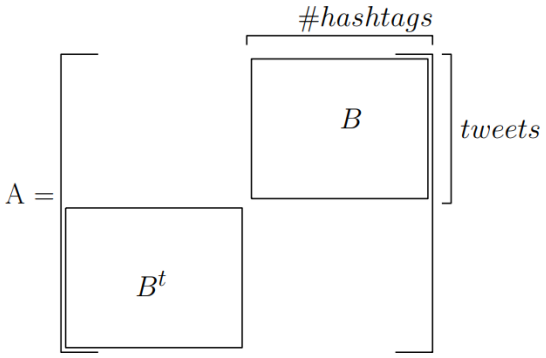
Giulia Rizzoli - Report

Initial Ranking: Code provided. Further details in data extraction report.

Bipartite Tweets-Hashtags:

The bipartite network of tweets and hashtags was build linking each tweet to the hashtags used in it.
In that way each row (tweet) sums up to the number of hashtags used in the post, and sums over a column (hashtag) states the total occurrences of a specific hashtag.
The algorithm randomly chooses 100k tweets per each dataset (prolife and prochoice).
Further **data reduction** was performed in order to deploy nodes (hashtags) which appear in the network only few times. This is done by setting a **degree threshold** which was arbitrarily set to be equal to 10, since the number of hashtags-nodes seems consistent.
In the end three different structure are build: one for prolife (100k tweets), one for prochoice (100k tweets) and the mixed matrix, which contains both (200k tweets).



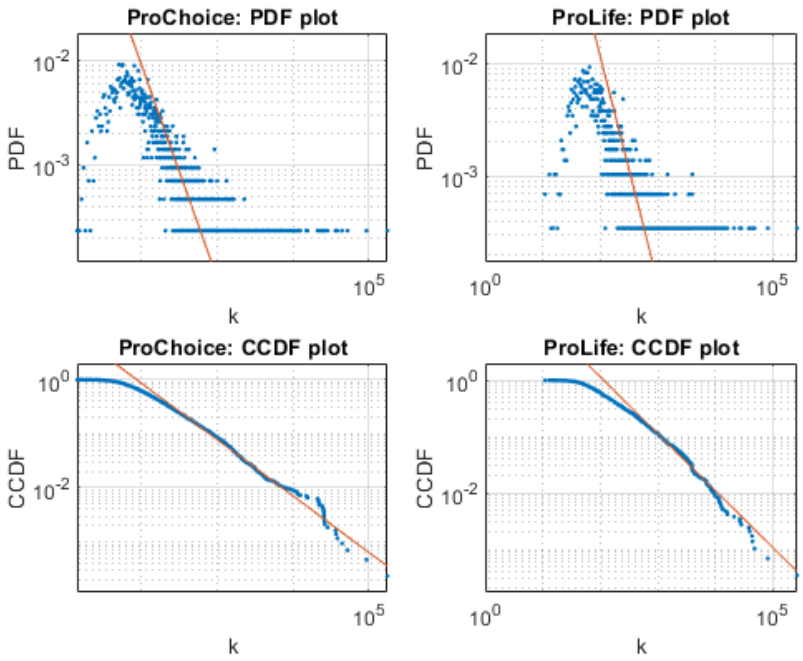
Projections network:

The projection network of hashtags it is simply obtained by multiplying the bipartite matrix to its transpose. Each cell contains the number of times two hashtags were used together in a tweet.

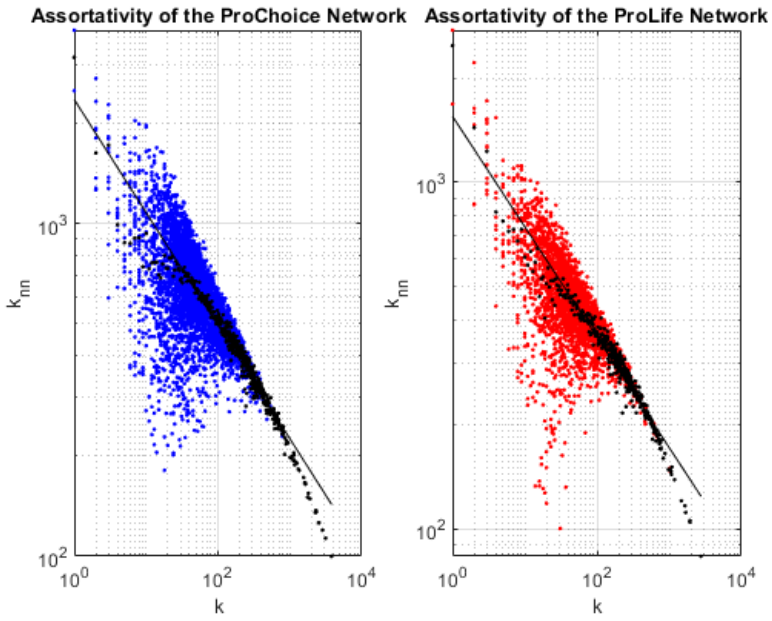
Aim of the study:

