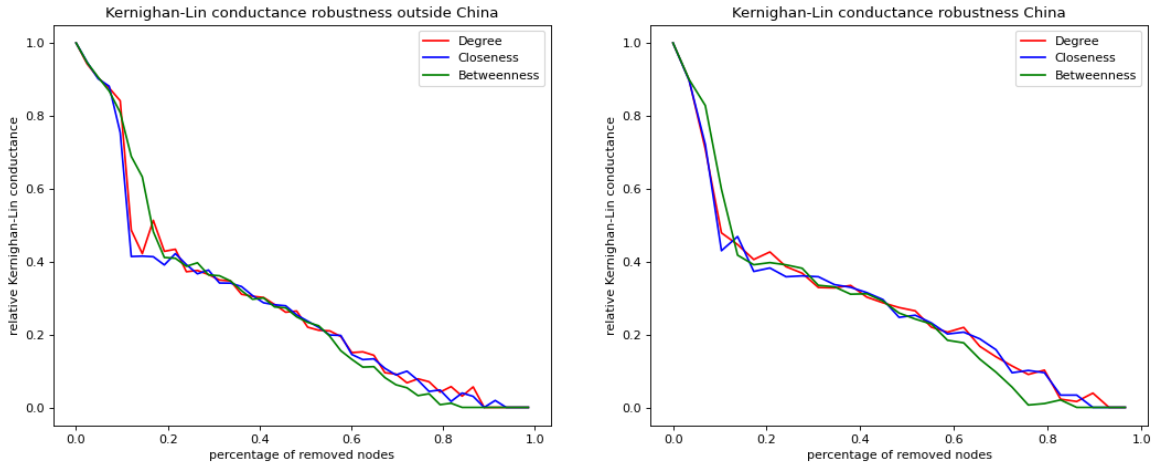


Figure 19: Kernighan-Lin estimation of the conductance as a function of the percentage of removed nodes under different strategies (batch = 100).

removals is responsible for more than a half of drop in conductance : this similarity among attacks may also suggest that there's a positive correlation among words' betweenness, degree, closeness and that the top 20% of important nodes according to this measures really work as a bridge between the communities divided by the optimal cut of the graph, as shown in fact in Section 3.4 [10].

3.7. Amen modeling and link strength prediction

In order to get some statistical significance from our analysis, we decided to test some probabilistic models that tries to capture the generative process behind the graph itself.
In particular, we tried p2-ergm and AME models [11]. These models are basically trying to fit the real probability distribution over the space of graphs after having observed the collected sample network.

More in details:

- p2 ergm model: as in all ergm models , [12], we're trying to capture a vector of sufficient statistics to describe the generative process behind the graph. In this case the statistics of interest are in degree, out degree, number of edges and the reciprocity as described in the following modeling equation

$$P(A|\mathbf{X},\mu,a,b,\beta,\rho) \propto \exp\Big[\mu \mathbb{1}\mathbb{1}^T A \mathbb{1}\mathbb{1} + \langle a, A\mathbb{1}\mathbb{1}\rangle + \langle b, A^T \mathbb{1}\mathbb{1}\rangle + \sum_{ij} \mathbf{X}_{ij}^T \beta + \rho Tr(A^2)\Big]$$

  The last one is the general model, but given that our graphs are symmetric and without features, the final model is simpler;

- Latent space models: it is an extension of the Anova decomposition (Row and column effect model), it introduces latent spaces to encode higher order effects between the nodes (like stochastic equivalence). In particular, the model we took under consideration is AME [11], described in the following

$$P(A|X,a,b,U,V) = \prod_{i,j} \sigma\Big(\beta^T \mathbf{X}_{ij} + a_i + b_j + \alpha(\mathbf{u_i}, \mathbf{v_j})\Big)$$

  where σ is the sigmoid function.
  Given the symmetry of our networks, in our case the situation is simpler since we have $\mathbf{a} = \mathbf{b}$ and $\mathbf{u_i} = \mathbf{v_i}$ for every i.

The advantage of having generative models like these two is that the distribution of any statistic $\psi(A)$ of the adjacency matrix can be approximated by simulations from the posterior distribution $P(A|A^{obs})$ (at least under the hypothesis of correct specification of the model).
Given the unsatisfying fit of the p2-ergm model, in the next analysis we focused just on the Amen one. In this context, we fitted two Amen models $P(A|A^{obs}_{\text{in\textbackslash out}})$, one for the "China" graph and one for the "outside China" one.
For this section the code had been written using R, in particular the main libraries used were [13],[14],[15].

3.7.1 Additive effects

After fitting the model on both of the graphs, we compared additive effects for the highest common Pagerank nodes as

shown in Figure 20. As shown in the latent space model equation, the bigger the additive effect the higher the "fitness" of the node (the higher the probability of receiving connections): in pratice if a node has a really negative additive effect it is less likely to receive connections according to the model.
Between the words with "opposite fitness effects" in the two networks we find: pandemic, president, coronavirus and spread.
In figure 20 we plotted the additive effects of the common highest pagerank words in the two graphs.

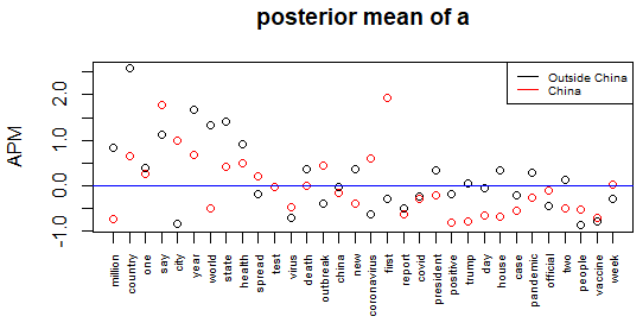


Figure 20: Plot of additive effects for the highest Pagerank nodes. Every word i has it own corresponding additive effect $a_i^{\text{in }\backslash\text{out}}$ for both the networks.

3.7.2 Latent effects

Latent effects has the same behaviour, the higher $\alpha(\mathbf{u_i}, \mathbf{v_j})$ (that in AMEN's case is just the scalar product $\mathbf{u_i}^T\mathbf{v_j}$), the higher the probability of having a connection between them according to the model. In Figure 21 we plotted the "best superposition" of latent nodal representations of the highest Pagerank words in the two graphs (latent effects are invariant under orthogonal transformations, so we found the one that minimized the least square distance between the two graphs' latent configurations, in order to try to match the latent representation of the two graphs as much as possible).
In Figure 22 we made instead a matrix plot of the difference in latent effects $U_{\text{out Ch}}U_{\text{out Ch}}^T - U_{\text{Ch}}U_{\text{Ch}}^T$ (matrix $U \in \mathbb{R}^{n\times d}$ is the matrix having has rows the latent representations of the nodes): if the difference is significantly different from zero it means that the interaction between the two nodes in question have a different effect in the two graphs.
Even thought the interaction effects here seems to be really labile, some differences are noticeable: (coronavirus,china), (first,covid), (positive, death), (covid,health), (covid,china), (covid,spread) and others. It may be important to notice that the word "China" may create a bias in the intepretation of these differences, since a Chinese journal is less likely to use autoreferential terms.

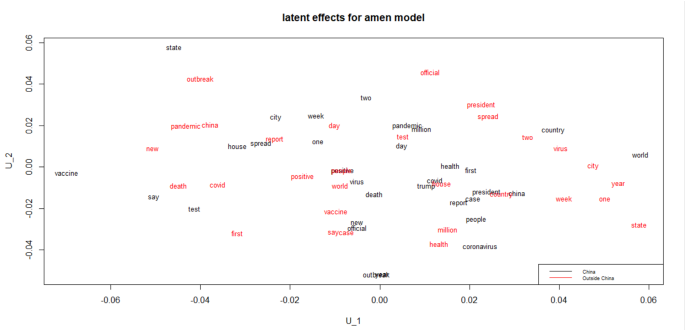


Figure 21: Plot of latent representation of the nodes. Every word i has its corresponding vector $\mathbf{u_i}^{\text{in }\backslash\text{out}}$ in the two graphs.

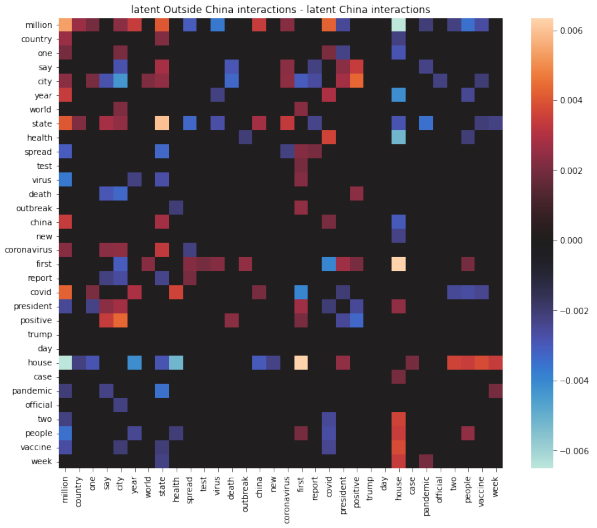


Figure 22: Thresholded delta of latent effects UU^T, Outside China minus China. Entry (i, j) of this matrix represents $\mathbf{u_i}^{\text{Out China}T}\mathbf{u_j}^{\text{Out China}} - \mathbf{u_i}^{\text{China}T}\mathbf{u_j}^{\text{China}}$.

3.7.3 Simulations

After doing that, we decided to compare the two posterior distributions by looking at the difference $P(A_{ij}|A^{obs}_{\text{out China}}) - P(A_{ij}|A^{obs}_{\text{China}})$ for the highest pagerank words. These two posteriors $P(A_{ij}|A^{obs}_{\text{country}})$ can be interpreted as a measure of the strength of the connection (i, j) according to the model.
The results of this difference are plotted in the matrix in Figure 23: if the value of an entry is positive, it means that these two words are more likely to be connected in the "outside China" graph (again, according to the model). Similar differences to the one noticed in the difference of latent effects can be read in the difference of connection probability

matrix in Figure 23.
Lastly, we decided to do a Monte Carlo simulation of the Pagerank in order to get some insights about its marginal distribution of the values. The results of this analysis are shown through a boxplot in Figure 24. As examples of significant differences found in the Pagerank values: coronavirus, president,positive,world,spread, death president and pandemic.

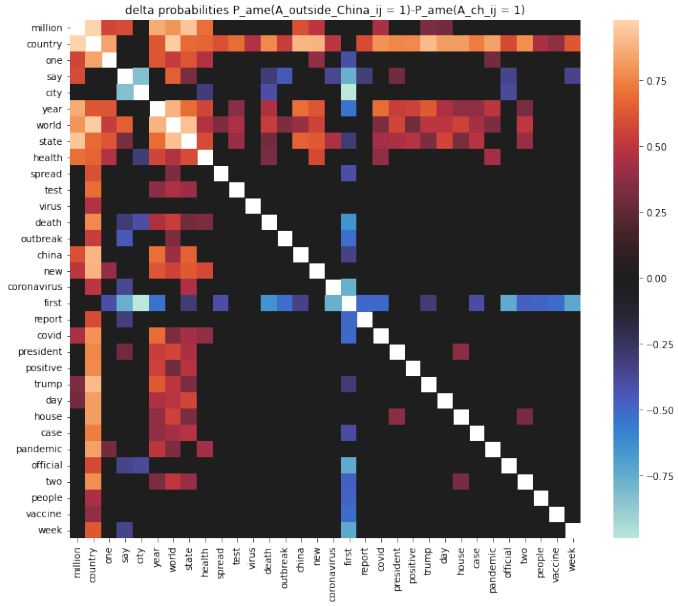


Figure 23: Thresholded differences of posterior connection probabilities, $P(A_{ij}|A^{obs}_{\text{out China}}) - P(A_{ij}|A^{obs}_{\text{China}})$. The thresholding has been done in order to set to zero differences that were in absolute values less than the mean absolute difference over all connections.

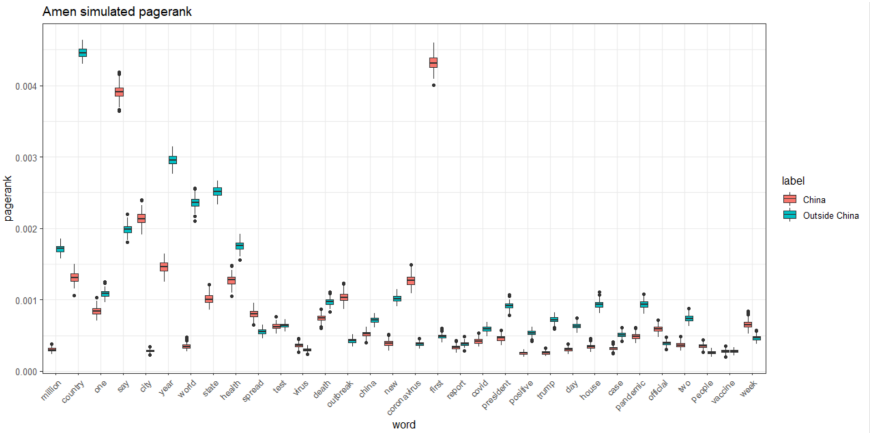


Figure 24: Boxplot comparison of the amen-simulated PageRank marginal distribution for the highest PageRank words in the two graphs.

4. Community Detection

In this section we will identify the hidden relationships that may exist among the nodes in a network, indicating communities of nodes. Since community detection can be performed with several different algorithms, we decided to implement six different strategies to carry out the task: Kernighan–Lin bipartition, dendrograms, Modularity maximization and Louvain method for discovering some relevant graph partitions, while Big CLAM and Clique percolation for highlighting possibly overlapping communities.

The results are evaluated from a qualitative point of view through the semantic interpretation of the communities, but also with a quantitative assessment based on some partition metrics and the relative frequency of the keywords (more details are given in Section 4.3.8 and Section 4.3.9.

4.1. Networks

In order to avoid biases and to identify relevant and meaningful communities, we considered the networks built by using only the search keywords "Coronavirus", "Covid", "Vaccine", therefore excluding "China", "Wuhan" and "WHO": these three keywords spontaneously showed up in the tweets. Since community detection is more interpretable and understandable on medium (about 500 or 400 nodes) and small-size (about 150 or 100 nodes) networks, we selected the most relevant nodes according to PageRank. Moreover, we decided to drop the search keywords from the final networks and eventually discard some isolated keywords in order to get a unique connected component.

In particular, we will consider two settings. First we detect the communities in the networks containing the keywords extracted from both the inside and outside China accounts' tweets, and then we verify if these communities actually identify the differences among inside and outside China. In the second setting, we perform community detection in two separate networks and then compare the results: on one hand we have the network containing the keywords extracted from the inside China accounts' tweets, on the other hand we have the same kind of network built with the outside China accounts' tweets. In both the setting we will consider the three selected period and all the periods together, for a total of four networks in the first setting and four pairs of networks in the second setting.

4.2. Communities visualization

For a better understanding of the results, we displayed the communities by using the procedure as follows. First, we create a weighted graph, referred as *community graph*, in which each node corresponds to a community and each edge is weighted according to the number of edges between the communities. This weighted graph is displayed with a circular layout. Therefore, each keyword of the original network will be assigned with a *community position* corresponding to the coordinates of the specific community graph's node that represents the community to which the keyword belongs. Then the keywords of the original network will also be assigned with a *within community position* corresponding to the coordinates obtained by displaying with a random layout the subgraph containing just keywords belonging to the same community. For each community, the subgraph is displayed individually, i.e. independently from the other subgraphs. At the end, the *community position* and the *within community position* are summed together in order to obtain the final coordinates of the keywords. These coordinates are then exported and used to initialize a graphical display in Gephi.

Initially the size of the nodes displayed were linked to their degrees in the final network. This was a simple and efficient choice to point out important node in each community. However in order to be more relevant to the real importance of those nodes in the original network their sizes in the finale display are relative to their score in the PageRank algorithm. Finally the nodes are colored according to their community and the display is cleaned using the remove overlap and label adjustment tools on Gephi.

4.3. First setting: inside and outside China grouped together

Here we apply all the aforementioned algorithms and consider medium-size networks, except for dendrograms which are more readable with small-size networks.

4.3.1 Kernighan–Lin bipartition

The Kernighan–Lin bipartition algorithm is a greedy algorithm introduced by [16] with the goal of partitioning the set of nodes V of an undirected graph $G = (V, E)$, into two disjoint subsets A and B of nearly equal size. To do so, it minimizes the sum T of the weights of the subset of edges that connect A and B. If the graph is unweighted, the algorithm assigns weight 1 to each edge.

To perform community detection with this algorithm, we used the implementation provided by the method `kernighan_lin_bisection`[1] of the NetworkX library.

The partitions performed by this algorithm are not particularly significant and informative, because the method relies on the strong assumption that the network's nodes are divided into exactly two communities of approximately same size.

[1] https://networkx.org/documentation/latest/reference/algorithms/generated/networkx.algorithms.community.kernighan_lin.kernighan_lin_bisection.html#networkx.algorithms.community.kernighan_lin.kernighan_lin_bisection

4.3.2 Dendrograms

A dendrogram represents a hierarchical graph clustering that is built to display the successive merges of nodes: it can be seen as a matrix of size $(n-1)\times 4$ in which each row contains the two merged nodes, their distance and the size of the resulting cluster, where n is the number of nodes in a graph. This can be obtained by applying to the graph an agglomerative clustering algorithm that performs greedy merge of nodes i and j based on their similarity $\frac{a_{ij}}{w_i w_j}$, where a_{ij} is the element of the adjacency matrix A representing the weight of the edge (i, j) and w_i, w_j are the degrees of the node i and j respectively.

To perform community detection with this algorithm, we used the implementation provided by the method `Paris`[2] of the Scikit-Network library.

For display and interpretability reasons, we chose to build the dendrograms on small-size networks. We tried different separation levels to have 2, 4 or 8 clusters, but here we consider just the ones characterized by the best metrics values, namely the dendrograms obtained with 4 clusters.

