Black Lives Matter *A Network Science approach to BLM debates on Twitter*

Erica Cau
e.cau@studenti.unipi.it
Student ID: 545126

Veronica Mesina
v.mesina@studenti.unipi.it
Student ID: 521783

Alfonso Ferraro
a.ferraro16@studenti.unipi.it
Student ID: 626006

ABSTRACT

The murder of George Floyd on 25 May, 2020, was the trigger but also the chance to highlight the colonial roots of the Western society and the intrinsic racism of its coercitives institution.
In this paper we conducted a systematic study – through Network Science[1] tools – of the relationships developed between Twitter users on this subject, taking a look to their discorsive productions and to the context of their respective communities.

KEYWORDS

BLM, Black Lives Matter, Colonialism, Politics, Racism, Social Media, Social Network Analysis, Twitter.

1 INTRODUCTION

As stated in the official website, Black Lives Matter is a movement founded in 2013 *in response to the acquittal of Trayvon Martin's murderer* with the aim of *eradicate white supremacy and build local power to intervene in violence inflicted on Black communities by the state and vigilantes.* The *BLM* acronym and the hashtag *#blacklivesmatter*, became wordspread in 2020, with the murder of George Floyd on May, 25th, by Derek Chauvin.
The purpose of this analysis is to exploit the relationships intertwined by users on this subject and to understand – through Natural Language Processing techniques – which are the topics that enlivened the debates on the movement. The first section will describe the chosen source of data, the data scraping process and how the network was created. Here, the network will be also analysed and compared with other synthetic models.
Then, the second and third section will identify clusters of similar users – the communities – and will study of their diachronic evolution through *dynamic community discovery.* After these steps, the analysis will be deepened by modeling the way ideas about this topic propagated through the network and then the focus will shift to identify those nodes that are more likely to form a connection in the network.
The last section will outline an extensive study on the main users communities – the ones identified in Section 4 – using NLP, namely *Topic Modeling* and the traditional *Sentiment Analysis.* The final aim, in fact, is to observe how the different communities are characterized in terms of discussed topics and general attitude in order to point out how they interact using Twitter. Then, we will conclude the whole work in Section 9, with a discussion of the more interesting results and the potential developments of the project.

2 DATA COLLECTION

Selected Data Sources

The first – and necessary – step to begin with the whole analysis is the *data collection* part. It has been decided to not use pre-existent datasets but to create a new original dataset that could better fit our needs. The chosen data source was *Twitter*, since it acts like a dynamic and public square, where even a quick thought may potentially reach an exponential number of people, unlike what happens with other online social networks, like *Instagram* or *Facebook.* In their cases, in fact, lot of profiles are private – so it is not possible to scrape what has been published by a user – and there's also another problem, that is that their algorithms tends to rank the posts that will be shown to the user accordingly to the user's personal navigation history or to the information about the post itself[2]: this means that many posts are condemned to an inesorable overshadow, that may also affect the overall quality of the collected data and, consequently, of the analysis, since many users may not have the chance to interact with posts of their interest, that would turn out in a loss of information. Moreover, Twitter is also very easy to scrape through Python, so it was like an obvious choice.

[1]**Project Repositories**
Data Collection: https://github.com/sna-unipi/2021---final-project-cau_ferraro_mesina/tree/main/data_collection
Analytical Tasks: https://github.com/sna-unipi/2021---final-project-cau_ferraro_mesina/tree/main/network_analysis
Report: https://github.com/sna-unipi/2021---final-project-cau_ferraro_mesina/tree/main/report

[2]See the Instagram blog for details: https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

Crawling Methodology and Assumptions

Moving to the more technical side, the Twitter scraper was instantiated through the Python library *Twint*[3]. The scraper was set in order to download a maximum of 5000 tweets per day in the timespan between February 26, 2020 and March, 1 2021. This specific timespan was arbitrarily chosen, because it was supposed to cover the two events of worldwide importance related to the Black Lives Matter movement: the *murder of George Floyd* on May 25, 2020 and the *2021 United States Capitol attack* on January 6, 2021. This should allow to study the temporal evolution of the *BLM* communities and the language used when they interact.
All the tweets have in common to contain in their raw text the hashtag *#blacklivesmatter*. At the very beginning, we tried to scrape all the tweets containing the acronym *#BLM*, but this idea was soon abandoned for the high number of unrelated tweets, since *blm* is also an indonesian word meaning *not yet*, that isn't specifically related to the topic of this research.
To sum up, it was downloaded a total number of 1260935 english tweets both of verified and not verified accounts[4], originally divided into four different *.csv* files, accordingly to the quarter in which they were produced.