- Provide insights for the abortion hashtags projection network.
- Identify more relevant hashtags in abortion network and understand how the network is influenced by them.
- Extract metrics to capture the general behaviour in the different dataset of prolife and prochoice. State differences between the two.
- Identify hashtags communities in the abortion network.

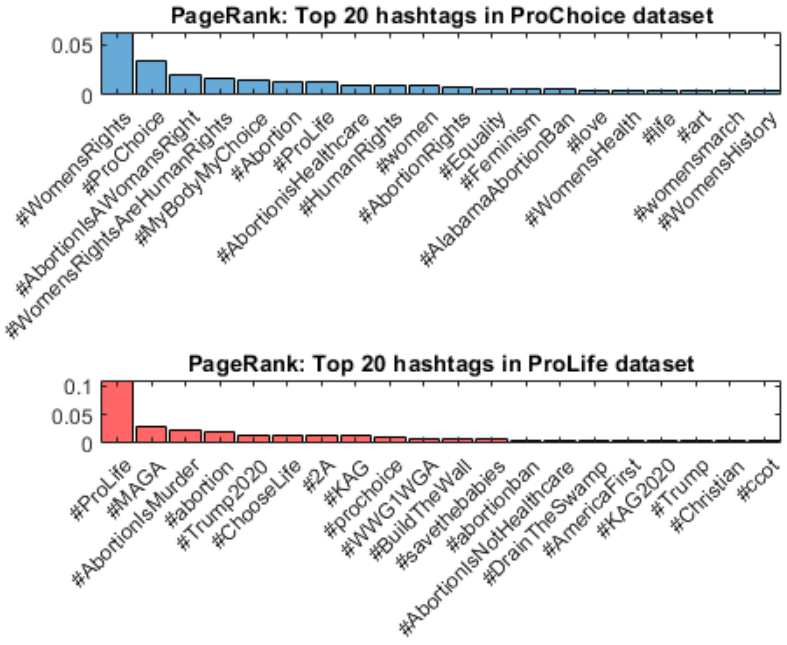
Projections are made for the three structures obtaining: projections for prolife dataset, prochoice dataset and for the mixed one.
Matrices were found to be fully connected.
Initial analysis was made on the distribution: the prolife and prochoice network has a **scale free** trend, with little deviance on big hubs, which indeed are present.
Further analytics are shown at the very end of the report.



Disassortativity in terms of degree could lead to conclusions that very connected hashtags tend to avoid being used with others big hubs. However, it could be also induced by structural reasons accounting that the network is scale free.



PageRank algorithm is performed for prolife and prochoice dataset. Results are reported in Figure. These outcomes are pretty much the same as in the initial ranking: it can be observed that, while in the prochoice data the hashtags are more focused on women's rights, equality and feminism, in prolife data instead it seems that politic orientations and movements are more relevant.

