Network science - AA 2020-2021

Teacher: Tomaso Erseghe

Network model for products

Alfredo Petrella - 1206627

February 8, 2021

1 Introduction

Both the release of the new Sony PlayStation 5 in November 12 and the announcement of the Pfizer CoViD-19 mRNA BNT162b2 vaccine effectiveness the morning of November 9 have had a global impact and have been strongly discussed on Twitter.

The aim of this project is to analyse, from different perspectives, the analogies and the differences between the two almost contemporary phenomena, even if they may seem uncomparable from one another at first sight.

2 Datasets presentation

Almost 1.5 million tweets (1500 API requests from 5 different Twitter accounts) have been downloaded and processed in order to build the two analysed networks, covering the entire ten days period following each event. In particular, given that only enterprise accounts can randomly sample from the entire Twitter dataset and filter by followers and friends number (in order to avoid bots and inactive users, for example), a non-trivial task was to find query which stroke a balance between the limited number of available requests and the bias of the datasets. Finally, the chosen queries are:

- for the tweets about the PS5:
  - *((#ps5) lang:en* to gather more tweets in the first two days and
  - *((#ps5 #sony) OR #playstation5) lang:en* for the other eight days;
- for the tweets about the vaccine:
  - *(#pfizer OR #pfizervaccine) lang:en* to gather more tweets in the first two days and
  - *((#pfizer #vaccine) OR #pfizervaccine) lang:en* for the other eight days.

For both the PlayStation 5 and the Pfizer vaccine, an undirected weighted hashtags network was then built, in which two hashtags are connected with an integer weight w if and only if they appear together in exactly w tweets. Moreover, every node of the network have different attributes

inherited by the tweets it was contained in, such as the first time it appears in the network and the average outcome of many sentiment analysis variables.

In order to perform the needed tasks, the used tools have been Python (NetworkX package and own code for some algorithms), LIWC and Gephi.

Figure 1: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks filtered by node degree in Gephi (about 5% of the whole networks, ForceAtlas 2 layout, nodes coloured by PageRank centrality).

3 Exploratory Data Analysis

From now on we will refer to the PlayStation 5 network as *PS-network* and to the Pfizer vaccine network as *V-network*.

Before starting with the networks analysis, a few checks have been performed in order to clean the data and familiarise with the dataset. Besides some non-English tweet and the majority of entries with missing location, it was important to note that:

- only 25% of the tweets about PS5 are original tweets, while in the vaccine case the retweets are less than 50%, which suggests an hypothesis about the attitude of the users with respect to the two topics: we will confirm later that, surprisingly, the engagement is stronger for the PS5 than for the vaccine;
- the V-network is slightly smaller than the PS-network but a little denser: this holds because the average number of hashtags per tweet is higher in the former (3.5 VS 2.9) but the tweets regarding the PS5 are three times more than the ones about the vaccine, even if the queries have been more selective.

4 Networks analysis

4.1 Networks overview

A summary of the main parameters of the networks is given in table 1, while the adjacency matrices are given in figure 2. For both networks, the nodes are sorted by decreasing degree; note that the apparently highlighted diagonal is due to the fact that hashtags with similar degree use to appear in the same tweets and not to self loops, which are not allowed by construction (the actual diagonal only contains zero values). Note also, in position 1253 and 812 respectively, the presence of one node with relatively high degree but few connections with the rest of the network: these positions correspond to the hashtags *#playstation5* and *#pfizervaccine*, unpopular siblings of the more spread *#ps5* and *#pfizer #vaccine*, added with the OR operator to the API query.

Different graphical representations of the networks will follow.

	Number of nodes	Number of links	Density	Power law
PS-network	$\|V\| = 9082$	$\|E\| = 102155$	$\rho = 0.002$	$\gamma = 2.34$
PS-network	$\|V\| = 7538$	$\|E\| = 71634$	$\rho = 0.003$	$\gamma = 2.26$

	Momenta	Distance
PS-network	$<k>$=22, $<k^2>$=14828, $<k^3>$=69626523	dist=4, $<d>$=2.15, $<d>_{exp}$=2.21
V-network	$<k>$=19, $<k^2>$=11544, $<k^3>$=48307846	dist=5, $<d>$=2.16, $<d>_{exp}$=2.19

	Clustering	Assortantivity	Robustness
PS-network	$<C>$=0.85, $<C>_{exp}$=0.01	$\mu = -0.15$, $\mu_{rand} = -0.024$	$f_{rand} = 0.97$, $f_{attack} = 0.31$
V-network	$<C>$=0.86, $<C>_{exp}$=0.01	$\mu = -0.16$, $\mu_{rand} = -0.023$	$f_{rand} = 0.98$, $f_{attack} = 0.22$

Table 1: Summary table.