4.3.3 Modularity Maximization

Modularity is a network metric that measures how well a network is partitioned into communities and it is obtained in the following way:

$$Q = \frac{(Q_1 - Q_2)}{2L}$$

where Q_1 is the number of edges falling within communities, Q_2 is the expected number of edges in an equivalent network obtained through a random rewiring, and L is the total number of edges.

Q_1 and Q_2 are obtained in the following way:

$$Q_1 = \sum_{ij} a_{ij}\eta(c_i = c_j)$$

$$Q_2 = \sum_{ij} p_{ij}\eta(c_i = c_j)$$

where a_{ij} is an entry of the adjacency matrix, η is the indicator function, c_i is the community of node i, and p_{ij} is the wiring probability.

Networks with high modularity have dense connections between the nodes within the same community, but sparse connections between nodes belonging to different communities. However, since Q grows with the size of the graph and with the number of well-separated clusters, it cannot be used to compare networks with very different sizes.

Modularity optimization algorithms are used to partition a network into communities that maximize the modularity value for that network.

To perform community detection through modularity maximization, we used the implementation provided by the method `greedy_modularity_communities`[3] of the NetworkX library.

4.3.4 Louvain method

The Louvain method is a scalable algorithm that consists in 2 phases, namely *local modularity optimization* and *community aggregation*, that are iteratively executed until the maximum modularity is achieved.

In the first phase, each node is removed and added in a different community until no significant increase in modularity is verified. In the second phase, all nodes belonging to the same community are merged into a single giant node added with a self-loop weighted according to the sum of all the edges inside the corresponding community before collapsing into the giant node, while the edges connecting giant nodes are the sum of the previous ones.

To perform community detection with this algorithm, we used the implementation provided by the method `community_louvain.best_partition`[4] of the NetworkX library.

4.3.5 Clique Percolation

The clique percolation method is an algorithm for detecting overlapping communities in a network. This method is based on the concept of k-cliques, which are fully connected sub-graphs composed of k nodes. A community is defined as the maximal union of cliques of size k that can be reached through adjacent k-cliques, i.e. k-cliques that share k-1 nodes. If a sub-graph fulfills this criterion, it will be considered a community, independently of what happens to another part of the network. This definition of community is therefore based on a local property of the network.

While based on a simple concept, the clique percolation algorithm does not have good performances. In fact, it is able to identifies communities that overlap only in a small subset of nodes, which is not very useful if we are considering a big network, like in our case.

To perform community detection with this algorithm, we used the implementation provided by the method

[2] `https://scikit-network.readthedocs.io/en/latest/tutorials/hierarchy/paris.html`

[3] `https://networkx.org/documentation/latest/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html#networkx.algorithms.community.modularity_max.greedy_modularity_communities`

[4] `https://python-louvain.readthedocs.io/en/latest/api.html`

k_clique_communities[5] of the NetworkX library.

We were not able to run this algorithm on the network that comprehends the three periods together, since it was computationally too heavy. Furthermore, since the overlapping nodes are not easily displayable in a graph, we report them in Tables 20, 21 and 22 of the Appendix.

We can observe that, for all the three networks, the algorithm detects one giant community plus some very small communities, usually of the dimension of the chosen k-clique. Overall, within the same network, the smaller communities have a big overlapping with the giant community, while only overlapping on a few nodes with the other smaller communities. The wide asymmetry we obtained in the communities' sizes may be due to the fact that the clique percolation algorithm is not able to identify an overlapping on a big network, but just on a small set of nodes. This makes this algorithm not suitable to detect overlapping communities in our networks, which are indeed quite big.

4.3.6 BigCLAM

BigCLAM stands for Cluster Affiliation Model for Big Networks and is an overlapping community detection method that scales to large networks, introduced by [17]. The idea is to take into account a graph $G = (V, E)$ with $n = |V|$ nodes and a guessed number of communities c, and solve the following maximum likelihood estimation problem with gradient ascent:

$$\max_{M} \prod_{(i,j)\in E} (1 - Q_{ij}) \prod_{(i,j)\notin E} Q_{ij}$$

where $Q = \exp(-MM^T)$, and $M > 0$ is the membership matrix of size $n \times c$ containing for each row the probabilities of the corresponding node to belong to a certain community.

To perform community detection with this algorithm, we used the implementation provided by the method big_clam[6] of the CDlib library.

4.3.7 Semantic interpretation

In this section we will discuss the semantic meaning of the detected communities. For each period, we will show the graphical displays of communities, according to the different algorithms that we considered.

This analysis is carried out as follows. First we compare the three methods having the poorer performances, namely Kernighan–Lin bipartition, dendrograms and BigCLAM. Indeed, as previously mentioned in Section 4.3.1, Kernighan–Lin bipartition is biased towards having exactly two communities having almost the same size, and this is a great disadvantage in our case: Covid-19 is a worldwide discussed topic and therefore it may be too simplistic to reduce the content of the news to two main topics. Moreover, the graphical displays of the dendrograms are easy to understand, but in general dendrograms rarely provide the best solution. Finally, the BigCLAM algorithm, in our networks, tends to select two giant communities leaving just few nodes outside of them and the resulting communities don't overlap, therefore it is comparable with a bipartition method. Probably, this is due to the fact that relevant overlapping communities are not present in our networks.

Finally, we compare the two strategies leading to the best performances, namely Modularity maximization and the Louvain method, both based on modularity optimization.

We decided to not discuss the semantic interpretation and to not show the visualization of the Clique percolation overlapping communities because, as explained in Section 4.3.5, the results are not much informative in our case.

Worst-performing algorithms

As we said in Section 4.3.2, the dendrograms are computed on a small-size network, while the other two algorithms are applied on a medium-size network. This holds in general, for all the periods.

Figure 25 displays the communities detected with Kernighan–Lin bipartition and BigCLAM on the networks referring to all the three selected periods together. Both the algorithms tend to form a giant community containing the hubs ("china", "say", "new", "case", "president", "trump" and so on) and another giant community with nodes having a low degree. In particular, we can observe that the two main communities of BigCLAM are similar to the communities detected by the Kernighan–Lin bipartition: they share, respectively 212 and 210 keywords. However, we are not going to semantically interpret these communities in details because they collect many different topics together. Also the small community detected by BigCLAM seems to describe different topics (vaccinations, elections, and the African continent) and therefore it is not relevant itself.

Figure 26 displays the communities detected with Kernighan–Lin bipartition and BigCLAM on the networks referring to the January-February 2020 period. Similarly as before, we have a giant community containing the hubs and another giant community containing just low-degree nodes. The two main communities detected by the BigCLAM algorithm are somehow similar to the communities detected by the Kernighan–Lin bipartition: in that they share many

[5] https://networkx.org/documentation/latest/reference/algorithms/generated/networkx.algorithms.community.kclique.k_clique_communities.html#networkx.algorithms.community.kclique.k_clique_communities

[6] https://cdlib.readthedocs.io/en/latest/reference/cd_algorithms/algs/cdlib.algorithms.big_clam.html

of their keywords (149 and 68 respectively). However, unlike the bipartition algorithm that is constrained to detect same-size communities, the BigCLAM algorithm is able to detect a relatively small community of hubs ("china", "outbreak", "case", "new", "novel", "death", "health") that seems to summarize the main topics of the discussion in the January-February 2020 period: the virus outbreak in China, the report of the new Covid-19 cases and the consequent public health emergency. The third very small community detected by this algorithm seems to focus on the humanitarian aid that many nations started to give to the countries in need due to the health crisis ("support", "help", "outside", "effort", "disease", "center").

Figure 27 displays the communities detected with Kernighan–Lin bipartition and BigCLAM on the networks referring to the September-October 2020 period. Again, for both the algorithms we have a giant community containing the hubs ("president", "trump", "say", "test", "positive", "new", "case", "report", "pandemic", "white", "house" and so on) and another giant community containing just low-degree nodes. In particular, the hubs describe the Trump's announcement (on October 2, 2020) that he and the first lady had tested positive for Covid-19. The third very small community detected by BigCLAM is not informative.

Figure 28 displays the communities detected with Kernighan–Lin bipartition and BigCLAM on the networks referring to the March-April 2020 period. Even in this case, both the algorithms detect two giant communities, with only low-degree nodes and one with hubs ("say", "new", "case", "report", "health", "country", "johnson", "people", "million", "death" and so on) that incorporate two main topics: Covid-19 report of infection and death cases and vaccines. The third and fourth very small community detected by BigCLAM are not informative, even if one seems to indicate an estimation of the contagion rate, probably referred to a Covid-19 variant ("risk", "month", "infection", "second").

In conclusion, the Figures 29, 30, 31 and 32 report the dendrograms relative to the three selected periods and all the periods together. The figures include a brief description of the detected communities.

Best-performing algorithms

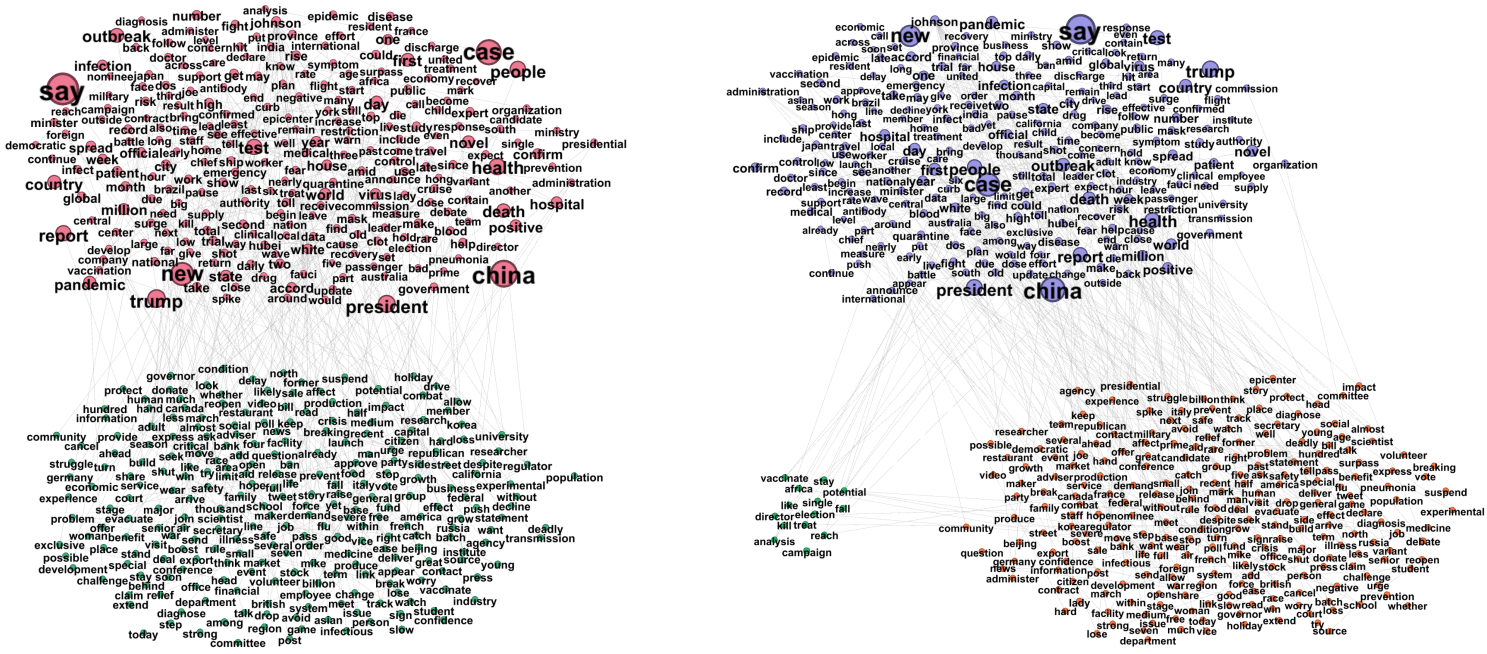
In general, Modularity maximization and the Louvain method find pretty similar communities: the number of communities may vary but, overall, the semantic content remains almost the same. We will leave out of the analysis some very small and not much interpretable communities. This kind of communities increases in number as we proceed from the earlier period to the later period: this could possibly mean that the Covid-19 debate was more polarized in some major topics during the January-February 2020 period, while getting more various and including much different topics in the March-April 2021 period. This is reasonable because at first the official channel news tended to report almost the same information about the health emergency, but as time passed they started to enrich and vary the content of their tweets.

The Figure 33 displays the communities detected with Modularity maximization and the Louvain method on the medium-size network referring to all the three selected periods together. As reported in Table **??** the main communities that has been found considering all the periods together are five and concern the Covid-19 outbreak, the USA politics, the worldwide report of cases of infections and deaths, the vaccines and the people's concerns about the pandemic from the point of view of the health and the economy. Besides these, there is a smaller and less discussed community describing the Diamond Princess case[18].

The Figure 34 displays the communities detected with Modularity maximization and the Louvain method on the medium-size network referring to the January-February 2020 period. As reported in Table **??** the main communities that has been found considering the January-February 2020 period are five and concern the Covid-19 outbreak, the official communications by the World Health Organization, the report of first Covid-19 cases of infections especially for what concerns Asian countries, the medical assistance to Covid-19 patients and the Diamond Princess case, which here has more relevance than before (the community contains more keywords and the keywords degrees are proportionally higher).

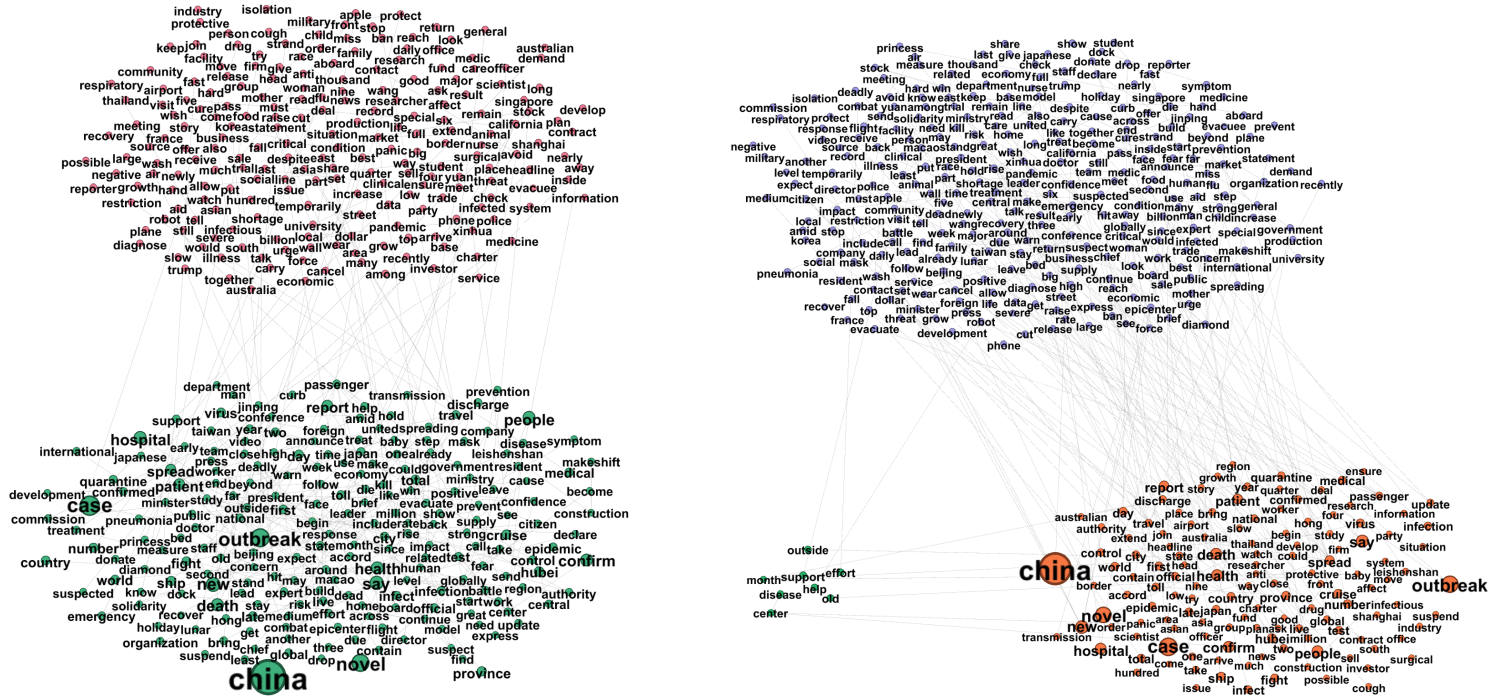
The Figure 35 displays the communities detected with Modularity maximization and the Louvain method on the medium-size network referring to the September-October 2020 period. As reported in Table **??** the main communities that has been found considering the September-October 2020 period are four and concern the worldwide report of cases of infections and deaths, the people's concerns about the future development of the pandemic, the vaccine campaign, production and inoculation on one hand and the vaccine debate and regulations on the other hand.

In conclusion, the Figure 36 displays the communities detected with Modularity maximization and the Louvain method on the medium-size network referring to the March-April 2020 period. As reported in Table 37 the main communities that has been found considering the March-April 2021 period are four and concern the worldwide report of cases of infections and deaths, the USA politics and public administration, the vaccine and the general debate about health, Covid-19 preventive measures, economy and topics of public interest. Besides them, some small communities concern the international travel bans, the spread of fake news, the Anthony Fauci interview [19], the Russian vaccine for animals, the French new lockdown and hospital emergency and some issues related to the Tokyo Olympics.