Data Understanding and Data Cleaning

After the scraping part, it was obtained a dataset whose records were described by 36 features: the most interesting ones were the date when the tweet was posted, the tweet itself and the number of likes, retweets (RTs) and replies. Other interesting attributes were also the username and the *user id* of the author of the tweet and a further particular feature that contained all the replies to the tweet with all the information about those specific users. This attribute was the key to build the whole network.
In addition to this, there were also other uninteresting features that were stored by Twint to keep all the information about the user/tweet localization or about the translation to another language: these attributes were dropped, since they mainly contained missing values and were therefore considered uninformative and useless.
Then, it was time for the final step in this preprocessing part, that is the drop of the missing values and/or duplicate values. The following list shows the number of tweets for each quarter after the data cleaning process.

- **First quarter (Q1)**: 46.327 tweets;
- **Second quarter (Q2)**: 346.969 tweets;
- **Third quarter (Q3)**: 339.622 tweets;
- **Fourth quarter (Q4)**: 528.017 tweets.

We also visualized the temporal distribution of the tweet (see Figure 1): as opposed to our initial assumption, the users tweeted more in January 2021 than in June 2020, so, apparently were more prone to talk about the Capitol Hill attack rather than about the murder of George Floyd.
At this point, it was extracted another dataset from this one that became an edgelist describing all the relationships (edges) between a user and another one (the nodes). The edgelist has been structured in such a way that it has n rows for each reply to the tweet: the first two columns store the *username* and the *user id* of the author of the tweet, while the third and the fourth column contain the *username* and *user id* of each user that answered to the tweet. Then, a fifth column was created with the specific purpose to store the *weight* assigned to each conversation (the *edge*) between two specific users: the *weight* was defined as the number of times that an interaction happened, normalized in order to obtain a value in the range 0,1.
The final edgelist is shown in Table 1.

Figure 1: Tweet timeseries

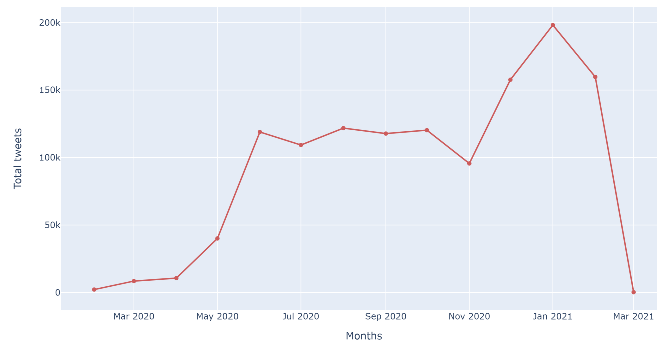


Table 1: Network edgelist

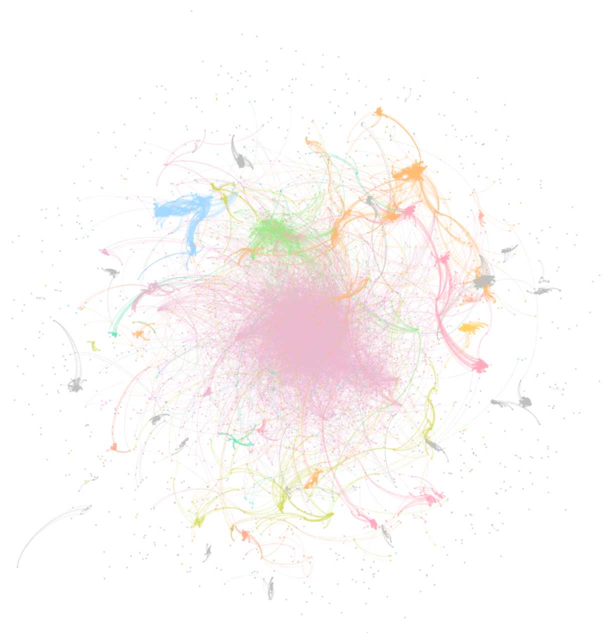
source	id_source	target	id_target	weights
pengologist	16548831	SeattlePD	25101704	1
info4u2know	47725408	thehill	1917731	1
soloyochapin	93071854	realDonaldTrump	25073877	1
info4u2know	47725408	realDonaldTrump	25073877	1

3 NETWORK CHARACTERIZATION

After obtaining the edgelist, the next step was the creation of the *Real-World Network* (hereafter *RW*) using the Python library *NetworkX*[5]. The graph was modeled as a *directed MultiGraph* (the `MultiDiGraph()` class in NetworkX), since

[3]Github repository: https://github.com/twintproject/twint
[4]As it will be discussed in Section 9, it wasn't possible – due to problems of the scraper itself – to extract a feature stating if the user was a verified one or not.
[5]Documentation: https://networkx.org/documentation/stable/index.html

Figure 2: Graph visualization using the *Force Atlas 2* layout on *Gephi*



this structure allows to have multiple directed edges linking the same nodes (i.e. multiple answers between the same pair of users). The graph includes 52795 nodes and 81063 edges, one of which is a self-loop. The graph was then visualized through Gephi[6], an open-source software for graph analysis and visualization. The chosen layout was the *Force Atlas 2* and the nodes and the edges were colored with different hues according to their modularity score[7] (see Figure 2). The choice of this specific layout model is justified by the overall balance between quality and performance, as stated in [6]. In order to compare the results of the analysis, four different synthethic graphs were also instantiated.