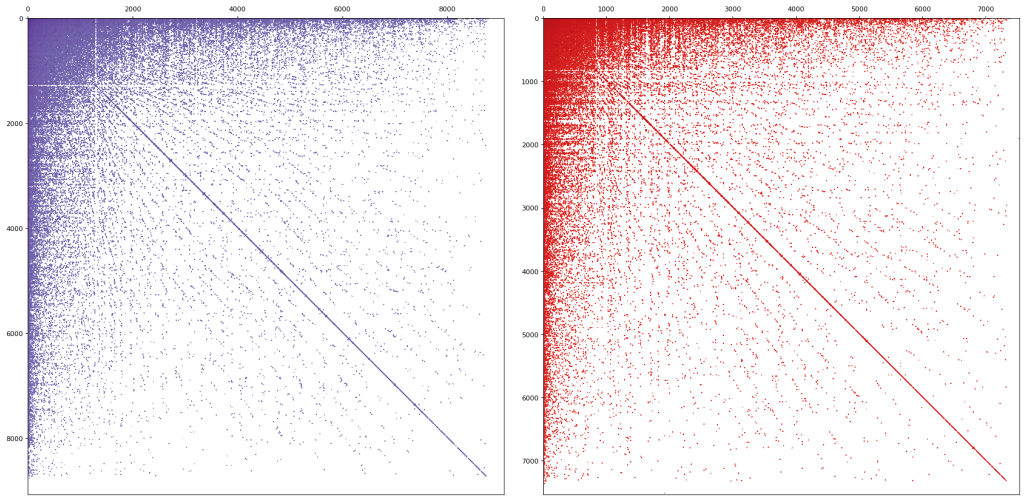


Figure 2: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks adjacency matrices.

4.2 Networks degree

Let us start from the degree distributions to understand the main properties of the networks. From the logarithmic plots in figure 3 it is easy to see that both the distributions follow the power law $p(k_i) = Ck_i^{-\gamma}$ with similar γ values, which have been calculated thanks to the maximum likelihood estimator obtained from the hypothesis over the distributions, considering $K_{min} = 12$ for both networks and then deriving $C = (\gamma - 1)k_{min}^{\gamma-1}$. In the plots is also represented the networks natural cut-off $K_{max} = K_{min}N^{\frac{1}{\gamma-1}}$, which suggests the absence of any structural disassortativity (more about it later on).

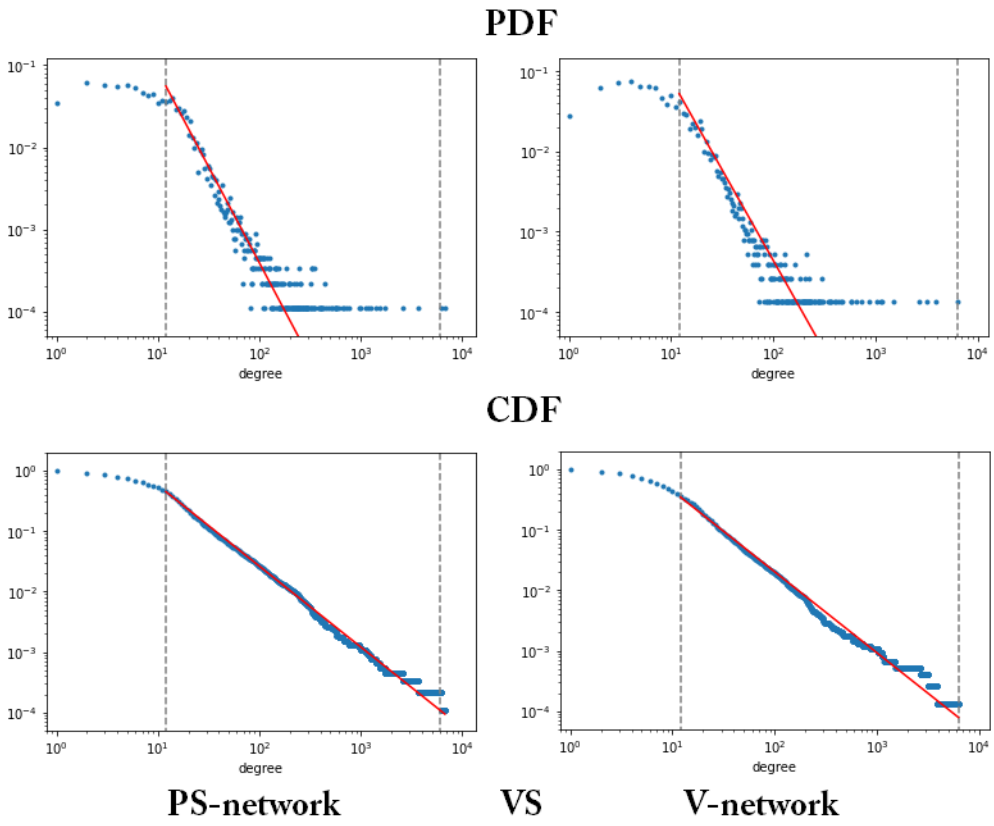


Figure 3: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks degree distributions in logarithmic scale.

Note that the values of γ tell us that the networks are similar to one another and scale free, which can also be seen through the higher order momenta in table 1 because they tend to explode as their degree increases.

The presence of big hubs (displayed later) is obviously due to the structure of the queries used to download the tweets (each tweet must contain at least one among the hashtags related to our tasks), which is the same reason that led to connected graphs without a few exceptions (the giant component of each network covers more than its 99.7% in both cases), while out-of-context and popular-but-unrelated hashtags constitute the small degree nodes.

4.3 Ultra-Small World property

Also from this point of view our two networks are pretty similar. In fact, computing the distance between each node through the breadth-first search algorithm, it has been found that the diameter values are 4 for the PS-network and 5 for the V-network, while the average path length is respectively $<d>_{PS}$=2.15 and $<d>_{V}$=2.16, in agreement with the expected values $\log\log N_{PS}$=2.21 and $\log\log N_{V}$=2.19. These properties as well can be easily explained looking at the queries: the short-

est paths between any two nodes, also for the most unrelated tweets, are the ones passing through the hubs.

4.4 Centrality measures

In this subsection the main centrality measures for the nodes of the network are compared and interpreted.