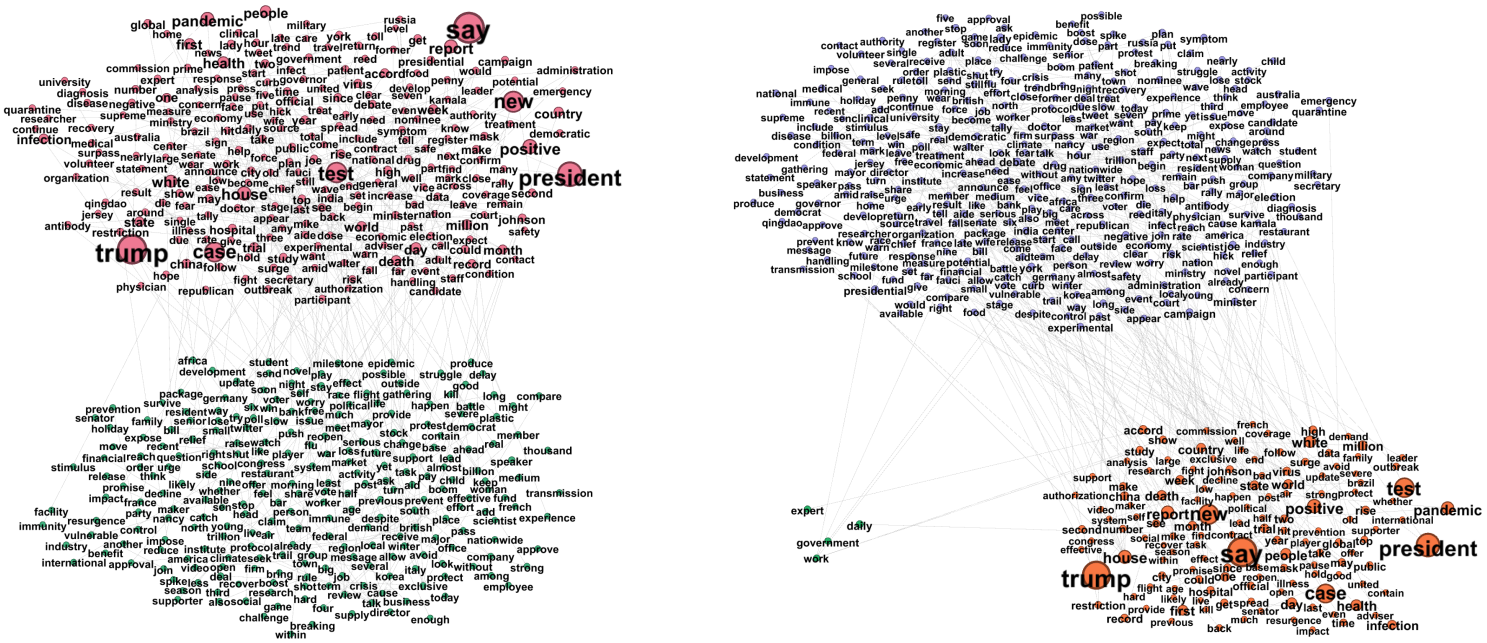
(a) Kernighan–Lin bipartition. (b) BigCLAM.

Figure 25: Kernighan–Lin bipartition and BigCLAM detected communities for all the three periods together.



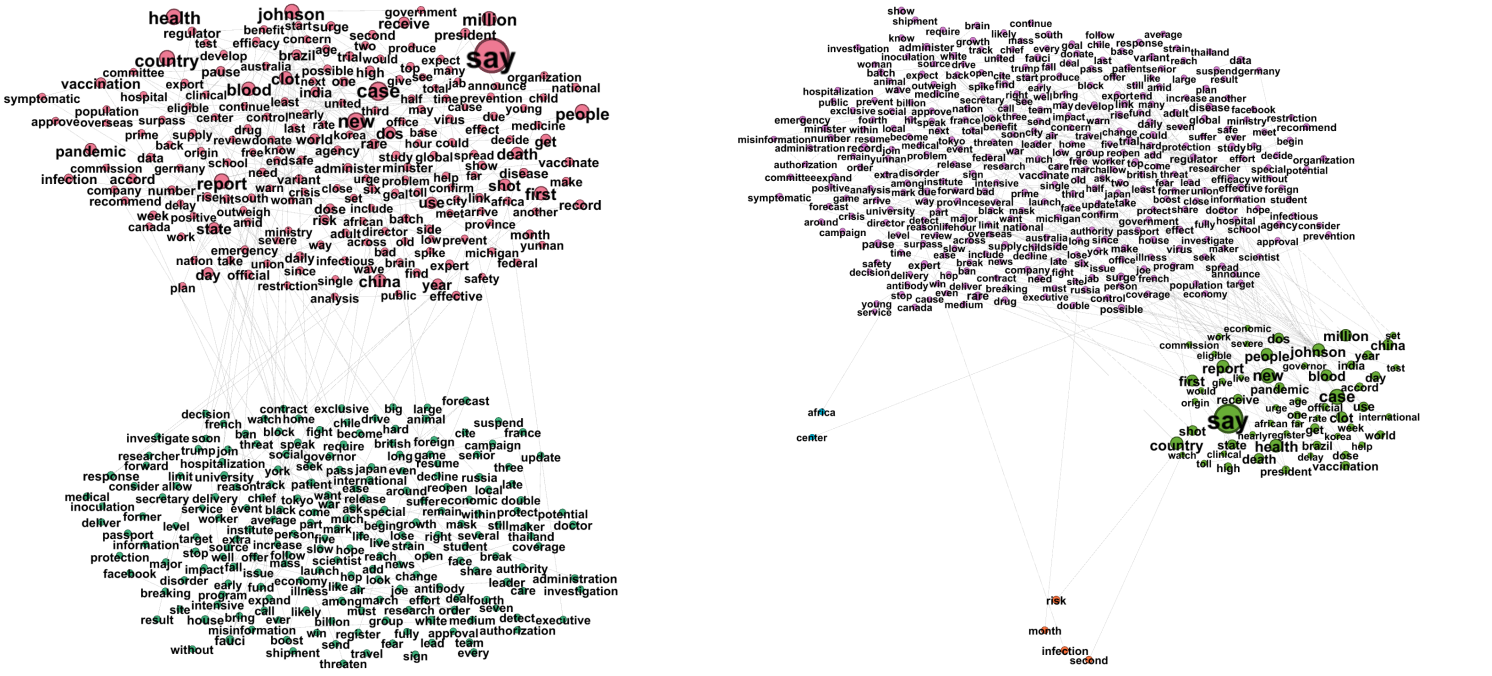
(a) Kernighan–Lin bipartition. (b) BigCLAM.

Figure 26: Kernighan–Lin bipartition and BigCLAM detected communities for the January-February 2020 period.



(a) Kernighan–Lin bipartition. (b) BigCLAM.

Figure 27: Kernighan–Lin bipartition and BigCLAM detected communities for the September-October 2020 period.



(a) Kernighan–Lin bipartition. (b) BigCLAM.

Figure 28: Kernighan–Lin bipartition and BigCLAM detected communities for the March-April 2021 period.

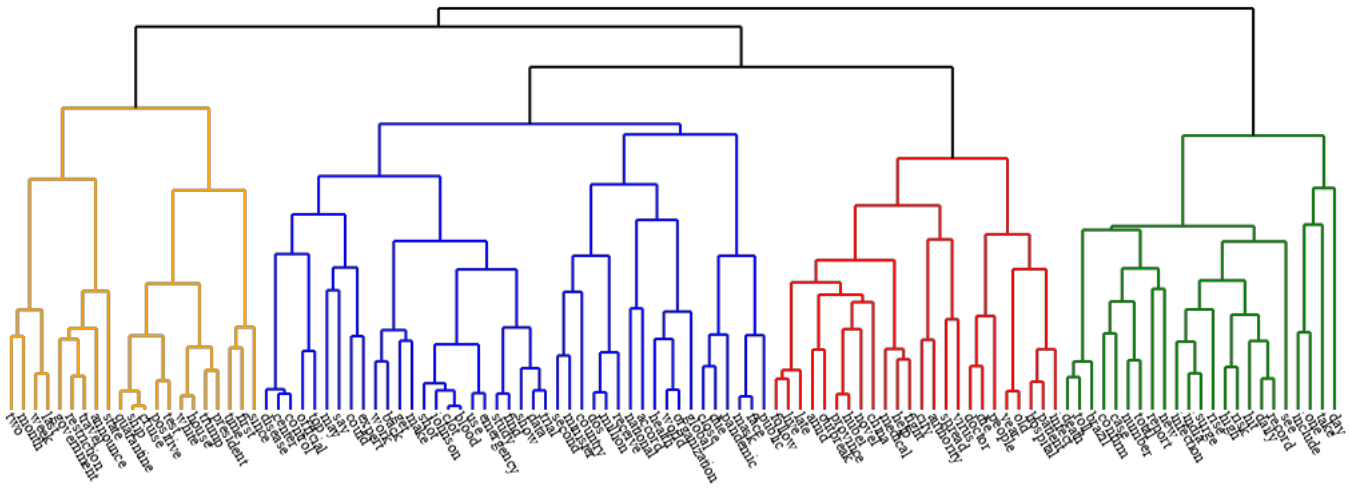


Figure 29: Dendrogram detected communities for all the three periods together. The relevant communities are four. The yellow community refers to the news about president Trump testing positive for Covid-19. The blue one refers to WHO's recommendations about safety measures and the Johnson&Johnson vaccine being paused after the report of possible side effects, such as blood clots. The red community is about the pandemic outbreak in China. Finally, the green community seems to focus on the international situations and on the increasing death tolls.

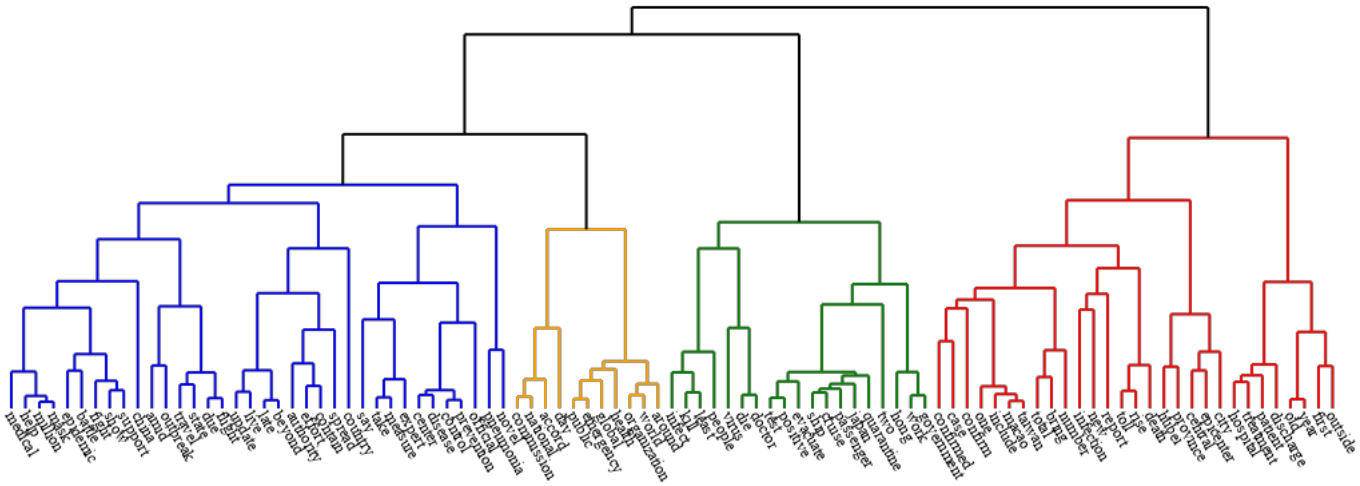


Figure 30: Dendrogram detected communities for the January-February 2020 period. The relevant communities are four. The blue community refers to the pandemic outbreak in China and to the need of adopting safety measures, such as masks and travel restrictions, to contain the pandemic. The yellow community refers to the WHO's declaration of global emergency. The green community is about the case of the Diamond Princess cruise ship, that stopped in Japan for the duration of the passenger's quarantine. Finally, the red community is about the first cases of Covid-19 in the Chinese provinces of Hubei and Macao.

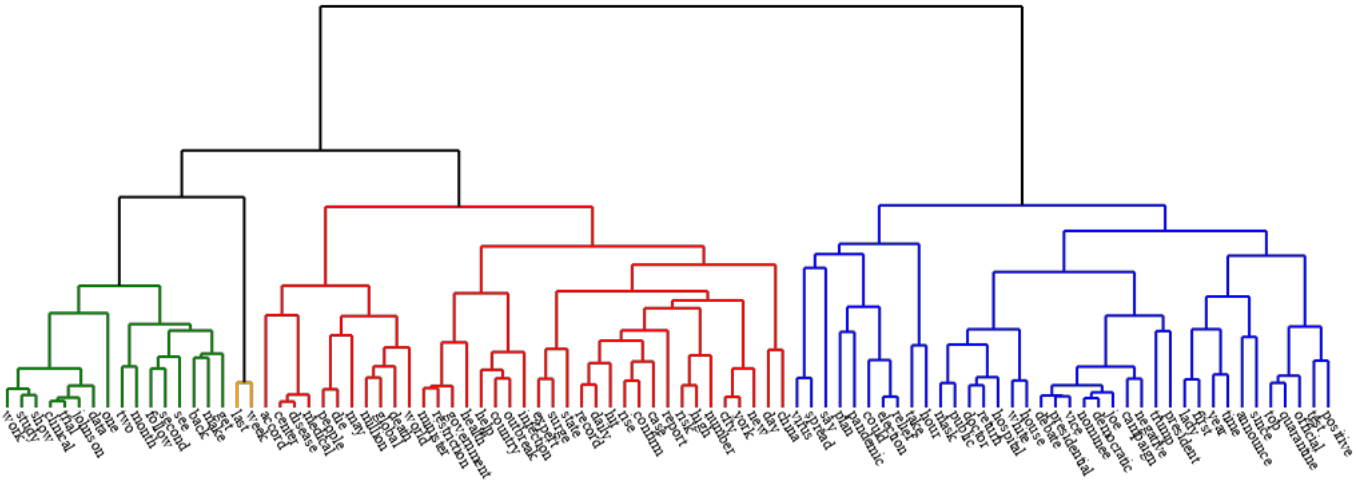


Figure 31: Dendrogram detected communities for the September-October 2020 period. The relevant communities are three. The green one refers to the vaccine research and, presumably, the time needed for the vaccine to activate our immune response and the duration of the vaccine coverage. The red one is about the health crisis and the consequent restrictions needed in the emergency. Indeed it contains Keywords referring to the reports about the rising of number of infections and deaths, and it also mentions the Chinese country and New York city. The blue one refers to the USA presidential elections and the Trump's announcement that he was positive-tested. The keywords also refer to the hospital administration, the safety measures and the pandemic plan.