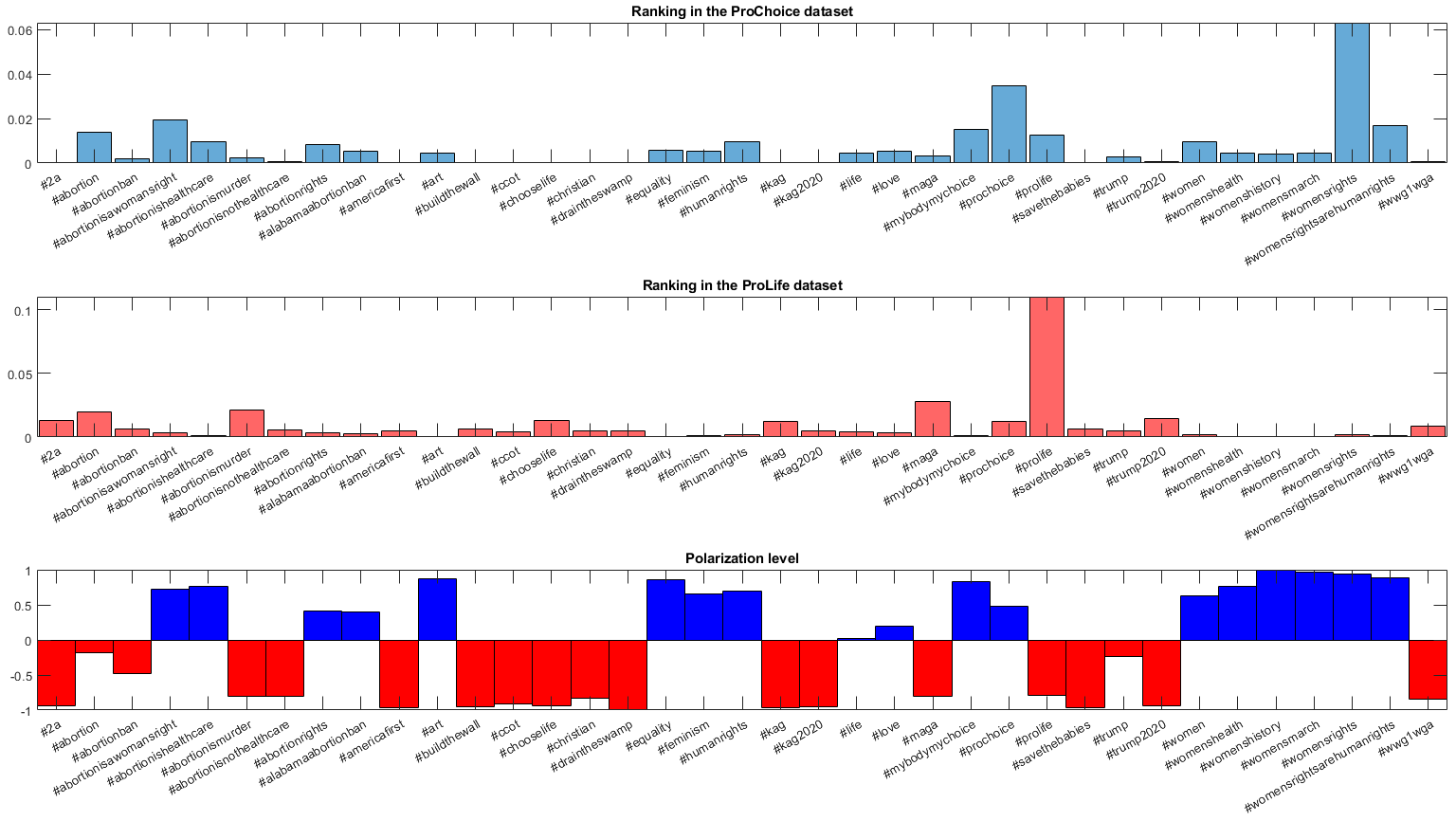


A more direct comparison is made between the two datasets by using the top ranked hashtags from the previous figure and show their PageRank weights once again.

To measure hashtags centrality among the two datasets a new measure was established: polarization.
Polarization vector is created for each hashtag as

$$P_i = \frac{W_{pc_i} - W_{pl_i}}{W_{pc_i} + W_{pl_i}}$$

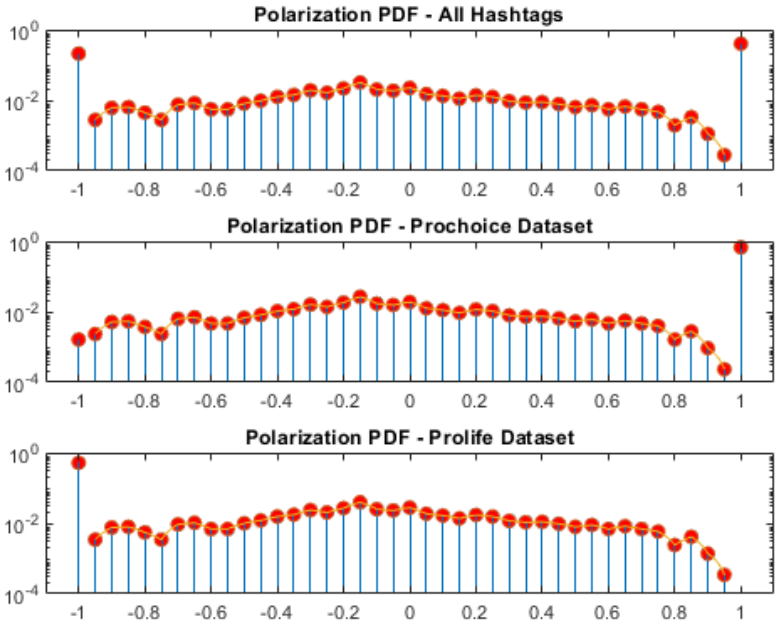
where W_{pc_i} and $W_{pl_i}$are the PageRank weights for prochoice and for prolife dataset.
If a hashtag is only present in one of the two datasets, the weight of the other one Is set to zero.
Results for the top hashtags are shown in figure.