In general, no big differences among the computed centrality measures have been encountered, due to the fact that the networks are undirected and that they have been built starting from the queried hashtags. Nevertheless, as expected, both PageRank (PR) and Hiperlink Induced Topic Search (HITS) highlight a few nodes which where not considered important by the other measures and which lead to interesting interpretations. Note that, being the graphs undirected, hubs and authorities found via HITS coincide.

In general, only some of the most relevant distributions are reported to keep the report more readable.

4.4.1 PS-network

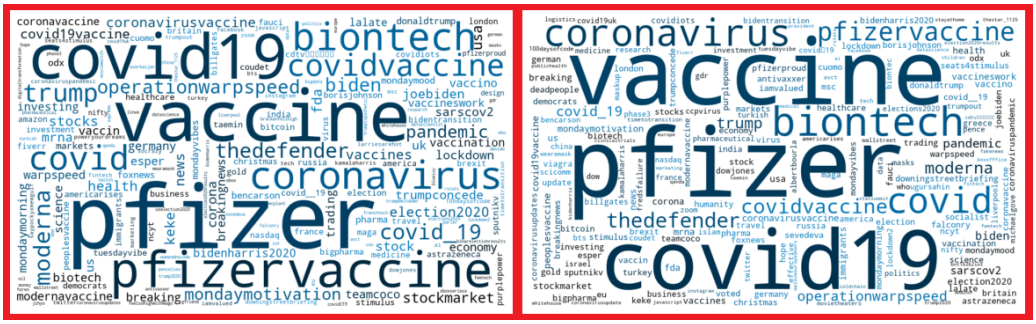
#ps5 and #playstation5 contend for the first two places in all the computed measures, followed by other trivially relative hashtags such as #gaming, #sony or #xboxseriesx, while in the lower positions some interesting differences show up: for example, looking at the closeness centrality, #twitch is quite high in the ranking, indicating that the streaming platform is an important node for the spreading of information regarding the PS5 (actually, the gaming world in general), while betweenness highlights #youtubedown, probably indicating a possible overlap between the gaming community and the youtube users one.

On the other hand, both PR and HITS rank at incredibly high positions (respectively 6 and 7, 4 and 5, figure 4) #rtx3070 (a GPU released by NVidia the month before) and #500cash: it turns out that a series of bot profiles have been publishing advertising contents during the days after the PS5 release, promising this trending products or 500 dollars as available prizes, and it could be interesting to investigate more about the nature of these bots and the reverse engineering algorithms they use to achieve these results. Besides, a decisive contribute in this area was given by the fact that the PS5 shortly became sold out on major retailing sites in Japan and the United States, naturally inducing to the introduction of lotteries to distribute limited stocks.

Figure 4: PS-network HITS ranking word cloud (first 200 hits), note #500cash and #rtx3070.

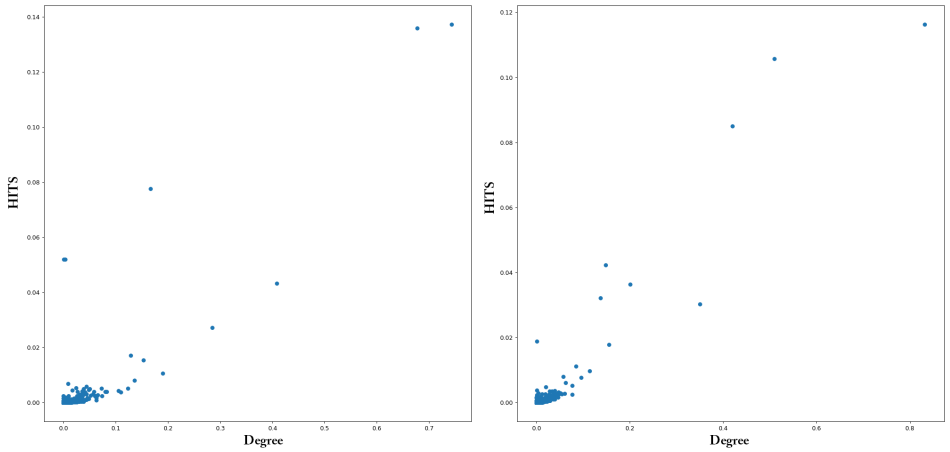
4.4.2 V-network

#pfizer, #vaccine and #covid19 are obviously the higher ranked, followed by similar ones, but in all the rankings (the degree centrality as well) it is immediate to spot socio-political and financial hits. One clearly needs to recall that a few days earlier (November, 3) the highly debated 2020 US elections were taking place, so part of this results should be considered noisier than usual. On the other hand, the strong involvement of D. Trump (running President of the United States of America in the period the data refer to) in CoViD-inherent decisions can't be denied, also regarding the Operation Warp Speed, relative hashtag of which high-ranks as well. HITS is the only algorithm which is able to separate the discussed sides, as it can be seen from the word clouds in figure 5 and the plot in figure 6, comparing the HITS results to the degree centrality distribution as a benchmark for the others: one can see that the relationship between the hubs and the other ranks is strongly non-linear on one side, and how the two measures differ more from one another with respect to the PS-network on the other.



PS-network VS V-network

Figure 5: V-network PR vs HITS ranking word cloud (first 200 hits).



PS-network **VS** **V-network**

Figure 6: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks comparison between the degree centrality and the hubs distribution.

Looking at the PR word cloud in figure 5, it is easy to spot many socio-political hashtags alongside with the ones inherent to the vaccine and to the news, such as politicians names, #usa, #elections, but also a lot of financial hashtags such as #stock and #investing.