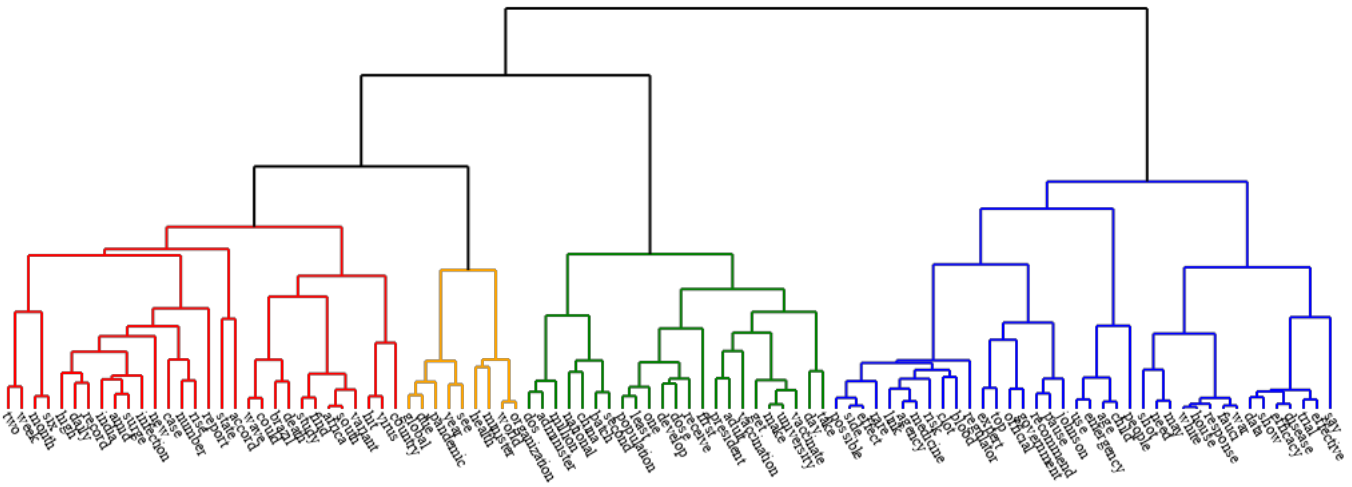
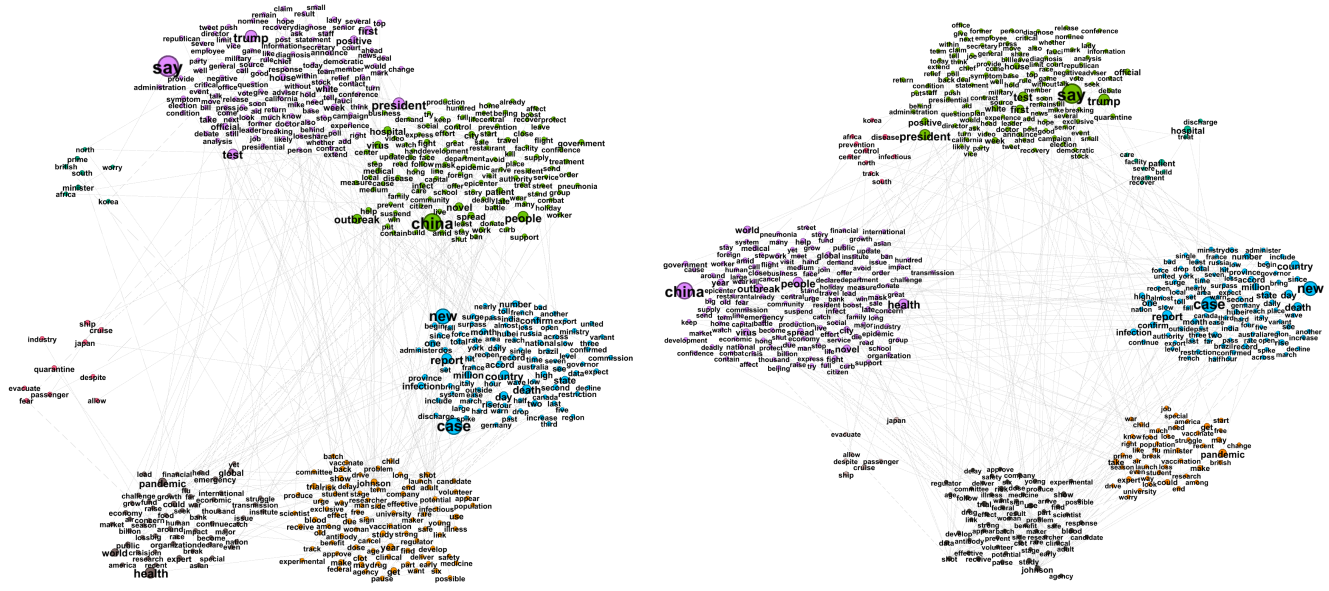
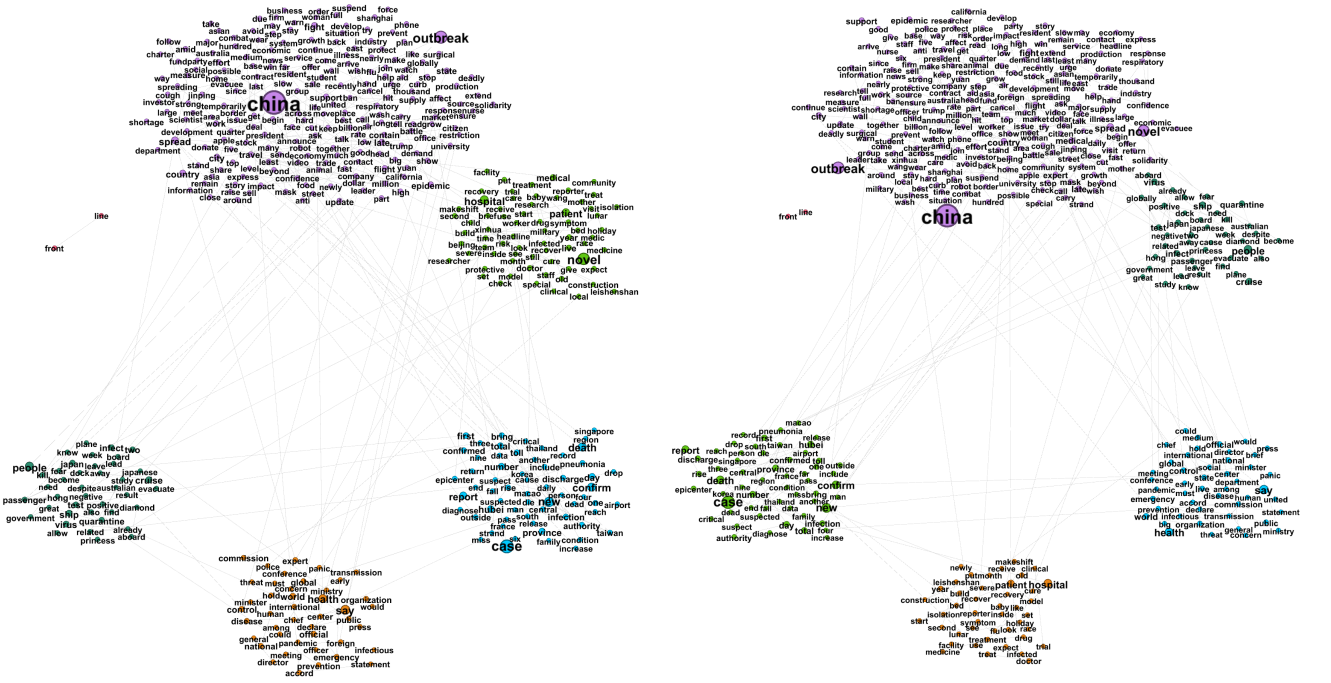


Figure 32: Dendrogram detected communities for the March-April 2021 period. The relevant communities are four.
The red one refers to the arise of a new variant in India and the increase of Covid-19 infection and death rates in the world, with a mention to South Africa and Brazil. The yellow one is tightly linked to the red one and indeed it focuses on the World Health Organization reports about the pandemic. The green one refers to vaccination campaign and the inoculation of the vaccine to the population, with a mention to the second dose. The blue one refers to debate about vaccines that involves the Anthony Fauci declarations, the child vaccination and the Johnson&Johnson vaccine paused after the report of possible side effects (blood clot).



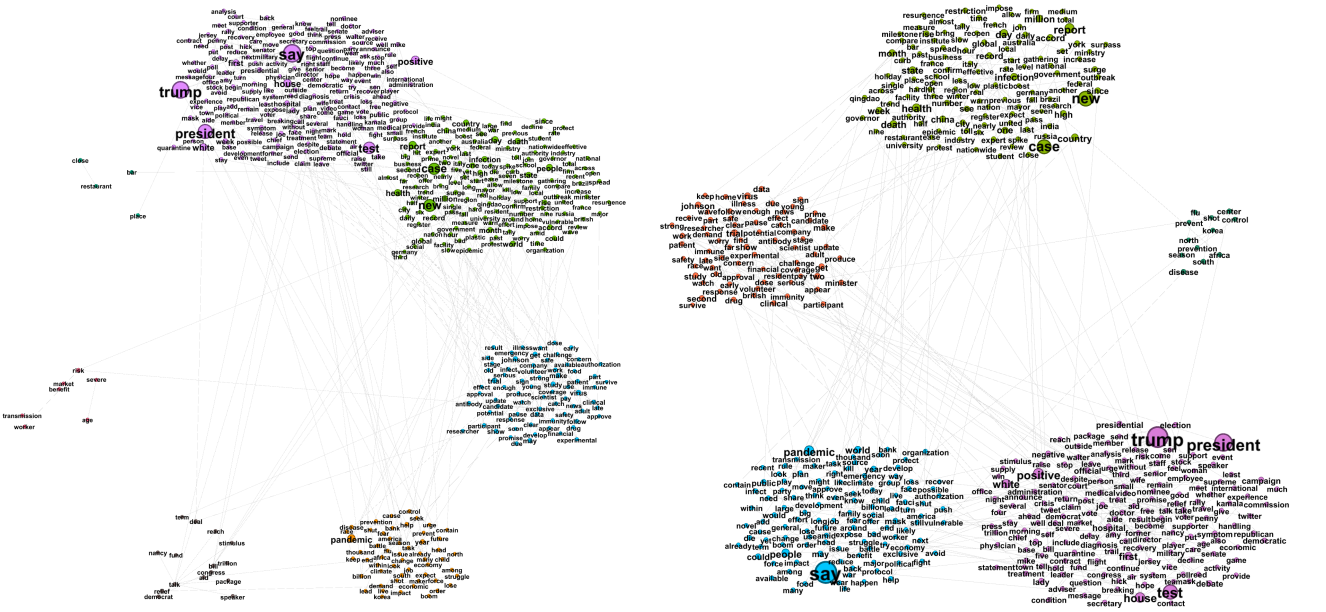
(a) Modularity Maximization. (b) Louvain method.

Figure 33: Best-performing algorithms detected communities for all the three periods together.



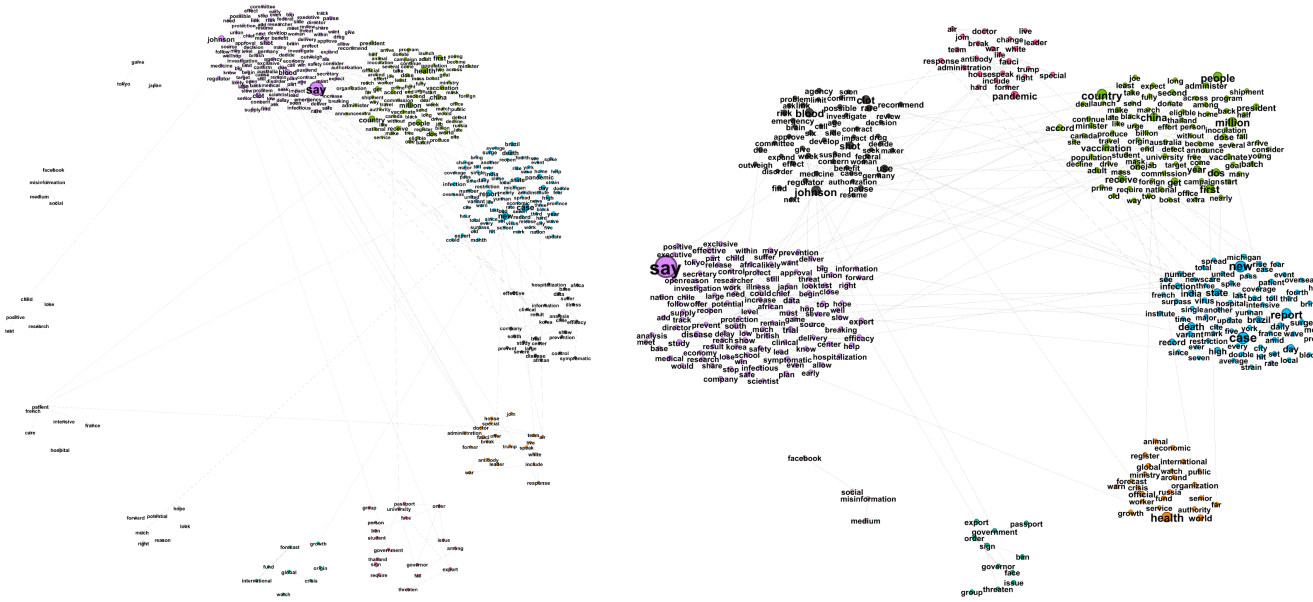
(a) Modularity Maximization. (b) Louvain method.

Figure 34: Best-performing algorithms detected communities for the January-February 2020 period.



(a) Modularity Maximization. (b) Louvain method.

Figure 35: Best-performing algorithms detected communities for the September-October 2021 period.



(a) Modularity Maximization. (b) Louvain method.

Figure 36: Best-performing algorithms detected communities for the March-April 2021 period.

Community name	Keywords summary	Semantic content
Covid-19 outbreak	**Common:** china, outbreak, people, virus, hospital, spread, government, pneumonia, epicenter, beijing, worker, die; **Louvain:** economic, capital, financial, fund, market **Modularity Max:** mask, patient, prevention, infect, disease	Description and effects of the pandemic: Louvain focuses more on the economic aspects, while Modularity Maximization focuses more on the medical aspects.
USA politics	say, president, trump, test, positive, debate, official, white house, provide, campaign, conference, election, vote, military, administration, first lady	Politic and administrative scenario in the USA during the pandemic. Reference to the fact that Trump turned out to be positive.
Worldwide report	new case, report, million, death, country, day, state, accord, infection, germany, russia, france, italy, brazil, canada, australia, hubei	Worldwide report of Covid-19 cases and deaths.
Concerns	**Common:** pandemic, expert, research, struggle, loss, war, break, flu; **Louvain:** worry, free, need, vaccination; **Modularity Max:** impact, challenge, crisis, health, emergency, concern, financial, economic, bank	Concern about the pandemic: Louvain focuses more on health struggles, while Modularity Maximization focuses on the economic crisis.
Vaccines	johnson, experimental, volunteer, develop, study, trial, shot, receive, antibody, committee, approve	Vaccine debate.
Diamond Princess	japan, ship, cruise, passenger, evacuate	In February 2020, the Diamond Princess cruise ship stopped in Japan for the duration of the passenger's quarantine. Negative tested passengers have been evacuated.

Communities found by Modularity Maximization and the Louvain method for all the periods together.

Community name	Keywords summary	Semantic content
Covid-19 outbreak	china, outbreak, spread; **Safety measures:** measure, mask, wash, protect; **Politics:** information, ban, president, trump; **Economy:** economy, dollar, investor, industry; **Community feelings:** solidarity, help, together, confidence, donate	Description and effects of the pandemic from different points of view: safety measures, politics, economy and community feelings.
Official communication	say, world health organization, global, statement, press, declare, official, international, disease, emergency, pandemic, panic	Official communication of the beginning of the pandemic by WHO.
Asia first cases report	new case, report, death toll, diagnose confirm, first, hubei, taiwan, thailand singapore, macao, korea	Report of the fist Covid-19 cases in Asia.
Medical care	hospital, patient, doctor, symptom, isolation, recovery, cure, clinical, medicine, treatment	Situation in the hospitals with Covid-19 patients.
Diamond Princess	japan, diamond, princess, ship, cruise, passenger, evacuate, quarantine, virus, aboard, people, test	In February 2020, the Diamond Princess cruise ship stopped in Japan for the duration of the passenger's quarantine. Negative tested passengers have been evacuated[18].

Communities found by Modularity Maximization and the Louvain method for the January-February 2020 period.

Community name	Keywords summary	Semantic content
Worldwide report	new case, report, china, death, health, infection, million, country, day, state, accord, infection, germany, russia, france, italy, brazil, australia, india	Worldwide report of Covid-19 cases and deaths.
USA politics and public administration	trump, president, white house, test, positive, kamala, first lady, administration, hospital, republican, democrat, campaign, diagnosis, announce, crisis	USA campaign for the presidential elections and public administration.
Vaccines	experimental, dose, johnson, antybody, drug, immune, coverage, virus, participant, study, trial, clinical, safety, approve, effect, concern	Discussion about vaccines: experimental results and concerns about the possible effects.
Concerns	pandemic, climate, battle, economy, boom, order, impact, change, future, fear, issue, struggle, lose, shut, help	Concerns about the future development of the pandemic.

Communities found by Modularity Maximization and the Louvain method for the September-October 2020 period.

Community name	Keywords summary	Semantic content
Worldwide report	new case, report, death, record, variant, infection, virus, day, spike, news, infection, wave, spread, state, michigan, france, brazil, india, yunnan	Worldwide report of Covid-19 cases and deaths.
Vaccine public administration	country, million, population, people, inoculation, vaccination campaign, boost, dose, produce, urge	Organization of the vaccine campaign, production and inoculation.
Vaccine debate and regulation	johnson, shot, drug, **Health:** risk, cause, clot, blood, problem, concern, impact, benefit; **Regulation:** regulator, approve, authorization, recommend, suspend, confirm	Debate about the risks and benefits for heath of the vaccines and communication of the vaccine regulation.
General debate about Covid-19	**Common:** hospitalization, symptomatic, disease, control, clinical, information, study, prevention, effective, data; **Louvain:** reopen, school, economy, supply, executive, nation, expert	General discussion about health, Covid-19 preventive measures, economy, public interests.
Anthony Fauci	fauci, trump, team, speak, white house, administration, leader, antibody response	Anthony Fauci was interviewed in March 2021 as the Director of the National Institute of Allergy and Infectious Diseases and one of the lead members of the White House Coronavirus Task Force in USA.
Travel bans	passport, ban, government, issue, order, export, threaten, group, sign	Description of the international regulations about travel bans.
Fake news	facebook, social, misinformation, medium	Spread of misinformation on digital medium, like social networks.

(a) Common communities.

Community name	Keywords summary	Semantic content	Algorithm
Russian vaccine for animals	world health organization, russia, animal, register, authority, forecast, official, ministry	Russia has registered the world's first Covid-19 vaccine for animals, according to the country's agriculture safety.	Louvain
French lockdown	pandemic, climate, battle, economy, boom, order, impact, change, future, fear, issue, struggle, lose, shut, help	French introduces new restrictions, brings in extra staff for the hospitals for the rising of the intense care patient numbers.	Modularity Max
Tokyo Olympics	tokyo, japan, game	North Korea refused to take part to Tokyo Olympics due to Covid-19 fears and the Tokyo Olympics faced a possible postponing.	Modularity Max

(b) Communities detected only by one of the two algorithms.

Figure 37: Communities found by Modularity Maximization and the Louvain method for the March-April 2021 period.

4.3.8 Metrics-based analysis

In this section we will compare the partitions $\mathcal{P}$ of our community detection algorithms by considering the following metrics for a graph $G = (V, E)$ with $n = |V|$ nodes:

- **Partition Coverage**: the ratio of the number of intra-community edges to the total number of edges in the graph, that is

$$C(\mathcal{P}) = \frac{|(i,j) \in E : C_i = C_j|}{|(i,j) \in E|} \in [0,1]$$

  where C_i, C_j are the communities to which nodes i, j respectively belong;

- **Partition Performance**: the number of intra-community edges plus inter-community non-edges divided by the total number of potential edges, that is

$$P(\mathcal{P}) = \frac{|(i,j) \in E : C_i = C_j| + |(i,j) \notin E : C_i C_j|}{n(n-1)/2} \in [0,1]$$

  where C_i, C_j are the communities to which nodes i, j respectively belong;

- **Modularity**: the number of edges falling within the communities minus the expected number of edges in an equivalent network having the edges placed at random (the mathematical formalization was given in Section 4.3.3).

By definition, an ideal clustered graph is characterized by clusters that are disconnected from each other: this yields a partition coverage of 1, as all edges of the graph fall within clusters.

A similar observation also holds for the partition performance: it counts the number of "correctly interpreted" pairs of vertices, that are couples of vertices belonging to the same community and connected by an edge, or couples of vertices belonging to different communities and not connected by an edge; therefore the higher the partition performance, the better the community detection.

For what concerns modularity, the more the number of internal edges of the cluster exceeds the expected number, the better defined the community; therefore, also for this metric, large positive values generally indicate good partitions.

All the previous insights about the metrics are taken from [20].