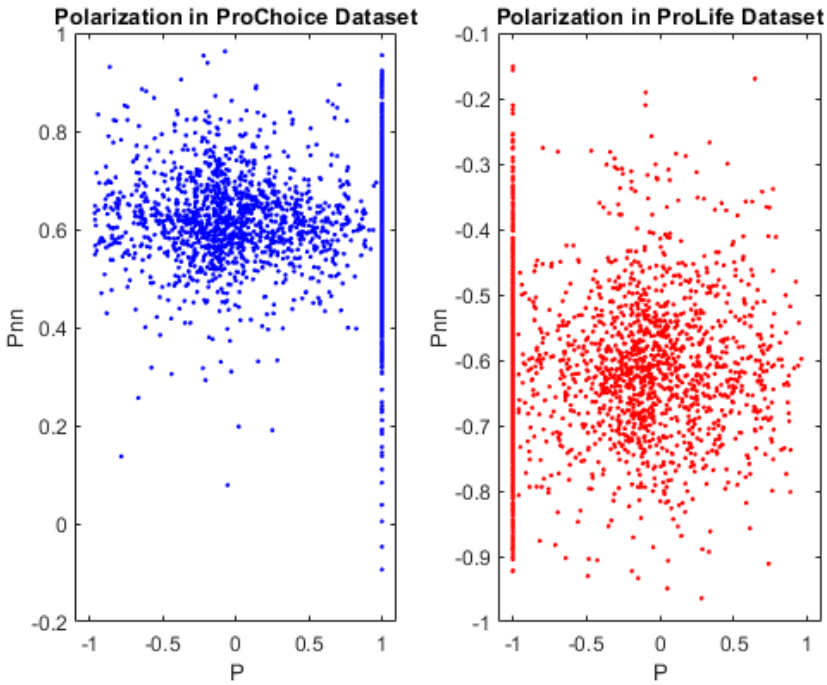
Polarization seems to be a good metric, since the hashtags are well balanced between the two classes (positive = prochoice, negative=prolife).
Especially it is noticeable that if a hashtag appears in both the datasets, as #prolife for example, it will end in the right one.
From the results among the top hashtags it seems that PR weight measures quite good the opinion for the main topic of abortion.

To capture deeply how the network is polarized it is computed the polarization **distribution** for each dataset, showing that there is a strong polarization at the extremities (+/-1), due to strong opinion hashtags. Note that this metrics could also be affected by the presence of hashtags that appear only in one dataset and it could be strongly related to the structure of the network.
The main findings in distribution are that while in prolife dataset distribution only prolife-opinion hashtags can be found (shown by the peak in -1), the prochoice dataset distribution describes the presence of both opinions, even if in a smaller amount.