On the other hand, HITS rank focuses on less nodes, highlighting one node in particular, which was less evident in the PR: it's #thedefender, referring to *Children's Health Defense*, an American nonprofit organisation founded and chaired by Robert F. Kennedy Jr., whose purposes are summarised on the official web page by one of its founder's quote: *The greatest crisis that America faces today is the chronic disease epidemic in America's children*, and this obviously points up the presence of a strong (and active on Twitter) network of antivax militants.

4.4.3 Networks comparison

The centrality distributions of the studied networks can help comparing them in a more formal and interesting way. To assess the similarity between two probability distributions p and q over the same sample space Ω, in fact, the Bhattacharyya distance is one of the most widely used statistical tools. It is defined as follows:

$$d_B(p,q) = -\log\left(\sum_{i=1}^{|\Omega|}\sqrt{p_i q_i}\right),$$

and it ranges from 0 to plus infinity.

Considering the set of all the hashtags appearing in at least one of the networks as a node, we can extend the previously computed distributions, setting to 0 the values related to the missing hashtags, and then compute the distance between them.

As for the evaluation of the resulting values, it must be said that the Bhattacharyya distance is usually used as a relative parameter to be minimised varying one distribution to match the other (it can be inferred, for example, from the lack of any kind of normalisation factor with respect to the cardinality of Ω), but it can help at a descriptive level.

Comparing the PageRank distributions we obtain $d_B(PR_{PS}, PR_V) = 1.69$, while comparing the hubs distributions we obtain $d_B(HITS_{PS}, HITS_V) = 2.23$, quite good results considering that the only explicit relation between our networks is the fact that the hashtags constituting the nodes come from tweets posted in an almost overlapping time period. Moreover, their difference is in line with our expectation: intuitively, the chance that two such unrelated networks share the same hubs is poor compared to the one that quite important hashtags for one network are shared and important also for the other one.

4.5 Clustering coefficients

Another strong similitude between the networks can be found in their clustering coefficients distributions. In particular, as shown in figure 7, almost all their nodes have a clustering coefficient around 1. This means that all the neighbours of the majority of the nodes are connected.

The average clustering coefficients, moreover, are $<C>_{PS}$=0.85 and $<C>_{PS}$=0.86, almost a hundred times higher than the expected clustering coefficient $<C>_{exp} = \frac{\log(N)^2}{N} = 0.01$, meaning that the nodes of the networks are much more connected than expected, once again thanks to the presence of the keyword-hubs and, from a practical point of view, of a common topic leading to a family of related hashtags which sooner or later will appear in the same tweet.

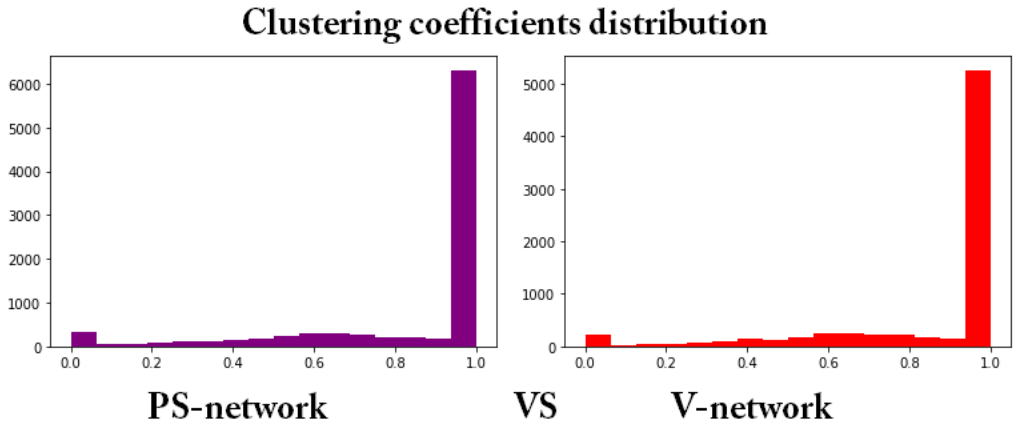


Figure 7: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks clustering coefficients distributions.

4.6 Assortativity

Looking at the average nearest neighbours degree, it is clear that both networks are disassortative, as shown in figure 8.

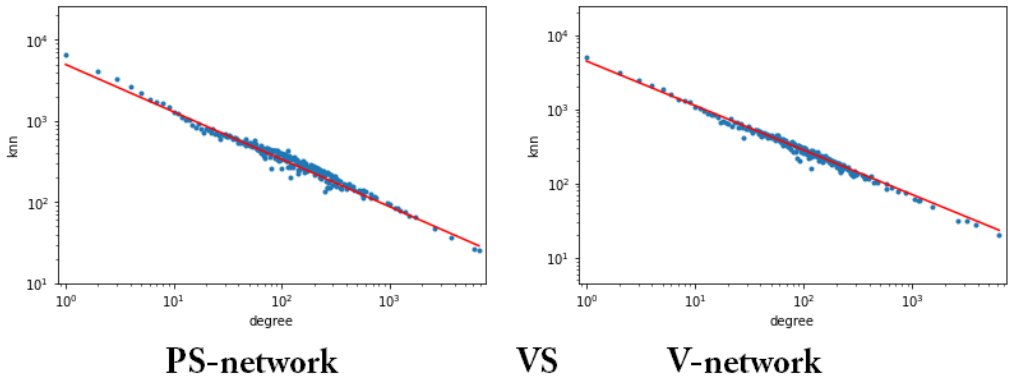


Figure 8: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks average kNN distributions.

We already suspected that it is not a structural property, but, to confirm that, the networks were randomly rewired, and the resulting average kNN distributions, shown in figure 9, leave no doubts.