Tables 18 summarize the results we obtained on four medium-size networks, one for each selected period, plus one for all the periods together. Note that we did not compute the metrics for the Clique Percolation algorithm since they are not thought for overlapping communities. Also, we did not display the metrics for the dendrogram algorithm, because this method has not been applied to the same networks as the other methods, but on some small-size networks, for a better interpretability of the graphical displays. In fact, our metrics should not be used to compare the quality of the community structure of networks that are very different in size.

From the results in the tables, we can observe that, while Kernighan–Lin bipartition algorithm reaches the highest partition coverage in all the networks, it also always have very low modularity values. Indeed, this algorithm minimizes the sum of the weights of the edges connecting the two detected communities and therefore it is built to leave only few and low weighted edges between the communities: this implies that $|(i,j) \in E : C_i C_j|$ is low for this algorithm, and consequently $|(i,j) \in E : C_i = C_j| = |(i,j) \in E| - |(i,j) \in E : C_i C_j|$ is reasonably high.

Overall, BigCLAM performs very poorly with respect to all the metrics. Indeed it doesn't significantly improve the bipartition method and doesn't even detect overlapping communities on our data. Therefore it actually doesn't enrich the analysis with interesting insights, in our case.

In conclusion, Modularity Maximization and Louvain show low values for the partition coverage metric, but have the highest values for partition performance and modularity. This was expected since they are both based on modularity optimization. Overall, these two methods obtain very similar results in terms of metrics performance.

4.3.9 Frequency-based analysis

To conclude, we analyse whether the communities are useful to distinguish the inside China tweets from the outside China tweets.

Indeed, after computing the communities, we wanted to investigate whether a single community actually reflect only the inside China tweets or only the outside China tweets. In order to perform this analysis, we choose to assign a keyword key to the inside China tweets (i.e. $where(key) = inside$) if the keyword is more frequent in those tweets than in the outside China tweets, and viceversa (i.e. $where(key) = outside$). Since the number of collected tweets inside China is generally lower than the the number of tweets outside China, we considered the relative frequencies and assigned a keyword to the inside China tweets if it was at least 5% more (relatively) frequent in those tweets than the outside China tweets, and viceversa. If this doesn't occur than the keyword is assigned to both the tweets (i.e. $where(key) = both$). Formally:

$$where(key) = \begin{cases} inside & f_{in}(key) - f_{out}(key) \geq 0.05 \\ outside & f_{out}(key) - f_{in}(key) \geq 0.05 \\ both & otherwise. \end{cases}$$

	Bipartition	BigCLAM	Modularity Maximization	Louvain method
Coverage	0.708	0.672	0.362	0.383
Performance	0.536	0.556	0.780	0.754
Modularity	0.007	-0.003	0.221	0.219

(a) All the periods together.

	Bipartition	BigCLAM	Modularity Maximization	Louvain method
Coverage	0.837	0.419	0.473	0.485
Performance	0.522	0.450	0.689	0.652
Modularity	0.005	-0.046	0.267	0.270

(b) January-February 2020.

	Bipartition	BigCLAM	Modularity Maximization	Louvain method
Coverage	0.723	0.433	0.576	0.524
Performance	0.518	0.411	0.729	0.750
Modularity	0.263	-0.039	0.301	0.298

(c) September-October 2020.

	Bipartition	BigCLAM	Modularity Maximization	Louvain method
Coverage	0.729	0.399	0.577	0.528
Performance	0.513	0.277	0.804	0.825
Modularity	0.207	-0.033	0.350	0.356

(d) March-April 2021.

Table 18: Comparison of the metrics for all the periods, for all the comparable algorithms.

where $f_{in}(key)$ and $f_{out}(key)$ are the relative frequencies of the keyword key in the inside China and outside China tweets, respectively.

In the experiments we carried out, for all the periods and all the algorithms, none of the detected communities can be completely identified (following the aforementioned procedure) with inside China or outside China tweets. Indeed, setting a threshold of 5% to establish the higher relative frequency of the keywords for the inside or outside China tweets forming our networks, we get the following results:

1. All periods together (506 nodes):
    - 91% common keywords
    - 5% inside China keyword: 'china', 'trump', 'accord', 'medium', 'japan', 'hubei', 'south', 'beijing', 'symptom', 'hong', 'russia', 'read', 'korea', 'germany', 'asian', 'brazil', 'dos', 'york', 'india', 'france', 'johnson', 'africa', 'italy'
    - 4% outside China keywords: 'australia', 'diagnose', 'evacuate', 'british', 'employee', 'cancel', 'hundred', 'seek', 'california', 'fauci', 'canada', 'french', 'america', 'rule', 'mike', 'question', 'joe', 'republican', 'democratic', 'breaking', 'poll'
2. January-February 2020 (430 nodes):
    - 83% common keywords
    - 9% inside China keyword: 'china', 'novel', 'medium', 'japan', 'arrive', 'hubei', 'express', 'give', 'wang', 'donate', 'measure', 'south', 'dollar', 'beijing', 'accord', 'discharge', 'east', 'macao', 'taiwan', 'symptom', 'hong', 'month', 'cure', 'mother', 'tell', 'read', 'child', 'model', 'xinhua', 'wish', 'jinping', 'korea', 'clinical', 'suspect', 'leishenshan', 'construction', 'headline'
    - 8% outside China keywords: 'great', 'announce', 'asia', 'trump', 'australia', 'slow', 'diagnose', 'evacuate', 'grow', 'united', 'restriction', 'base', 'shanghai', 'cancel', 'hundred', 'sale', 'apple', 'thousand', 'japanese', 'record', 'wall', 'lunar', 'evacuee', 'california', 'diamond', 'princess', 'france', 'airport', 'singapore', 'thailand', 'australian', 'wash', 'asian', 'quarter', 'investor'
3. September-October 2020 (500 nodes):
    - 79% common keywords
    - 5% inside China keyword: 'italy', 'confirm', 'accord', 'africa', 'india', 'brazil', 'hour', 'china', 'trump', 'york', 'russia', 'qingdao', 'united', 'restriction', 'organization', 'medium', 'prime', 'twitter', 'white', 'walter', 'reed', 'joe', 'johnson'
    - 16% outside China keywords: 'france', 'sign', 'raise', 'fall', 'germany', 'australia', 'french', 'warn', 'child', 'update', 'fear', 'south', 'korea', 'slow', 'worker', 'researcher', 'claim', 'rule', 'employee', 'struggle', 'democratic', 'stock', 'resident', 'breaking', 'sen', 'kamala', 'british', 'republican', 'mike', 'democrat', 'penny', 'america', 'congress', 'supporter', 'symptom', 'happen', 'speaker', 'nancy', 'amy', 'supreme', 'court', 'age', 'old', 'lose', 'seek', 'fauci', 'effort', 'thousand', 'already', 'senate', 'win', 'term', 'result', 'political', 'turn', 'poll', 'question', 'appear', 'catch', 'trail', 'adult', 'exclusive', 'senator', 'north', 'hick', 'pause', 'play', 'woman', 'expose', 'talk', 'voter', 'jersey', 'game', 'party', 'feel', 'pay', 'authorization', 'review', 'financial'
4. March-April 2021 (398 nodes):
    - 78% common keywords
    - 7% inside China keyword: 'brazil', 'accord', 'dos', 'restriction', 'authority', 'approve', 'organization', 'africa', 'china', 'south', 'base', 'india', 'michigan', 'prime', 'watch', 'confirm', 'institute', 'african', 'united', 'chile', 'medicine', 'york', 'johnson', 'yunnan', 'russia', 'animal'
    - 15% outside China keywords: 'bad', 'black', 'lose', 'child', 'passport', 'result', 'france', 'hospitalization', 'tokyo', 'seek', 'british', 'germany', 'hop', 'canada', 'urge', 'deliver', 'governor', 'ask', 'facebook', 'review', 'thailand', 'game', 'fund', 'contract', 'japan', 'benefit', 'win', 'expect', 'double', 'consider', 'student', 'joe', 'medium', 'french', 'add', 'service', 'cite', 'threaten', 'trump', 'update', 'union', 'scientist', 'speak', 'problem', 'suffer', 'researcher', 'korea', 'exclusive', 'investigation', 'war', 'australia', 'analysis', 'want', 'breaking', 'source', 'infectious', 'fauci', 'decision', 'coverage'

Even if the difference in the frequencies slightly increases as we go from the earlier to the later pandemic period, almost all (from 78% to 91%) the keywords in our networks are equally frequent in the tweets inside China and outside China. This clearly lows the possibility of success of the algorithms to identify communities that entirely belongs to one of the two kinds of tweets. Moreover, the keywords grouped and listed before don't seem to form any interpretable or relevant community and therefore we can deduce that the inside and outside China news channel accounts report similar topics in their tweets. This is

conformed by the fact that, more or less, for each one of the main communities (not considering the very small ones and the non-meaningful ones) detected by each algorithm in each period, the percentage of keywords that is common to the inside and outside China tweets doesn't go lower than 65% and the remaining keywords are almost evenly distributed among inside China and outside China tweets (according to their relative frequency).

4.4. Second setting: inside and outside China separated

Since, in general, the communities analysed in the previous setting resulted to not be useful to distinguish the inside China tweets from the outside China tweets, in this new setting we want to compare the community detection directly performed on the inside China tweets' networks and the one performed on the inside China tweet's networks. In this case, for simplicity, we only consider the Louvain method applied on small-size networks.

4.4.1 Semantic interpretation

The Figure 38 displays the communities detected on the inside China and outside China networks, both referring to all the three selected periods together.

However, we will discuss in details just the three single periods.

January-February 2020

The Figure 39 displays the communities detected on the inside China and outside China networks, both referring to the January-February 2020 period.

This first period is a period that could be summarised by '*report*', '*case*', '*hospital*' and '*China*'. According to Figure 39, most of the words are regarding the description or report of the pandemic situation, the case increased, the novel coronavirus, these themes appeared both inside and outside China but with different themes.

By figuring out the communities inside China, we can observe there are 5 communities. The main theme in this period inside China is a large number of reports of the novel virus and cases, the confidence to control this bad situation. Meanwhile, the 6 communities outside China are not only focused on the reporting situation but also the impact of coronavirus.

Comparing the purple communities inside and outside China, except for the '*China*', '*outbreak*', the nodes inside china present words in positive and hopeful sentiment, for example, '*effort*' '*battle*', '*solidarity*', '*confident*', '*support*' '*win*' and '*expert*'. In these tweets published by Chinese official media, it shows not only the effect, the bad situation, the number increased but also the expressed Chinese government gratitude to the public, to all over the world their confidence to control the medical emergency and comfort who is involved and concerned. This confidence shows in the public health response, an adequate health workforce were mobilized and deployed to a new temporary hospital. The Leishenshan Hospital increased access to care during the surge in Covid-19 infections, facilitated timely treatment, and transferred Covid-19 patients between GWs and ICUs within the hospital, all of which are potential contributors to lowering the CFR. Patients in the ICU experienced a much higher CFR and a greater burden of health care cost than those in GW [21]. At the same time, the WHO has named the disease Covid-19, short for "coronavirus disease 2019."

However, in the purple community outside China, there are not only the messages about coronavirus, also the main theme of the results can be conceptualised as: reporting tough, panic situations. We can find some different expressions: excessive and oriented, for example, '*risk*', '*warn*', '*deadly*', '*illness*', '*fear*'. Based on this, outbreak managers can strategies communication by sharing messages that respond to public concerns and feelings especially fear [22]. In return, it also caused public concerns, panic and xenophobia. There are some words regarding the economy, '*flight*', '*company*', '*business*'. The impact on the economy has just appeared. Due to the infectivity of coronavirus, China has controlled and reduced the number of international flights.

September-October 2020

The Figure 40 displays the communities detected on the inside China and outside China networks, both referring to the September-October 2020 period.

In this period, the number of deaths worldwide exceeded one million. We ran community detection separately. There are four communities within the tweets sent by the Chinese official news account.

The first community (Figure 40a, purple) contains many words. Most of the words are about the report of the Covid-19 situation at the time, both at home and abroad since there are words like '*report*', '*case*', '*new*', '*death*, '*health*'. This community also highlights that the number of dead surpasses a million at that time. The second community (Figure 40a, orange) in this period is about the economic situation since there are words like '*economy*', '*recovery*'. The third community (Figure 40a, blue) may concern with Africa's situation, and the last community (Figure 40a, green) is more about politics. There are words like '*Trump*', '*minister*, '*president*', and so on.

However, the topic of communities of the official account outside China may be different. We detected three communities. The first community (Figure 40b, orange) here is about the report of the situation of Covid-19. The

second one (Figure 40b, purple) is more about the political situation during the pandemic. There are many words about the president and campaign. The last community (Figure 40b, green) may be related to the vaccines since there are words like '*Johnson*', '*adult*'.

From the community detection, we can see that both news accounts inside and outside of China concern the Covid-19 situation and the politics. However, there are also some differences between them. The topic of economics is appeared in the inside China community and not in the outside China community. Moreover, within this community, we can also find the words like '*tourism*', '*holiday*', '*flight*'. There was a National Economic Performance Launch in this period. According to Fu Linghui, spokesperson of the National Bureau of Statistics, "China's economy turned positive for the first time in terms of cumulative growth, and there was great interest in the economic performance in October. Judging from the situation in October, the national economy continued its stable recovery." (*The Information Office of the State Council held a briefing on the operation of the national economy in October 2020.*) From this community, we can see that the economic situation was a hot topic at that time.

On the other hand, the topic of vaccines appears in the community outside China but not in the community inside China. At this stage, in September, the Johnson & Johnson vaccine entered the final phase of clinical trials. However, in October, Johnson Johnson's vaccine development was called off because of unexplained symptoms in vaccine recipients. So at this stage, the vaccine was also a hot topic.

However, although they both focus on the Covid-19 situation and politics, there are differences between them inside China communities. The Covid-19 situation is the biggest community within all the communities, while the biggest community of outside China communities is the one about politics. From this, we can see that although both sides have the topic of politics and the Covid-19 situation, the Chinese media focus on the Covid-19 situation with the keywords '*case*', '*report*', '*new*' and '*health*', this may be because of the serious Covid-19 situation around the world. In contrast, the foreign media focus on the political situation under the pandemic with the keywords '*Trump*', '*say*', '*president*' and '*test*', this may be because September/October 2020 falls in the middle of the US elections.

March-April 2021

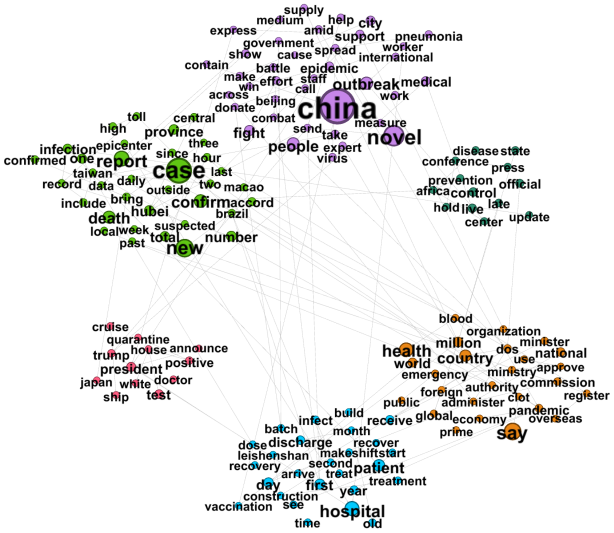
The Figure 41 displays the communities detected on the inside China and outside China networks, both referring to the March-April 2021 period.

In this period the debate was mainly about the vaccination and of course the different variants of vaccines available along with the number of cases.

We can see that there are five communities within the tweets by Chinese official news accounts along with five other communities which belong to tweets by the news accounts outside China. Each community is distinguished by a specific colour.

Going back to those inside China the biggest communities are identified by green, purple and blue. By looking at the green one we can see two keywords are highlighted, '*Health*' and '*Milion*'. There's nothing very surprising here since people's health is the main argument when we are talking about Covid-19. On the other hand '*Milion*' probably refers to the number of cases as well as the high population within China. In the purple community, we can see '*Report*', '*Case*' and '*New*' are highlighted. It's much clearer here that they are referring to the number of new cases which are being reported in this period. However, in the blue community, we can see that the discussion is mainly focused on China and the pandemic situation there since there is '*China*' as the most highlighted keyword followed by '*Country*', other keywords are emergency, arrive, origin, etc. Other two smaller communities (dark green and orange) mainly show the significance of vaccination by highlighting keywords such as '*Receive*', '*First*' and '*Say*'. we can interpret receive and first to the first dosage of vaccines received by people inside China.