- **Barabási-Albert** (BA): using 52795 nodes and m set to 2;
- **Erdős–Rényi** (ER): a directed graph with 52795 nodes and its p parameter equal to 0.000005;
- **Watts–Strogatz** (WS): a graph with a total amount of 52795 nodes and the parameter k set to 4 and p set to 0.2;
- **Configuration Model** (CM): by setting, for each node of the model, the corresponding degree of the node in the RW network.

[6]Gephi: https://gephi.org/
[7]Note that *modularity* is calculated by Gephi using *Louvain*, so the colors approximately corresponds to the communities identified in Section 4.

Table 2: Comparison between the RW network and the synthethic models (BA, ER, WS, CM)

Model	Number of nodes	Number of edges	Average Degree	N° self-loop
RW	52.795	81.063	3.07	1
BA	52.795	105.586	3.99	0
ER	52.795	13.786	0.52	0
WS	52.795	105.590	4.0	0
CM	52.795	81.063	3.07	22

Table 3: Average, highest and lowest degree of the nodes in the RW and in the models.

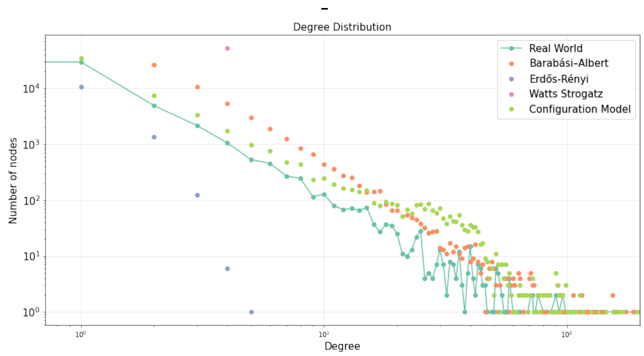
	RW	BA	ER	WS	CM
Average degree	3.07	3.99	0.52	4.0	0.52
Highest degree	949	819	5	4	949
Lowest degree	1	2	0	4	1

Degree distribution analysis

The first measure to be analysed was the degree of the nodes in the RW network; furthermore, the analysis was further developed by calculating the average *in-degree* and *out-degree* of each node; the results were then compared with the ones obtained with synthetic models.
As it can be observed in Figure 3, the models are very different from the RW network, since they follow a different distribution with the only exception represented by the *Configuration Model*. The most different ones are the ER and the WS model, since they both follow a Poissonian distribution, while the BA and the CM are – to some extent – quite similar. The CM model is necessarily the same of the RW network because of its intrinsic feature of having nodes that reproduce exactly the same degree of the RW network nodes.
The explanation of the similarity of the BA network, must be searched in the nature of the Barabási-Albert model itself, since this kind of model tries to simulate the presence of *hubs*, i.e. nodes with higher degree than the other ones in the network, through the means of ***preferential attachment***, similarly to what happens in a RW scenario.

Then, we shifted the focus on the exploration of the nodes with the highest degree: these are mixed users, both notable – like Donald Trump and Joe Biden – and non-verified ones. It is interesting to note that often these non-verified accounts also have the hashtag *#BLM* or other references to the movement itself in their *Twitter bio*. Among the highest-degree nodes, there's also the official *Black Lives Matter* profile, which corresponds to the third most *tweeted* profile.
In conclusion, it was also found out that the top 20 of the

Figure 3: Comparison between degree distributions (RW, BA, ER, WS, CM)



users that tweeted about *BLM* were all non-verified users – so they have a high out-degree value.

Connected component analysis

The second part of this network analysis was about exploiting the *connected components*, defined – accordingly to [2] – as *subgraphs whose nodes can be reached from one another by following the edges of the network*, both in the RW network and in the synthetic models. From the network visualization shown in Figure 2, it was possible to visually distinguish the **biggest connected component** in the center of the image: by deepening the analysis, it was assessed that it is composed by 38444 nodes (corresponding to the 72.81% of the whole network). This is not the only connected component in the network, but just one out of the overall number (3138). We observe a different behaviour in the ER model – which has 39009 smaller connected components, with a biggest component composed by just 4 nodes – and also in the CM, which is characterized by 4018 connected components. The BA and the WS are, instead, a different case, having both just one big connected component.