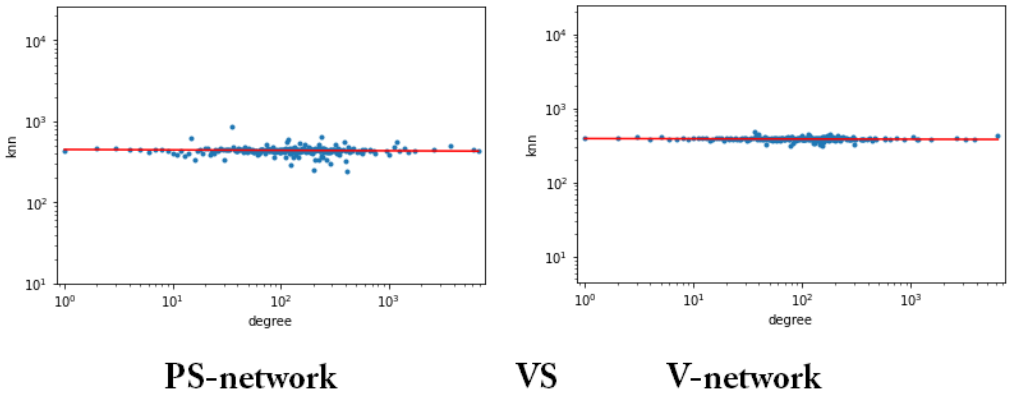


Figure 9: Randomly rewired PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks average kNN distributions.

Concretely speaking, the natural disassortativity of this network is probably due to way hashtags are usually utilised: most of the hubs refer to the same concept in different ways, so it is sufficient only one of them in order to describe the main topic, and then other less important hashtags are used to complete the tweet instead of other hubs.

4.7 Robustness

Being our networks scale free, we expect them to be very robust to random node removal but fragile in the case of an attack, and our expectancy is proven accurate by the plots in figure 10 and figure 11, displaying the average of 50 different random executions of the script. In order to calculate the breaking points, the average of the 50 inhomogeneity ratios $\kappa_i = \frac{<k^2>}{<k>}$ was used, which indicates that a giant component can't be identified anymore in the network when its value goes below 2.

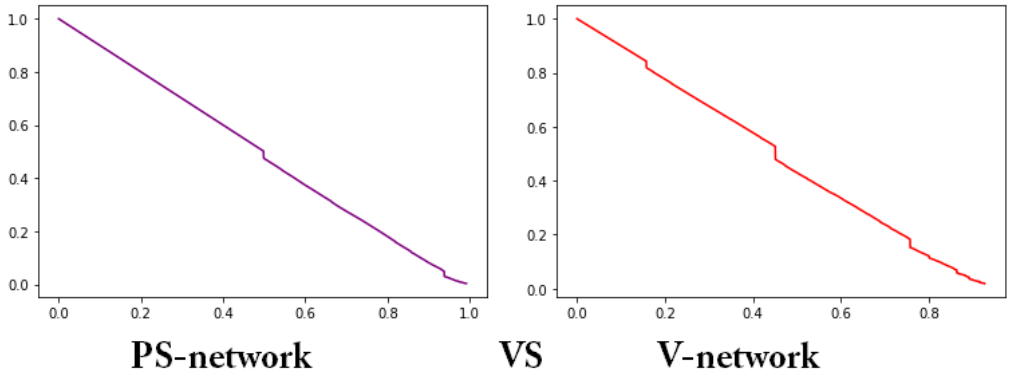


Figure 10: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks robustness to random node removal.

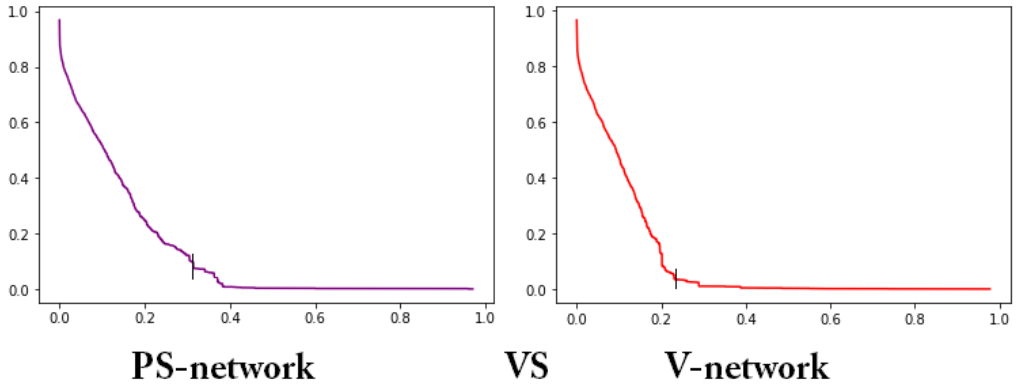


Figure 11: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks robustness to hubs removal.

We can actually note that the PS-network is slightly more robust to the attacks, probably due to the fact that many bot profiles use unrelated trending hashtags in their tweets together with the more related ones, but in general both networks strongly rely on the most popular/queried hashtags to remain connected.

Thanks to the sentiment analysis performed with LIWC, it is also possible to study the networks robustness in relation to the average tone expressed by each hashtag: starting from the data about the tweets, in fact, to each hashtag have been associated the average of the metrics relative to the tweets they where in. In particular, an interesting outcome is reported below: nodes have been randomly removed starting, respectively, from the ones associated with positive feelings and from the ones associated with negative emotions, resulting in the PS-network breaking before the V-network in the first case and vice versa. This shows, as we will clarify later, that the PS5 has been gladly welcomed while the vaccine has brought with it some negative emotions.