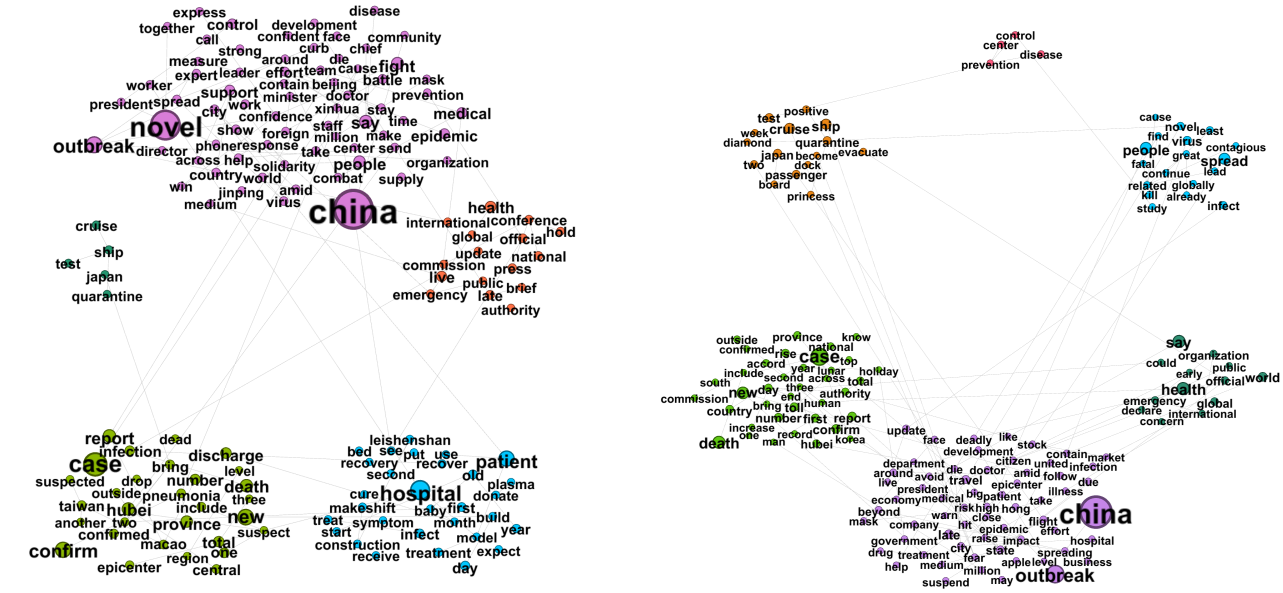
On the other side there are five other communities within tweets by official news accounts outside China, on the red community we can see the sign of vaccination procedure since the highlighted keywords are '*Johnson*', '*Blood*' and '*Emergency*'. We can link them to the vaccination outside China since Johnson is a brand name of a Covid-19 vaccine which is mainly used among western countries. On the blue community again the most highlighted keyword is '*Say*' same as the orange community inside China, we can relate it to different reports regarding the pandemic and various claims in this regard. The other three communities(light green, purple and dark green) are also emphasising the vaccination phase as we can see bold words such as '*Shot*', '*Vaccinate*', '*People*' and '*Case*' however interestingly there are some new bold words mentioning different countries such as '*India*', '*Brazil*' and '*China*'. These country keywords are most likely emphasising the differences in these countries regarding handling the pandemic situation since these three countries are quite different in this regard.



(a) Inside China.

(b) Outside China.

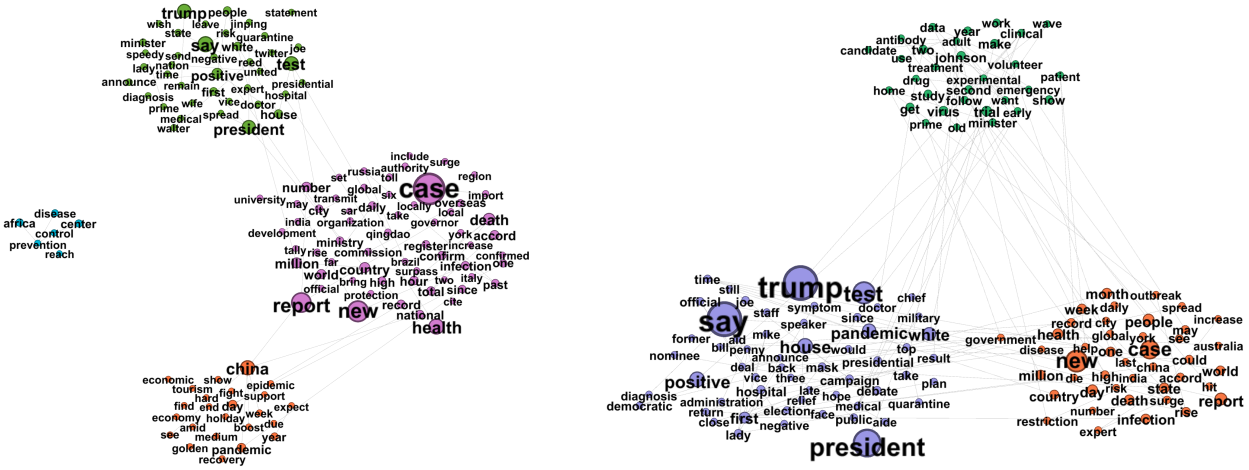
Figure 38: Louvain method detected communities for all the three periods together.



(a) Inside China.

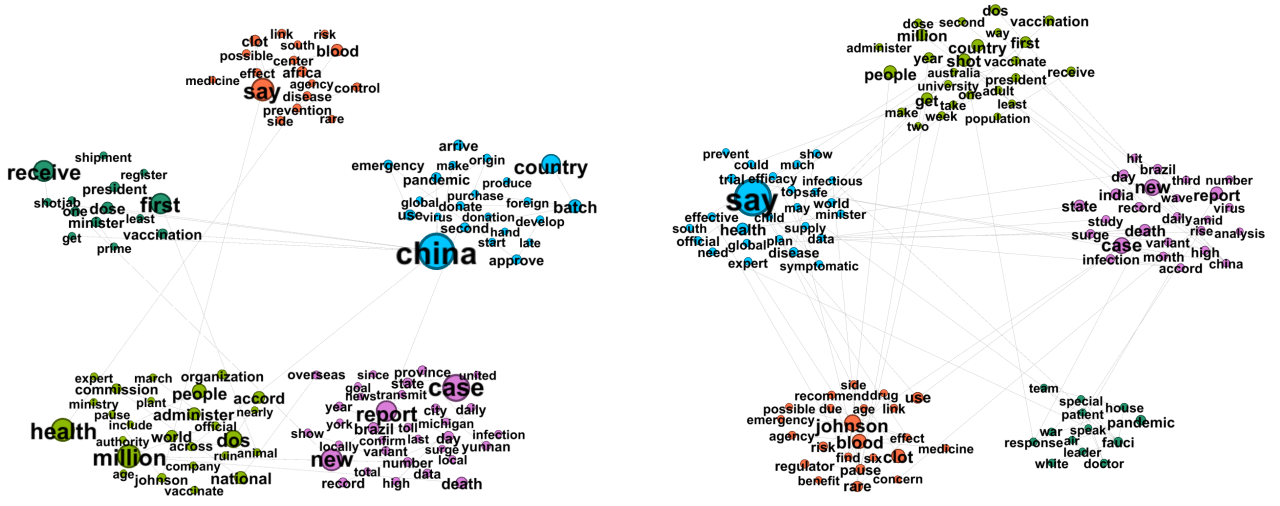
(b) Outside China.

Figure 39: Louvain method detected communities for the January-February 2020 period.



(a) Inside China. (b) Outside China.

Figure 40: Louvain method detected communities for the September-October 2020 period.



(a) Inside China. (b) Outside China.

Figure 41: Louvain method detected communities for the March-April 2021 period.

5. Sentiment Analysis

In this section we will investigate the differences in the medias discussion(inside and outside China) based on the analysis of sentiment. We will introduce the method used to project the results of the LIWK software. We will compare the results of LIWC on our dataset with the reference one and we will spot the difference in values of the markers tweets over the different time periods. Afterwards we will compare through linear regression different marker. In conclusion we will focus on the Negative emotions marker and visualize the results.

5.1. Methods

5.1.1 LIWC analysis

To perform this kind of analysis we exploited LIWC (Linguistic Inquiry and Word Count), a text analysis application capable of effectively and efficiently measuring the structural, emotional and cognitive components present in written speech samples.
In particular, we fed the entire corpus of collected tweets into LIWC2015, the latest version of the application, and analysed the output, taking into account some of the variables provided as output by the algorithm. The variables considered in this study are listed below, with a brief explanation of the information they bring [23].

- **Analytical thinking**: a high number indicates formal and logical thinking; lower numbers indicate more informal and personal thinking.
- **Clout**: A high number is a sign of high competence and confidence; low numbers indicate a more hesitant and humble style.
- **Authentic**: higher numbers are associated with a more honest and personal text; lower numbers suggest a more cautious and detached form of speech.
- **Emotional tone**: a high number indicates a more positive and optimistic style; a low number indicates more anxiety, sadness or hostility.
- **Negative emotions**: the higher this value, the more the text conveys negative emotions, including anxiety, anger and sadness. This category is referred to as *'negemo'*.
- **Positive emotions**: a high number indicates a large presence in the text of words conveying positive emotions. This category is referred to as *'posemo'*.
- **Focus present**:indicates the presence of words with a temporal orientation focused on the present.
- **Focus future**:indicates the presence of words with a temporal orientation focused on the past.
- **Affiliation**: High values of this variable suggest the presence in the text of words such as *ally, friend, social*, indicating cohesion within a group.
- **Empowerment**: This variable is computed as the mean of the LIWC2015 variables *power, achieve, reward, insight and cause*. Previous studies show that aggregating these variables gives a good estimate of agency. A high value of this variable is related to intelligence, skill, creativity, achievement, power, mastery, and assertiveness while a low value indicates weakness, submissiveness and incompetence. The estimation of empowerment with this technique has been exploited effectively in [24].
- **Death**: high values indicate a large presence of words related to death.
- **We**: measures the content of pronouns and verb forms in the first person singular. This variable is important because a high number of pronouns such as 'we', 'us' and 'our' is an indicator of group identity and positive political propaganda [25].
- **They**: measures the content of pronouns and verb forms in the third person plural. In contrast to the variable *we*, a high value here is often an indicator of negative political propaganda, exclusion and detachment [25].

5.1.2 Projection on the single words

Since we are dealing with institutional tweet, the sentiment inside the data will not be explicit from single words used. Therefore to capture the sentiment given by the social disclosure of the tweets into the single words, we use a PageRank based approach.

Starting from the cleaned tweets, we construct the non normalised adjacency matrix of the bipartite network linking tweets to words $\mathcal{B}$. Afterwards we build the row-normalised and the column-normalised-and-transposed counterparts to $\mathcal{B}$, respectively $\mathcal{B}_r$ and $\mathcal{B}_T$. Let us call with $\widetilde{m}_t$ the initial LIWK marker tweets and with m_w the aimed marker of words.

The PageRank-inspired projection solve the following equation through the power iteration method, where $\alpha \in (0, 1)$ is a constant which regulates the information' spread.

$$\begin{aligned} m_t &= \alpha \mathcal{B}_T \mathcal{B}_r m_t + (1-\alpha)\widetilde{m}_t \\ m_w &= \mathcal{B}_r m_t \end{aligned}$$

Other two ways to capture the sentiment from the tweets LIWC marker have been tried. The first consists on associating to each word the summed score of the tweets in which it is contained. Since the latter approach is affected by the degree of each words, the second method consists of averaging the value of LIWC associated to the tweets containing them, which as we will see has very similar results to the projection one.

In conclusion to understand the structural relevance of the words with higher marker in the projected network, we repeat the node removal strategy of the Section 3.6, this time removing first the words with higher projected sentiment.

5.2. Results

5.2.1 Comparison with reference values

We compared the values of the variables assigned to our tweets with reference values found in [23] for two categories: articles published online by the New York Times and a collection of tweets gathered from public profiles. The decision to adopt these two categories as reference comes from the fact that the corpus of tweets used in this research comes from official news twitter accounts only. Therefore, it is expected that the values assigned by LIWC to the variables will be approximately between those of Twitter and newspaper articles. A comparison of the average values obtained on the collected tweets from Chinese news accounts, those from American news accounts and the two reference categories (Twitter and New York Times) is presented in Table 19.

	China	Outside China	NY Times	Twitter
Analytic	95.19	93.69	92.57	61.94
Clout	63.16	60.92	68.17	63.02
Authentic	28.27	20.88	24.84	50.39
Tone	34.31	32.28	43.61	72.24
Negemo	1.46	1.40	1.45	2.14
Posemo	1.41	1.10	2.32	5.48
Focus Past	1.73	1.54	4.09	2.81
Focus Present	5.03	4.22	5.14	11.74
Focus Future	0.58	0.67	0.80	1.70
We	0.14	0.30	0.38	0.74
They	0.27	0.19	0.68	0.47
Affiliation	1.03	0.92	1.69	2.53
Death	0.56	0.51	0.22	0.19

Table 19: Comparison between LIWC2015 Statistics [23] and the corpus LIWC2015 analysis.

Two of the values that deviate the most from the expected ones concern the *posemo* variable, which is lower than the reference and the *death* variable, which is much higher than the reference. This trend confirms the effectiveness of LIWC on our dataset, for which results of this type could easily be expected.

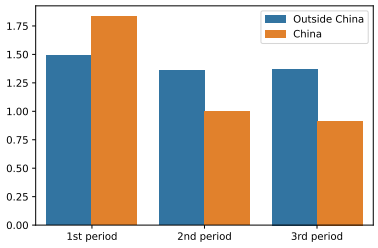
5.2.2 Analysis of the single time periods

Once the tweets were analysed as a whole, the next step was to look at the different periods separately, comparing the average value of the variables of tweets inside and outside China. Figure 42 shows this comparison from which similarities and differences between inside and outside China emerge.

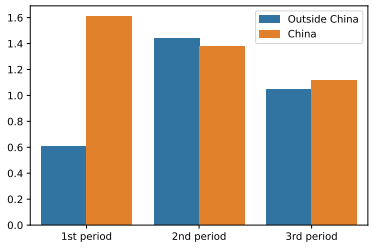
In both cases, the variables *negative emotions* (fig. 42a) and *death* (fig. 42f) are much higher in the first period and tend to decrease in the following two periods. This trend may be due to the desire to reassure the population after the first period, in which panic broke out all over the world also because of the excessive amount of alarmist news. In addition the chinese media tends to write tweets containing more numbers as shown in Figure 42k. On the contrary, the variables *empowerment* (fig. 42g), *positive emotions* (fig. 42b) and *affiliation* (fig. 42j) increase in the second two periods for tweets outside China, showing the desire to report the contingent situation as under control. These variables, for China's tweets, are already very high in the first period, demonstrating (apparent) positivity and control even in a very critical moment. The most substantial differences are in the variables *we* (fig. 42h) and *they* (fig. 42i). The value of *we* is much higher in tweets from outside China for all three periods. The value of *they*, on the other hand, is much higher in tweets from China in the first period and then stabilises in subsequent periods. This result will be discussed below.

In fig. 43 is shown the pie charts representing the percentage of more positive, negative or neutral tweets inside a period for both inside and outside China. Those values are computed from the Positive and Negatives emotions with the assumption that they are comparable. A tweet is considered neutral in case the value of Positive and Negative content is equal within a 0.1 accuracy. As one can see in the first period outside China the Positive Sentiment tweet number are drastically low, in correspondence with an increase of Negative one. Inside China we have similar scenario in which negative tweets are prevalent and double in size respect to the other periods. We will later analyze this situation.

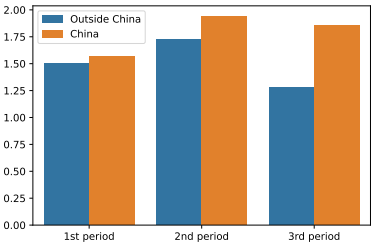
To further investigate the differences between tweets from inside and outside China, we analyzed the variables *we* and *they*, which indicate belonging and distancing respectively, by associating them with the variables *positive emotions* and *negative emotions*. Figures 44 and 45 show the linear regression (with confidence intervals) of *they* versus *negemo* and *we* versus *posemo*, respectively, first as a whole and then dividing the single periods. We observe that



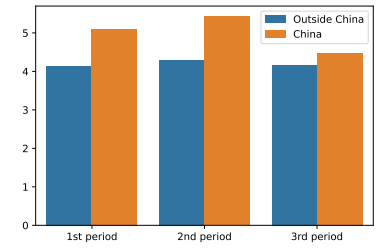
(a) Negative emotions.



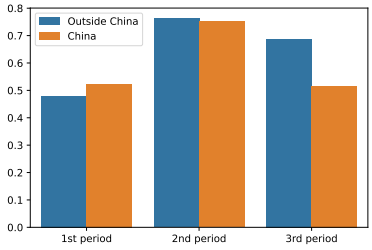
(b) Positive emotions.



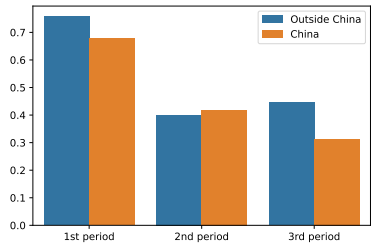
(c) Focus past.



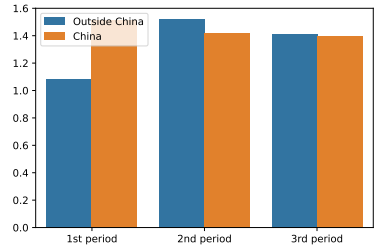
(d) Focus present.



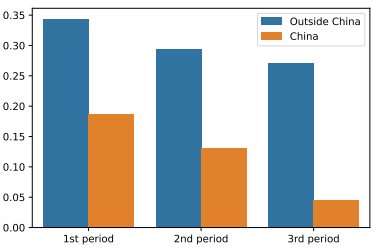
(e) Focus future.



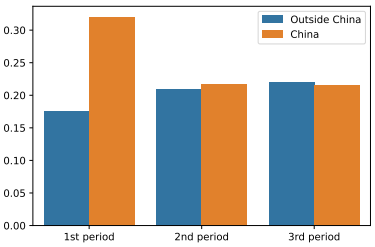
(f) Death.



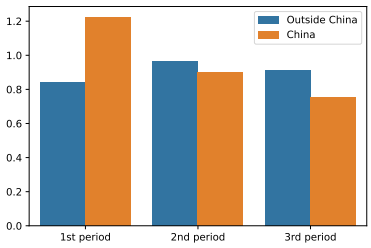
(g) Empowerment.



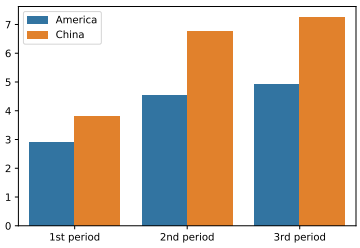
(h) We.



(i) They.



(j) Affiliation.