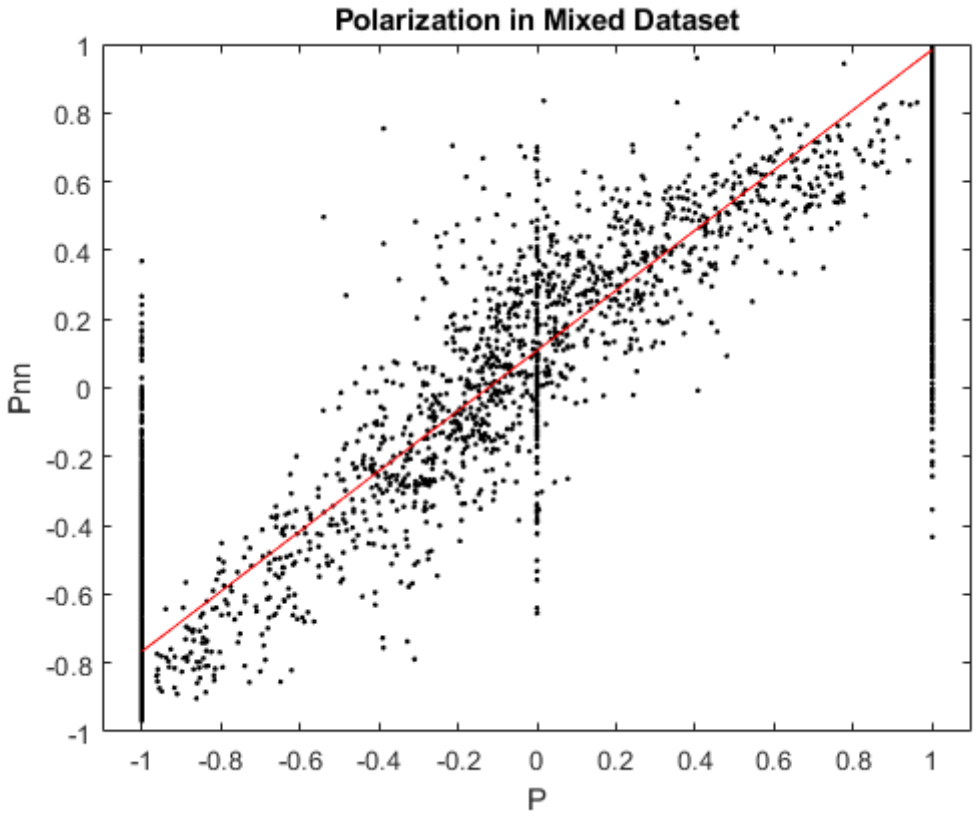
Note that the figure is in logarithm scale, since the probability in the (-1,+1) interval are really low, compared to the two extremities.



The evidence of the absence of a debate in prolife could be also seen by computing the nearest neighbour for polarization, which is it compared to prochoice in the following figure. Only in prochoice network the strongest polarized hashtags (positive in prochoice) reached the opposite side (negative valued hashtags, prolife).
In this figure it is also clearer the presence of nodes which are extremely polarized and connected to nodes which ranges (mostly) from 0 to 1 in case of the prochoice and -1 to 0 in prolife dataset.