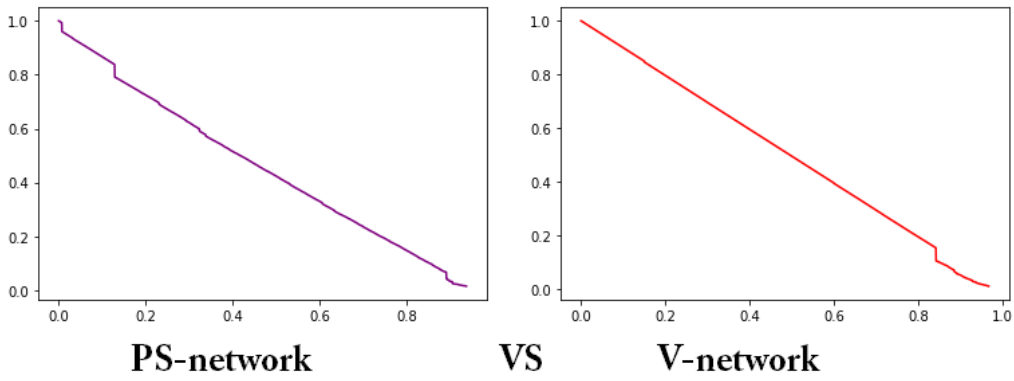


Figure 12: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks robustness to random node removal, starting from the positive ones.

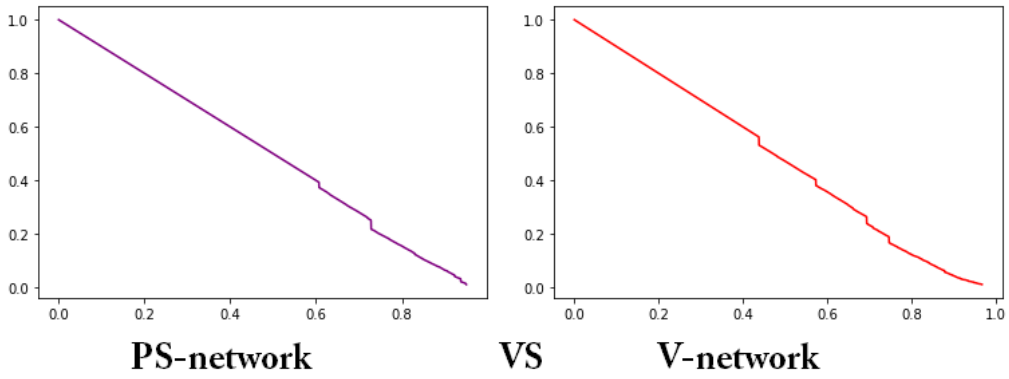


Figure 13: PlayStation 5 (on the left) and Pfizer vaccine (on the right) networks robustness to random node removal, starting from the negative ones.

4.8 Community detection and Gephi analysis

In this section we mainly focus on interesting insights over the networks obtained clustering and inspecting the graphical representation of the results, as well as on some of the most captivating outcomes of the sentiment analysis performed via LIWC on the tweets.

Different clustering algorithms have been tested with the help of the NetworkX package, but the most computationally efficient (both in time and space given the networks size) and interpretable approach happened to be the heuristic method based on modularity optimization implemented in Gephi and described in detail in [4]. A trial and error approach to the hyper-parameters tuning task was finally adopted in order to select the best resolution and nodes minimum degree for each of the two networks. It is clear that running the algorithm over less higher-degree nodes gives more interpretable results, but it also doesn't exploit most of the information contained in the network, so it was chosen to run the algorithm filtering out only the most sporadic nodes and than looking at the main hashtags on the biggest communities, also at the cost of losing some accuracy. The size is proportional to the PR centrality of each node.

Figure 14 shows the main communities obtained for the PS-network in different colours. It is clear that the green community is the most closely related to the PS5 and its spread in the market, while the white one refers to the competitors and to some popular games, probably discussed in review tweets. The orange cluster is somehow peculiar, gathering hashtags about the *Marvel's Spider-Man: Miles Morales* video game, released the same day of the PS5, and the hashtags relative to the previous model of the console (think to promotions and advertising): probably they were grouped together due to the fact that they were present since the beginning of the data collection and, for this reason, more connected than the average random alternative, which is what modularity-based algorithms consider. The light blue community is instead about contemporary but unrelated topics, while the purple one identifies the hashtags more often used among gamers, with a strong predominance of streaming platforms-related nodes, denoting a big number of online events organised to discuss about the new console.