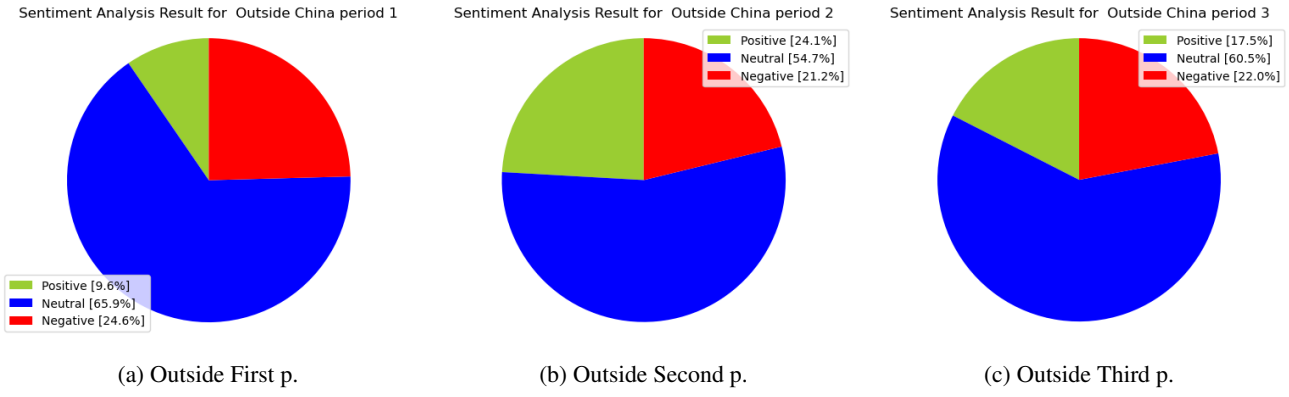
(k) Numbers.

Figure 42: Comparison of the LIWC variables between the periods.

in both graphs the slopes are almost always positive, suggesting an association between the two pairs of variables.

In particular, it can be seen by looking at Figure 44 that, considering the overall graph the *they-negemo* association appears much stronger inside China (considering the slope). By analyzing the single periods we can see that this is due to a particularly strong association in tweets inside China during the first period. Thus, this result shows that especially in the first considered period, in tweets from Chinese accounts the use of a higher number of words belonging to the category *they* is associated with a higher presence of words expressing *negative emotions*. This trend explains the spike in the *they* value visible in Figure 42i.

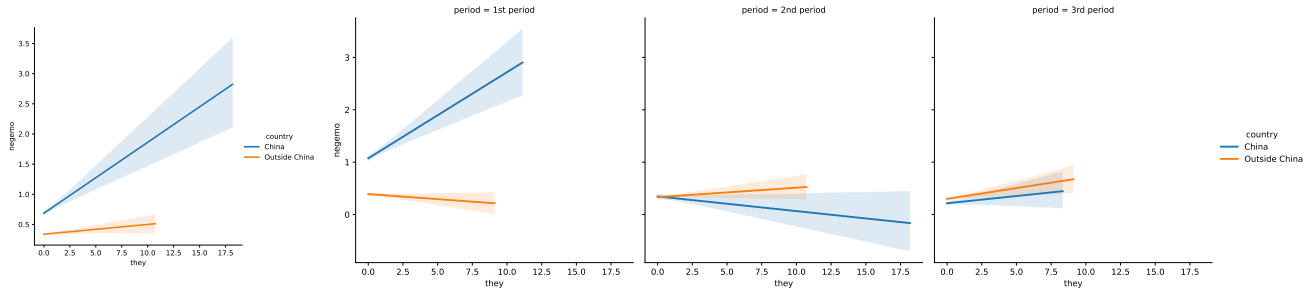
Figure 45 shows that the variable *we*, which indicates be-



(a) Outside First p. (b) Outside Second p. (c) Outside Third p.

(d) Inside First p. (e) Inside Second p. (f) Inside Third p.

Figure 43: Pie chart with positive-negative-neutral prevalent tweets percentage content during the different periods inside and outside China.



(a) Overall comparison.

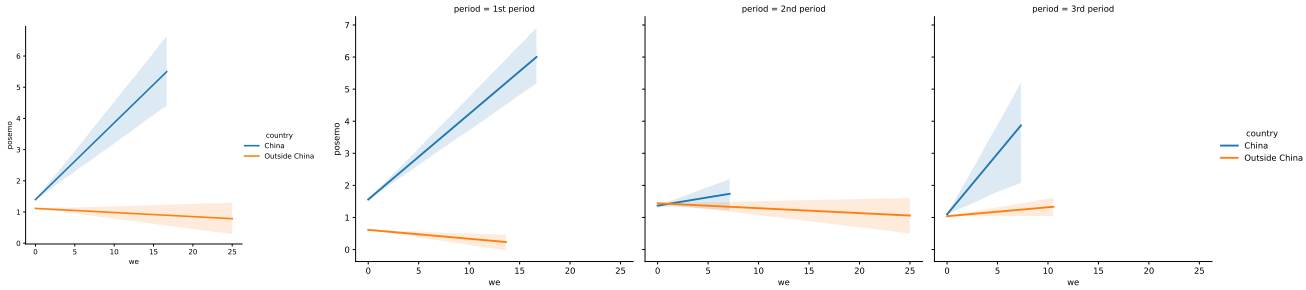
(b) Single time periods comparison.

Figure 44: Linear regression with confidence intervals for they versus negative emotions.

longing, is strongly associated with the variable *posemo* in tweets from inside China. This association is more evident in the first and third periods and can be appreciated also from the overall diagram. For tweets from outside China, however, the association is much less pronounced: the slope is near to zero (blue regression). This result seems to be in contrast to what is depicted in Figure 42h in which we can appreciate much higher *we* values for tweets outside China. However, it was not possible with the linear regression technique to motivate this specific trend.

5.2.3 LIWC words marker

In order to solve the equation in Section 5.1.2, we use $\alpha = 0.85$ and 20 iteration. Aiming to compare qualitatively the goodness of the results, we associate to each words in the network the sum of the tweets marker containing it, m_w^*. In Figure 46 we showed the 20 obtained words with higher projected Negative emotions in China during the first period on the left and the 20 words with higher m_w^* on the right. As one can notice using the projection method words which usually does not carry high negative feelings shows



(a) Overall comparison.

(b) Single time periods comparison.

Figure 45: Linear regression with confidence intervals for we versus positive emotions.

up. For instance the words "*bat*", "*soup*", "*goodwill*". Those are all related with tweets that clearly aim to defend the China from the "*echoing*" of being the source of Covid-19, "*criticizing*" the "*unfriendly*" U.S comments as not a sign of "*goodwill*". The marker obtained through summing the scores is not optimal since words with higher degree appear in more tweets therefore they get higher score. For this reason we compare the projection results with the markers obtained by the association of each word with the average score it obtains in the tweets LIWK result. We show it in the case of outside China during the first period, Figure 47a. As we can see the two methods gets the same ranking, in fact what we are doing through the PageRank projection is similar to average the marker value of each word through the network.

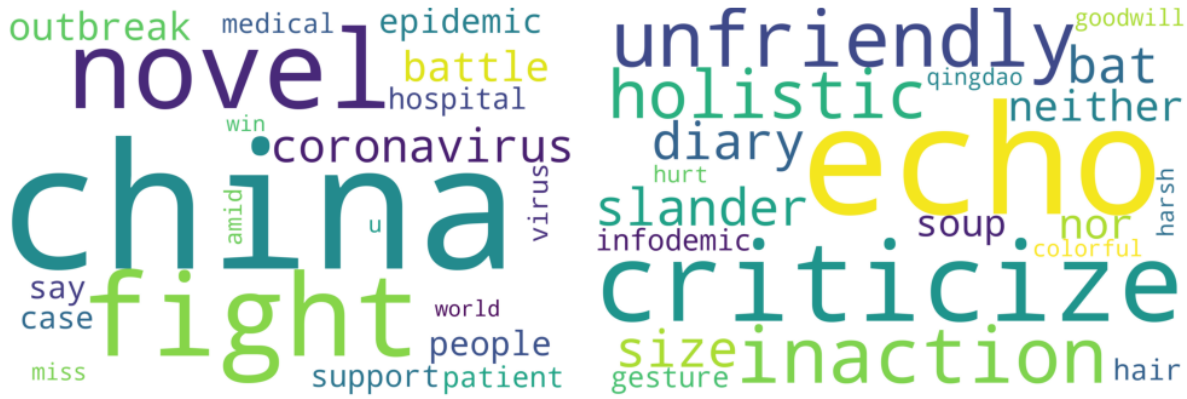
5.2.4 Projected marker attack

In order to understand the structural relevance of the words with higher projected sentiment on the network and to see if they depends on the times it appear on the tweets(degree), we implement an attack strategy as in Section 3.6. So starting with the co-occurency network we delete the words with higher sentiment keeping track of the giant component and components number. Here we show the results for China and America in the first period obtained with the markers Negative and Positive emotion (Figure 48). As one can notice from the graphs the network is more robust to the sentiment projection results. During the first part of the attack the behaviour is similar to a random failure. Therefore we can say that the obtained results does not depends much on the degree centrality of each words.

5.2.5 LIWC projection on words

After having developed a method to project the values of the LIWK variables on individual words, we continued the investigation by analysing the results obtained with this technique. First of all, we built a network where nodes are words and edges are links connecting two words that appear together in the same tweet. The weights of nodes are the values assigned by our algorithm to the words according to a specific variable. Then, we represented the network obtained using Gephi: we noticed that by separating the communities and highlighting the nodes with a higher weight, it is possible to identify the different news topics that contain words with a higher value of the considered variable.
The outcome of the procedure just described with the variable *negemo* is reported below.
Figures 49, 50 and 51 compare the networks obtained by considering tweets inside and outside China separately, for each of the three selected periods. The colour of each node represents the community it belongs to, while the size is directly proportional to the value assigned to each word (node) by the projection algorithm described in Section 5.1.2.

For the first period, looking at tweets from Chinese news accounts (Figure 49a), we see in the biggest (purple) community words such as *'unfriendly'*, *'bat'*, *'criticize'*, *'inaction'*, all of which are contained in tweets complaining about the behaviour of America, which spreads fake news and blames China for the pandemic. The orange community also reports Asian people being discriminated against and accused of carrying the virus to Europe.
The community (green) with the words *crazy, mourn* and *slander* contains words related to news that both describe as absurd the theory that the virus was created by China and mourn doctor Li Wenliang, a doctor who passed away after being infected with the coronavirus.
For tweets outside China (Figure 49b), the word with the highest value is *'flyer'*. When analyzing the news, we notice that this word often appears in news reporting that when the pandemic broke out, flyers were distributed in California urging people not to go to Asian restaurants, blaming them for spreading the virus. This word belongs to a community (the purple one) in which also the other most important words are used in the same news or in news with different topics, all negatively related to China. The other



(a) Words with higher summed tweets negative emotions.

(b) Words-cloud projected negative emotions.

Figure 46: Negative emotion word clouds of the January-February 2020 period in China.

(a) Words with higher averaged tweets negative emotions.

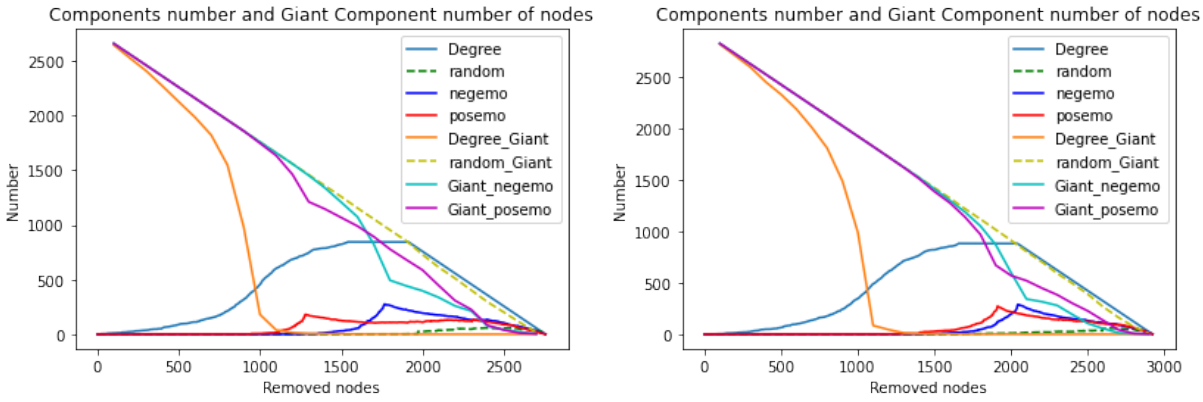
(b) Words-cloud projected negative emotions.

Figure 47: Negative emotion word clouds of the January-February 2020 period outside China.

most important community is the one with words *silenced, shameful, outspoken* and *squabble*. These words refers to news about the death of doctor Li Wengliang and what happened to some people escaping from Wuhan. In this case, however, news reports that the doctor died in suspicious circumstances, after being censored for being one of the first to warn of the coronavirus. In this case, the news accounts report the same news inside and outside China but with a completely different tone.

Considering the second period, we observe for the tweets outside China (Figure 50b) , communities represented by the words *wage, gang, lonely* (grey) or *misery, trauma, tragedy* (black). They are related to news reporting the distress of the various nations forced to suffer the economic and social effects of the pandemic. Again, the community with the words *distress, psychological, treating* (dark green) refers to the negative psychological effects on the population. With regard to tweets from Chinese accounts (Figure 50b), however, the principal communities contain words mainly related to news reporting the suffering of other nations, and only rarely news reporting the situation in the country. For example, the community with words *defeat, difficulty, stress, terrible* (purple) is linked to news reporting the difficulties of the USA during the peak of the infection, while the community with the words *scar, actively and devastating* is linked to news reporting the critical situation in Africa.

As regards the third period, in tweets outside China (Figure 51b) there are still communities linked to news expressing the negative effects of the situation: *battleground, turmoil, crematorium, assault* (purple), *impending, doom, implore* (dark green), *depression, psychosis, dementia* (orange). Alongside communities of this type, also recognizable in previous periods, there are others related to vaccines such as *rage, discourage, aggravate, countermeasure, dismay* (green), in which are concentrated words related to



(a) Outside China.

(b) Inside China.

Figure 48: Node removal by higher sentiment in China's in the January-February 2020 period.

(a) Inside China.

(b) Outside China.

Figure 49: Negative emotions projection on words, January-February 2020.

news about misinformation against vaccines or *manslaughter, dominant, consistently* (blue) related to the topic of pressure on hospitals.

For tweets in China (Figure 51a), confirming the trend of previous periods, there are no communities linked to news about the internal pandemic situation in the country. The community with the words *victimize, violation, killing, increasingly* (purple) is linked to the topic of discrimination against Asian people in America and Europe. The community with the words *impend, doom, hesitancy, avoid* (blue) is linked to news about the hesitation to vaccination by the population of some nations and the gravity of the situation in the USA. Finally, the community with the words *fake, counterfeit, regrettable* (green) refers to news about the spread of fake vaccines.

(a) Inside China.

(b) Outside China.

Figure 50: Negative emotions projection on words, September-October 2020.

(a) Inside China.

(b) Outside China.

Figure 51: Negative emotions projection on words, March-April 2021.

6. Discussion

January-February 2020

In the first period, accompanied by the increasing reported cases, the damage and tragedies happened in China, the panic and the xenophobia emotion are immersed. In this phrase, the media coverage is not only focused on the case number but also the effect and issues that happened.

The main idea of the outside China official media in this period is talking about the stress. In the sentiment analysis, we can observed that while the outside china media are reporting that the a flyer (Figure 49b, purple) bring the discrimination to Asian people and the law against the fake news about coronavirus in south Asia should be reconsidered, it also means that the panic emotion of public causing the fake news transmitted. However, Inside China media are against the fake news, criticizing the U.S unfriendly comments, denying the coronavirus is caused by the bat soup, against the discrimination and racism.(Figure 49a purple)

At the same time, the outside china news media are also emphasizing the symptoms (Figure 49b, black) and the imminent threat to public health. Inside the china news media with a mitigated situation, they show the confidence to control this health emergency by citing a cartoon(Figure 49b, purple), which is informative and attractive video that can be spread fast as an efficient communication strategy. The death of Dr.Li Wenliang, who raised the alarm about the coronavirus outbreak but initially was investigated for 'spreading rumors', stimulated the public anger to local government. The media inside China express a sad emotion with 'mourn' to this issue, avoid to talking about the people's anger and upset. However, the media outside China shows not only the 'mourn' but also the 'outspoken', 'silenced' to this issue, talking about the rights of speaking under a strict discipline circumstance.

September-October 2020

From the community detection, we know that the official news account inside and outside China focus on different topics in this period. This may be because of the different situations inside and outside of China. In the second phase, the situation in China changed for the better, with a reduction in daily additions. So, media coverage may be more inclined to cover topics such as economic recovery and the world epidemic situation. However, at the same time, because of the seriousness of the world epidemic, the greatest amount of communication about the Covid-19 situation was seen in this phase.

In the outside China official account, it is evident that the topics covered are more on vaccines and politics. The US elections in September/October 2020 was a significant issue for the President and the government in response to the Covid-19 pandemic.

The sentiment analysis shows that the negative emotion is expressed in different aspects inside China and outside China's official news account. In this period, we can see that the official news accounts outside of China express their concerns that the virus can trigger heart conditions due to inflammation (Figure 50b, blue). Moreover, they also show the negative emotions on the wage and unemployment (Figure 50b, grey). In addition, the effects of Covid-19 on the elderly (Figure 50b, red) and the work situation and stress of nurses and doctors (Figure 50b, dark green).

On the other hand, the official news account inside China expresses the negative emotions concerning the Covid-19 situation in the US, Africa, and the world (Figure 50a, purple, blue, grey). What is more, the concerns about the relationship between countries. And they are helping Africa during the Covid-19 situation (50a, black).

We can see that in this period, the official news account outside of China is more focused on the negative effects brought by the Covid-19. Moreover, the official news accounts inside China are more focused on the Covid-19 situation worldwide. It might be due to the fact that in this period, the number of deaths worldwide exceeded one million, so the news outside of China may focus more on the consequences brought by Covid-19. In China, however, there were fewer than 50 new Covid-19 cases per day in September/October of 2020, so the news about Covid-19 may be more focused on the world's situation.