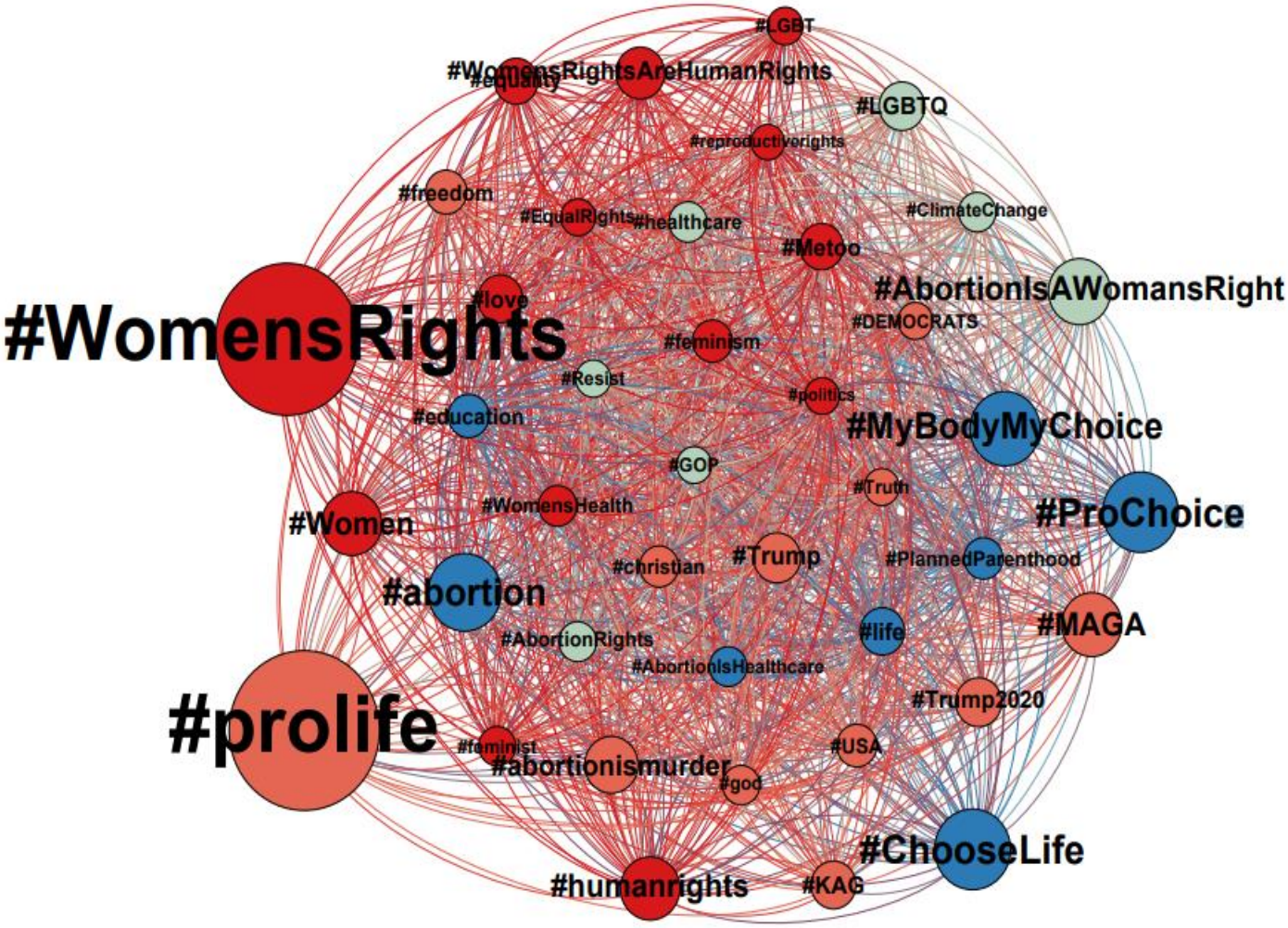


Then the same idea is applied to the mixed dataset.
From polarization a linear behaviour among nodes could be captured, in which strongly polarized nodes can be identified, either prochoice and prolife. That makes sense since the mixed dataset contains hashtags from both the datasets.
The network clearly shows the opposing parties, even though they share some nodes.
Nodes with strong polarization are mainly connected to others with the same class.
It can be also found a set of neutral hashtags which is linked to polarized nodes from both the classes.
Further tests could be trying to apply this metric to other abortion datasets, in order to see if it keeps information, meaning that it could be applied in a more general way, or if it is deeply depending on the structure of the network.



More general features of the overall behaviour in the mixture dataset are then exploited using Gephi.
The importance of a node is established by performing ranking which measures betweenness centrality.
Prolife hashtag result in having an enlarged importance. Such finding matches what was obtained by studying the two datasets separately: since in the prochoice dataset the prochoice hashtags were also co-used with prolife ones, #prolife's result in being more influent.
Then communities are identified using the modularity approach: the prolife and prochoice opinion-hashtags were divided. Prolife are reddish nodes, while the blueish are prochoice.
Most evident biased nodes in community detection were #WomensRights and affiliates, which in the previous analysis were found to be strongly prochoice.
The community division not only identifies the two main popular opinions, but also divide the dataset in small clusters (7 in total) with topic similarities: for example, #Woman, #Metoo, #WomanHealth belong to the same cluster, while #Abortionismurder, #God and #Christian are associated to another.

The insights statistics about nodes (i.e. centrality measure, community belonging) are saved in a csv file. The network of hashtags is only the projection of the full matrix and it could lead to loss of significant information. Also, there could be other parameters, different from hashtags, to state user-opinion and community belongings.
Final figure is the result of filtering by degree, to see only the more relevant nodes in the network.



Main Conclusions:

Pros: In the abortion network strong opinion hashtags can be clearly identified (figure above).
Polarization could be a good way to accounting hashtags opinion, by which tweets can be divided.

Con: No real information about tweets content.

Main Analytics Report:

	ProChoice	ProLife	Mixed
Avg. Degree	602,2893756	748,2451791	644,2308244
Gamma	2,041314015	2,009035284	2,090230704
Assortativity	-0,339774362	-0,318718056	-0,314190758
Avg. Distance	1,306803696	1,204921254	1,289299379
Clustering Coef.	0,552839964	0,574848731	0,545588344