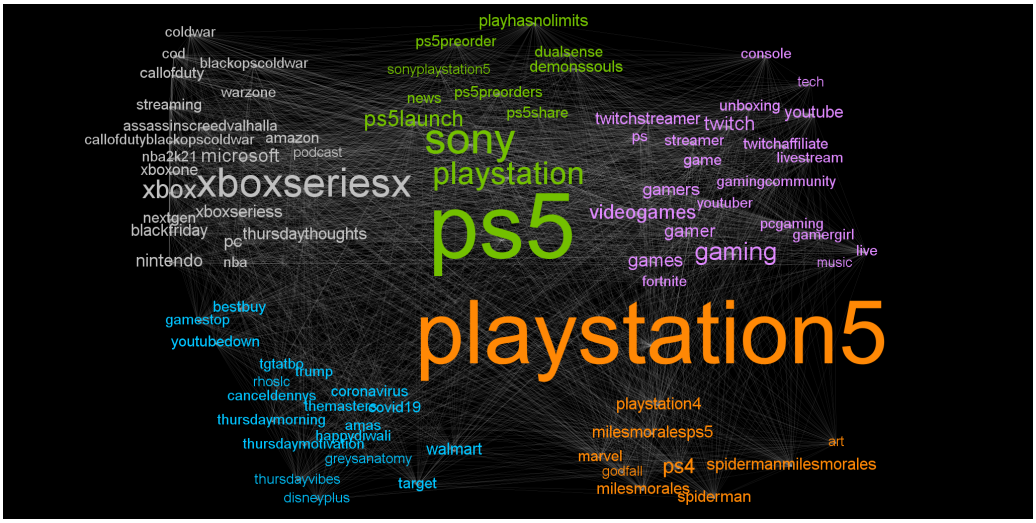


Figure 14: PlayStation 5 network's modularity communities, resolution = 0.4, about 2% of the network is visible.

The main clusters obtained for the V-network are, instead, shown in figure 15. The red community here contains the two most ranked nodes, together with hashtags which remind to the news world, regarding the management of the CoViD emergency. In particular, note #operationwarpspeed, the public–private partnership initiated by the U.S. government to facilitate and accelerate the development, manufacturing, and distribution of CoViD-19 vaccines, therapeutics, and diagnostics, and #fauci, referring to the famous American immunologist which has nowadays become chief medical advisor to President Joe Biden. The orange community contains more scientific keywords and the word *hope*, which in all the performed analysis is notably associated with science, while the green one corresponds to the PS-network light blue one. The smallest communities have been separated probably due to the high frequency of the corresponding hashtags, while, finally, the purple community is evidently related to the financial world.

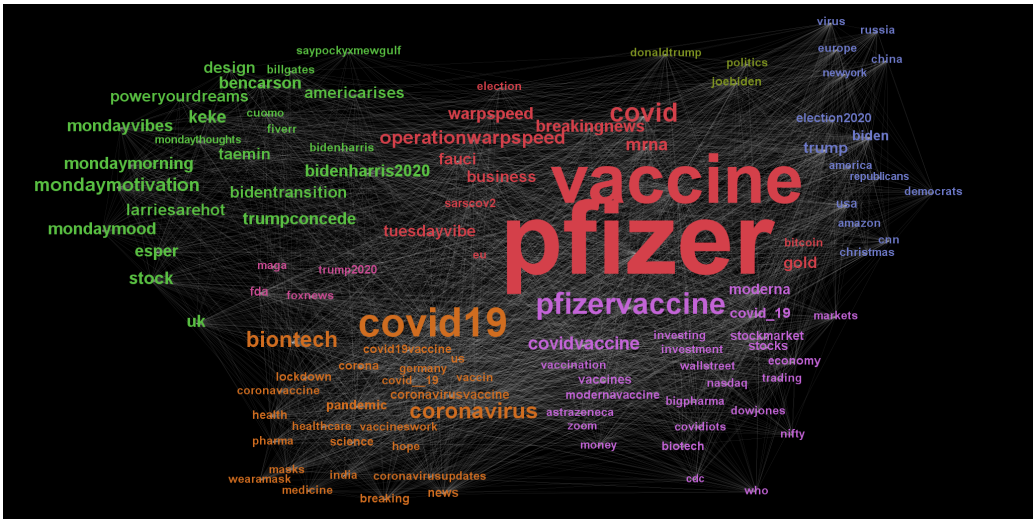


Figure 15: Pfizer vaccine network's modularity communities, resolution = 0.4, about 2% of the network is visible.

As for the sentiment analysis, the results regarding almost all the indexes computed by LIWC are generally not as informative as expected, probably due to the way they are projected over the hashtags from the tweets: the variance of the average sentiments is clearly low due to the huge number of data points, and in many cases the information is quite neutral, but let's investigate the exceptions.

In figure 16 can be seen that the sentiment regarding the PlayStation 5 is overall positive, with a particularly high value for the #xboxseriesx hashtag, but it doesn't necessarily means that this console is considered to be better: in fact, the analysed tweets contain PS5-related hashtags and, eventually, #xboxseriesx, and it is way more likely for a comparison to be performed praising the preferred product than criticising the least favourite one and so, independently on which console a user prefers, the sentiment projected on the tweet will be positive. Note also #gamergirl, denoting an usually positive approach to the topic, and the average negativity of the hashtags related to war video games, probably deriving from a more vulgar register, in the top left corner.

Figure 16: PlayStation 5 network' sentiment from negative (red) to positive (blue).

In figure 17 the sentiment regarding the Pfizer vaccine is shown: one of the most interesting results is probably the correspondence of the previously described science-related community to an area associated to generally positive emotions, which boosts up our previous hypothesis about the #hope hashtag position. On the other hand, the only slightly positive feelings about #vaccine denote a balance between the concern about the side effects of the vaccine on one side and the excitement the announcement brought. What is strongly visible, moreover, is high emotional engagement due to the elections, where the sentiment is generally negative. Finally, it is interesting to observe that the financial community is characterised by positive emotions while the tone of antivax tweets is generally negative, and that #breakingnews and #foxnews, referring to two of the most famous information channels, are respectively related to extremely positive and negative emotions, both showing up from the content of the news and from the reactions of the users.