At this period, we can conclude that because the situation is different in China and abroad, the focus of official media coverage on Twitter differs between China and abroad. At the same time, they both expressed negative feelings towards Covid-19, also because the situation of the epidemic was different and their positions were different, and their focus was different. In this period, the number of deaths worldwide due to Covid-19 has reached one million, so both the Chinese and foreign media have shown their concern for the situation in the world.

March-April 2021

As we have seen on Figure 41, there are obviously some differences between the highlighted tweets coming from news accounts inside China and those from other countries. However of course there are also some significant similarities between them owing to the fact that after all they all belonged to the same period which was highly debated as the vaccination period.

We have seen the different signs of this phase in both outside and inside communities emphasising on keywords such as 'receive', 'dose' and 'johnson'. As for the differences we can mention that the media inside China were emphasising mostly on domestic situations rather than focusing on for-

eign conditions, for this reason we can once again underline that one of the most highlighted keywords from inside media was mentioning China, which is not quite surprising considering the political situation and media censorship within this country.

On the other hand, official news outside China tended to emphasise more on the global situation regarding the pandemic as we have seen in the outside communities the significance of 'Reports' and covering the global situation regarding handling the pandemic as they were covering other countries such as India and Brazil along with going on with the vaccination phase.

7. Conclusions

From the keywords of different time periods, we can see that there are always differences in the focus of Chinese and foreign media reports in the same time period. After the sentiment analysis, we also can see the differences in emotion these news accounts want to express to the public, which influenced the public mood, brought changes in real societies, then reflected by these accounts' tweets. This cycle continues, and the impact of the media on society during the epidemic has gradually deepened.

At the beginning of Covid-19, close attention to the epidemic in China is a common feature of Chinese and foreign accounts. Even though panic broke out all over the world at the beginning, Chinese accounts showed much higher positive emotions than outside China accounts. Outside China accounts accused the China government, blaming China for the outbreak and spreading the virus. However, Chinese accounts showed great confidence and hardly no criticism in fighting the epidemic, denied the accusations against China and paid more attention to the discrimination against Chinese in abroad, which may be largely influenced by outside accounts.

In the last two periods, negative emotions of outside China accounts were higher than Chinese accounts, this is related to the different focus of their reporting. During the Sep and Oct 2020, the negative global impact of the Covid-19 on the economy and society has received extensive attention from both Chinese and foreign accounts. Trump's expression and U.S. election became another theme amid Covid-19 for both inside and outside of China accounts. For Chinese accounts, the recovery of domestic economic and people's lives became the main aspects of the positive report.

In March and April 2021, vaccine-related topics were the focus of Chinese and outside China accounts reports. Chinese accounts reported more on the increase in domestic vaccinations, while accounts outside of China questioned the effectiveness of Chinese vaccines.

In conclusion, the differences in reporting between Chinese and foreign accounts throughout the epidemic, which to a certain extent resulted in a biased opinion between audiences in China and other countries. From the number of followers, reposts and comments, it can be seen that the influence of Chinese accounts is relatively weaker than others, and critical reports on China dominated the early days of the epidemic, which to a certain extent deepened the contradictions and hatred against Asians. With the widespread of Covid-19, the focus of these accounts shifted from China to the world, and this contradiction has been alleviated, but still exists.

8. Division of the work

- Alavi Seyedhamidreza: [Section 1] background, [Section 4] interpretation of community detection in the second setting of period 3, [Section 6] discussion of period 3
- Bigarella Chiara: [Section 2] data collection, [Section 4] community detection, Clique Percolation, partition metrics, semantic interpretation of the communities in the first setting
- Cogato Matteo: [Section 2] data collection, text extraction from tweets, [Section 3] centrality measures (PageRank, betweenness, closeness) and network visualizations in Gephi, PageRank vs HITS and other correlations, remarks on the exploratory analysis
- Gicquel Thomas: [Section 2] write part of the "Data collection" subsection in the report, [Section 3] tweets ranking according to likes and number of retweets, [Section 4] community visualization (Gephi)
- Liang Yiling: [Section 1] literature review of the topic, [Section 4] interpretation of community detection in the second setting of period 2, [Section 6] discussion of period 2
- Liu Yichen: [Section 1] methodology, [Section 7] conclusions
- Mellino Daniele: [Section 2] text extraction from tweets, [Section 5] LIWC sentiment analysis, [Section 3.6] robustness by node removal degree-betweenness-closeness
- Poletti Silvia: [Section 2] data collection, data aggregation, tweets cleaning (just one little modification to the final code), [Section 4] community detection, community visualization (Python algorithm), Dendrograms, community frequency-based analysis, semantic interpretation of the communities in the first setting
- Rayri Akram: [Section 1] write the "Python libraries" subsection in the report, [Section 2] data collection, write part of the "Data collection" subsection in the report,tweets cleaning and text extraction for top 10 liked and retweeted tweets inside and outside China.
- Trolese Francesco: [Section 2] tweets cleaning (NLTK), [Section 5] LIWC sentiment analysis, LIWK sentiment projection on words and visualization
- Yang Jingwen: [Section 1] introduction (research question), [Section 4] interpretation of communitiy dedection in the second setting of period 1, [Section 6] discussion of period 1
- Zangrando Emanuele: [Section 2] text extraction from tweets, tweets cleaning, [Section 3] keywords importance according to frequency, PageRank, HITS and TF-IDF, centrality measures,centrality measures correlation, Amen modeling and link strength prediction, conductance robustness.

9. Source Code

The source code of our work can be found at: `https://github.com/ChiaraBi/covid-tweets`.

References

[1] D. Ammar, "Racism and Xenophobia Towards China and People with Chinese Ethnicity Following COVID-19: A Content Analysis of Replies to Donald Trump's Controversial Tweets," vol. 2, no. 1. [Online]. Available: https://journals.macewan.ca/crossingborders/article/view/1986

[2] L. Garrett, "COVID-19: The medium is the message," vol. 395, no. 10228, pp. 942–943.

[3] F. M. Rodrigues de Andrade, T. B. Barreto, A. Herrera-Feligreras, A. Ugolini, and Y.-T. Lu, "Twitter in Brazil: Discourses on China in times of coronavirus," vol. 3, no. 1, p. 100118. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590291121000140

[4] H. Budhwani and R. Sun, "Creating covid-19 stigma by referencing the novel coronavirus as the "chinese virus" on twitter: Quantitative analysis of social media data," *J Med Internet Res*, vol. 22, no. 5, p. e19301, May 2020.

[5] T. Da and L. Yang, "Local covid-19 severity and social media responses: Evidence from china," *IEEE Access*, vol. 8, pp. 204 684–204 694, 2020.

[6] D. Wang and Y. Qian, "Echo chamber effect in rumor rebuttal discussions about covid-19 in china: Social media content and network analysis study," *J Med Internet Res*, vol. 23, no. 3, p. e27009, Mar 2021.

[7] #COVID19: Social media both a blessing and a curse during coronavirus pandemic - Department of Sociology. [Online]. Available: https://www.yorku.ca/laps/soci/engagement/covid19-social-media-both-a-blessing-and-a-curse-during-coronavirus-pandemic/

[8] R. W. Orttung and E. Nelson, "Russia Today's strategy and effectiveness on YouTube," vol. 35, no. 2, pp. 77–92.

[9] K. Rebello, C. Schwieter, M. Schliebs, K. Joynes-Burgess, M. Elswah, J. Bright, and P. N. Howard, "Covid-19 News and Information from State-Backed Outlets Targeting French, German and Spanish-Speaking Social Media Users," p. 8, 2020.

[10] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[11] S. Minhas, P. D. Hoff, and M. D. Ward, "Inferential approaches for network analysis: Amen for latent factor models," *Political Analysis*, vol. 27, no. 2, p. 208–222, 2019.

[12] E. Kolaczyk and G. Csrdi, *Statistical Analysis of Network Data with R*. Springer Publishing Company, Incorporated, 2014.

[13] P. Hoff, B. Fosdick, and A. Volfovsky, *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*, 2020, r package version 1.4.4. [Online]. Available: https://CRAN.R-project.org/package=amen

[14] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: https://igraph.org

[15] C. T. Butts, *sna: Tools for Social Network Analysis*, 2020, r package version 2.6. [Online]. Available: https://CRAN.R-project.org/package=sna

[16] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, 1970.

[17] J. Yang and J. Leskovec, "Overlapping community detection at scale: A nonnegative matrix factorization approach," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, 2013, p. 587–596.

[18] E. e. a. Nakazawa, "Chronology of covid-19 cases on the diamond princess cruise ship and ethical considerations: A report from japan." *Disaster medicine and public health preparedness*, vol. 14, no. 4, pp. 506–513, 2020.

[19] "Transcript: Anthony fauci on face the nation, march 28, 2021," https://www.cbsnews.com/news/transcript-anthony-fauci-on-face-the-nation-march-28-2021/.

[20] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, p. 75–174, 2010.

[21] Y. Cai, Y. Chen, L. Xiao, S. Khor, T. Liu, Y. Han, Y. Yuan, L. Cai, G. Zeng, and X. Wang, "The health and economic impact of constructing temporary field hospitals to meet the COVID-19 pandemic surge: Wuhan Leishenshan Hospital in China as a case study," vol. 11, p. 05023.

[22] M. B. Karmegam D, "What people share about the covid-19 outbreak on twitter? an exploratory analysis." 2020.

[23] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[24] C. Suitner, L. Badia, D. Clementel, L. Iacovissi, M. Migliorini, B. G. S. Casara, D. Solimini, M. Formanowicz, and T. Erseghe, "The rise of # climateaction in the time of the fridaysforfuture movement: a semantic network analysis."

[25] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

A. Appendix

Clique Percolation community detection

We report below the list of the keywords belonging to the overlapping communities detected by the Clique Percolation algorithm on three networks. Each network refers to a specific period of time and has been built after the removal of the search keywords. Furthermore, for each network we also highlight the overlapping nodes.

January-February 2020, k=7

- **Community 1:** ['around', 'make', 'princess', 'global', 'epicenter', 'dock', 'director', 'cruise', 'evacuate', 'high', 'include', 'rate', 'lead', 'late', 'infection', 'world', 'far', 'passenger', 'end', 'human', 'local', 'support', 'pneumonia', 'epidemic', 'hospital', 'case', 'city', 'prevention', 'state', 'macao', 'brief', 'commission', 'build', 'resident', 'test', 'measure', 'face', 'already', 'international', 'video', 'united', 'total', 'virus', 'three', 'hong', 'quarantine', 'center', 'mask', 'contain', 'public', 'take', 'find', 'country', 'great', 'cause', 'expert', 'diamond', 'confidence', 'japanese', 'declare', 'accord', 'chief', 'fear', 'send', 'live', 'could', 'effort', 'die', 'home', 'level', 'see', 'death', 'staff', 'medical', 'bring', 'government', 'central', 'dead', 'fight', 'citizen', 'kill', 'suspect', 'treat', 'old', 'authority', 'disease', 'show', 'health', 'across', 'hubei', 'press', 'new', 'novel', 'doctor', 'ship', 'beijing', 'outbreak', 'board', 'prevent', 'confirm', 'risk', 'ministry', 'president', 'time', 'spread', 'recover', 'organization', 'minister', 'battle', 'patient', 'infect', 'follow', 'rise', 'supply', 'report', 'number', 'express', 'suspected', 'two', 'china', 'province', 'week', 'drop', 'help', 'use', 'flight', 'hit', 'update', 'globally', 'close', 'people', 'first', 'start', 'medium', 'development', 'another', 'become', 'travel', 'taiwan', 'solidarity', 'one', 'call', 'treatment', 'official', 'amid', 'work', 'deadly', 'say', 'due', 'related', 'like', 'month', 'confirmed', 'conference', 'discharge', 'toll', 'hold', 'positive', 'emergency', 'second', 'stand', 'japan', 'year', 'concern', 'day', 'win', 'million', 'continue', 'outside', 'control', 'worker', 'combat', 'least', 'national']
- **Community 2:** ['global', 'outbreak', 'china', 'economy', 'say', 'impact', 'novel']

Communities	Overlapping nodes
Communities 1 and 2	['global', 'china', 'novel', 'outbreak', 'say']

Table 20: Summary of the overlapping nodes among the communities detected by the cliques percolation algorithm applied on the network built on the period of January-February 2020, after the removal of the search keywords.

September-October 2020, k=9

- **Community 1:** ['around', 'make', 'aide', 'global', 'penny', 'come', 'fall', 'high', 'include', 'rate', 'lead', 'know', 'long', 'adviser', 'late', 'infection', 'world', 'record', 'far', 'recovery', 'symptom', 'state', 'hospital', 'case', 'city', 'increase', 'need', 'begin', 'diagnosis', 'commission', 'set', 'test', 'face', 'election', 'since', 'university', 'daily', 'mike', 'total', 'tally', 'relief', 'pandemic', 'virus', 'negative', 'three', 'quarantine', 'center', 'public', 'mask', 'ahead', 'take', 'find', 'mark', 'would', 'joe', 'country', 'walter', 'democratic', 'nominee', 'expert', 'australia', 'accord', 'chief', 'nearly', 'volunteer', 'dose', 'hick', 'live', 'could', 'home', 'die', 'issue', 'see', 'expect', 'johnson', 'death', 'staff', 'trial', 'white', 'return', 'vice', 'government', 'medical', 'fight', 'potential', 'announce', 'five', 'warn', 'still', 'event', 'authority', 'show', 'disease', 'health', 'across', 'economy', 'qingdao', 'new', 'get', 'doctor', 'outbreak', 'house', 'lady', 'single', 'trump', 'risk', 'confirm', 'president', 'may', 'time', 'clinical', 'plan', 'statement', 'leave', 'spread', 'recover', 'military', 'organization', 'minister', 'former', 'next', 'study', 'follow', 'york', 'infect', 'rise', 'hope', 'bad', 'report', 'part', 'kamala', 'number', 'debate', 'two', 'reed', 'china', 'week', 'many', 'presidential', 'four', 'office', 'help', 'use', 'hit', 'also', 'remain', 'close', 'people', 'first', 'half', 'large', 'fauci', 'ministry', 'result', 'start', 'surge', 'tell', 'travel', 'one', 'treatment', 'official', 'well', 'amid', 'work', 'data', 'say', 'india', 'month', 'want', 'toll', 'positive', 'back', 'emergency', 'second', 'early', 'campaign', 'last', 'year', 'clear', 'candidate', 'day', 'source', 'south', 'million', 'wave', 'continue', 'top', 'hour', 'administration', 'restriction', 'wear', 'economic', 'national']
- **Community 2:** ['study', 'adult', 'sign', 'old', 'safe', 'show', 'appear', 'work', 'say']
- **Community 3:** ['senate', 'trump', 'president', 'test', 'nominee', 'say', 'court', 'supreme', 'amy', 'republican', 'positive']

- **Community 4:** ['condition', 'house', 'treat', 'white', 'trump', 'president', 'hospital', 'say', 'doctor']

Communities	Overlapping nodes
Communities 1 and 2	['study', 'work', 'say', 'show']
Communities 1 and 3	['president', 'doctor', 'say', 'trump', 'hospital', 'white', 'house']
Communities 1 and 4	['president', 'test', 'positive', 'say', 'trump', 'nominee']
Communities 2 and 3	['say']
Communities 2 and 4	['say']
Communities 3 and 4	['president', 'say', 'trump']

Table 21: Summary of the overlapping nodes among the communities detected by the cliques percolation algorithm applied on the network built on the period of September-October 2020, after the removal of the search keywords.

March-April 2021, k=7

- **Community 1:** ['air', 'war', 'leader', 'report', 'doctor', 'fauci', 'pandemic', 'special']
- **Community 2:** ['make', 'global', 'high', 'risk', 'president', 'plan', 'record', 'world', 'infection', 'minister', 'study', 'end', 'rare', 'rise', 'batch', 'bad', 'report', 'number', 'state', 'case', 'city', 'hospital', 'develop', 'two', 'link', 'china', 'adult', 'commission', 'infectious', 'nation', 'week', 'set', 'help', 'use', 'hit', 'benefit', 'dos', 'since', 'people', 'daily', 'first', 'variant', 'recommend', 'vaccination', 'fauci', 'pandemic', 'virus', 'start', 'surge', 'clot', 'take', 'one', 'find', 'population', 'country', 'possible', 'official', 'age', 'agency', 'amid', 'administer', 'new', 'expert', 'data', 'say', 'india', 'pause', 'accord', 'brazil', 'vaccinate', 'drug', 'goal', 'month', 'dose', 'approve', 'could', 'shot', 'emergency', 'give', 'regulator', 'see', 'johnson', 'second', 'death', 'trial', 'concern', 'year', 'last', 'government', 'day', 'announce', 'federal', 'africa', 'receive', 'million', 'wave', 'blood', 'top', 'medicine', 'germany', 'six', 'show', 'disease', 'health', 'least', 'across', 'national', 'get']
- **Community 3:** ['response', 'live', 'speak', 'fauci', 'house', 'team', 'white']
- **Community 4:** ['committee', 'rare', 'blood', 'clot', 'side', 'possible', 'risk', 'brain', 'effect', 'say', 'due']
- **Community 5:** ['effective', 'trial', 'show', 'disease', 'efficacy', 'data', 'new', 'say']

Communities	Overlapping nodes
Communities 1 and 2	['fauci']
Communities 1 and 3	['report', 'fauci', 'pandemic']
Communities 1 and 4	['trial', 'data', 'say', 'new', 'show', 'disease']
Communities 1 and 5	['possible', 'blood', 'rare', 'clot', 'risk', 'say']
Communities 2 and 3	['fauci']
Communities 2 and 4	-
Communities 3 and 4	-
Communities 3 and 5	-
Communities 4 and 5	['say']

Table 22: Summary of the overlapping nodes among the communities detected by the cliques percolation algorithm applied on the network built on the period of March-April 2021, after the removal of the search keywords.