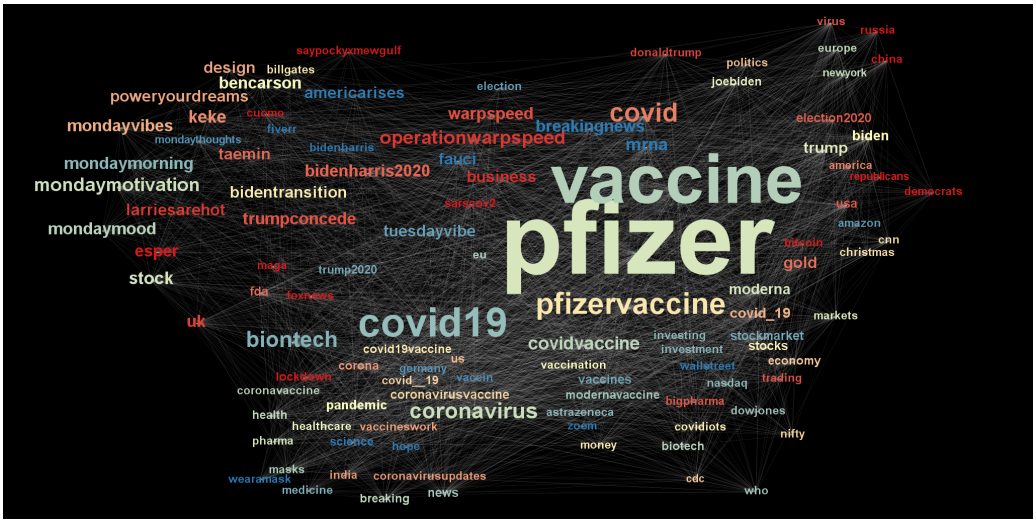


Figure 17: Pfizer vaccine network' sentiment from negative (red) to positive (blue).

Finally, a surprising result has been found by comparing both the degree of average affection in the two networks and the corresponding level of analytical approach in the tweet text. In figure 18 we can see that the emotional component (on the right) is slightly stronger than the analytical one (on the left), showing a balance between enthusiasm, disappointment and reaction to the performance of the new console, while figure 19 highlights a completely different setting: incredibly, the emotional component is way less present than the analytical one, underlying the presence of lower emotional involvement for the first version of a series of products that are going to return the entire humanity its normal habits than for a 26-years-reputation gaming console lots of people grew along with.

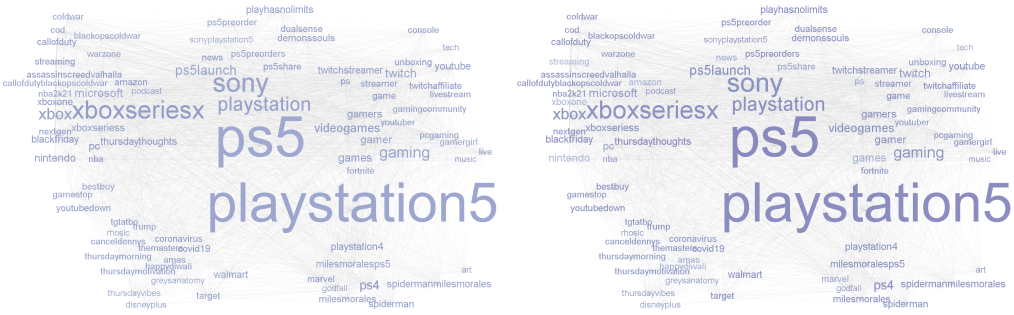


Figure 18: PlayStation 5 network comparison between analytical approach (on the left) and affect (on the right) in the *same normalised scale*.

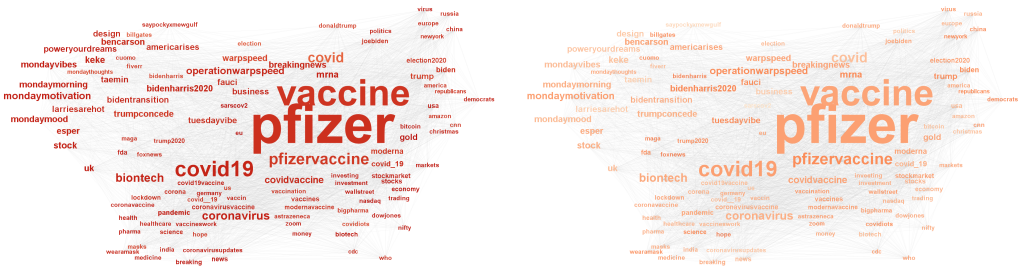


Figure 19: Pfizer vaccine network comparison between analytical approach (on the left) and affect (on the right) in the *same normalised scale*.

5 Conclusions

In view of the above study, the network are definitely more similar to each another than expected from many perspectives, and many individual results help to clarify aspects regarding the other network.

Clearly, a lot could still be said and studied, such as a link prediction task starting from the resulting joint network to discover which links would result more likely to have appeared on Twitter during those days, or the implementation of cutting edge community detection algorithms to improve the quality of the resulting communities, as well as deeper insights about the sentiment of each tweet and the other projection networks the initial data allow to build. For now, I just hope you enjoyed the reading at least a fraction of what I did thanks to this project.

References

[1] Albert-L´aszl´o Barab´asi et al. *Network science*. Cambridge university press, 2016.

[2] Extracting Linguistic Patterns from Texts with LIWC ("luke") for Analysis
`https://scalar.usc.edu/works/c2c-digital-magazine-fall-2016-winter-2017/extracting-linguistic-patterns-from-texts-liwc-analysis`

[3] Mastering Gephi Network Visualization
`http://gephi.michalnovak.eu/Mastering%20Gephi%20Network%20Visualization.pdf`

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008(10) P1000