Path analysis

The path analysis – unfortunately – is a very demanding task in terms of computational resources, since its complexity is equal to $O(N^3)$, where N is the number of nodes in the network. In order to overcome this problem, we considered only a sample of 25000 nodes taken from the biggest connected component and calculated their average shortest path. The RW network has a shortest path length around 6.24, while in the synthetic networks the results are slight different. The BA network returned an average shortest path equal to 5.61 – a value that is even lower in the CM (4.97). The synthetic model that deviates the most from the RW network is the ER, since its average shortest path is aroud 1.6.

Table 4: Density and average clustering coefficient (RW, BA, ER, WS, CM)

	Density	Avg clustering coefficient
RW	$2.90 \cdot 10^{-0.5}$	0.01
BA	$7.57 \cdot 10^{-0.5}$	0.0
ER	$4.94 \cdot 10^{-0.5}$	0
WS	$7.57 \cdot 10^{-0.5}$	0.5
CM	$7.59 \cdot 10^{-0.5}$	0.0

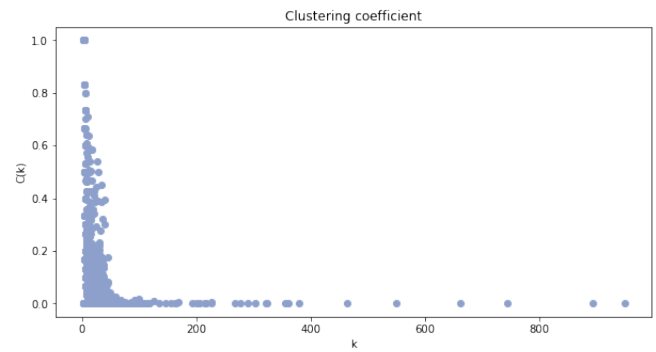


Figure 4: *Clustering coefficient*

Clustering coefficient and density analysis

As most real-world network, even this one tends to be *sparse*, so it is characterized by a low clustering coefficient, as it can be seen in Table 4. All the models fail in approximating the original clustering coefficient. The most similar is the WS model, since this kind of synthetic network tries to keep an high clustering coefficient, by placing the nodes equally distanced from each other and linking them to their k nearest neighbours.

Centrality analysis

The most important nodes of the network were computed using the following centrality measures.

Degree centrality

This simple centrality measure takes into account the number of neighbours of each node. In this specific case, the results were mostly uninteresting, since the only two notable nodes are the account of **Donald Trump** and **Joe Biden**.

Closeness centrality

This measure, instead, gave more interesting results – that were also closer to our expectations. The most central nodes were mostly users tied to politics or to the information field, as shown in Figure 6. Beside those ones, there were also few accounts involved in the sports field: one of them was

the official account of the *National Football League* (*@NFL*) – involved in a controversy for not taking a strong position on the issue and for some racist controversies surrounding the football society itself[8] – while the other one was the verified account of the Formula One driver *Lewis Hamilton* (*@LewisHamilton*), who became a symbol of the *BLM* movement for his statements on the matter[9].
Among all of the relevant accounts, one of the most interesting ones was that of Andy Ngo *@MrAndyNgo*, an independent conservative journalist with far-right political ideals, who became famous after publishing a video where he was attacked by an *Antifa* group while filming an *Antifa* rally in Portland, 2019.
The *closeness centrality* also detected an account that we deemed to be noise (*@golfyang*), since it was almost inactive and, when it was active, was not involved in discussions about BLM.

Betwenness centrality

The following centrality measure extracted as main nodes several users not really involved into the BLM, although with some exceptions. It is interesting to notice that even the betweenness centrality considered as important some users already recognized by the other centrality measures, as can be seen in Figure 5. In this case, the account of Alexandria Ocasio-Cortez (*@AOC*) is among the most important nodes of the network, along with the user *@ayrinweloveyou*, a victim of racism. Even here, there are users involved in the field of information even though of lesser importance: one interesting account is the official Twitter profile of *ELLE Magazine*, where the *BLM* is surely intertwined and narrated by famous female – and/or femminist – voices, like Beyoncé or Meghan Markle.

Harmonic centrality

This measure returned results similar to the ones of the *Closeness centrality*, probably due to the likeness between the two measures. No notable differences could be found, except for the swapped rank of a few accounts.

Page Rank

The results of *Page Rank* may be considered are a summary of all the others results: there are both verified and not verified accounts, some of whom even not strictly related to BLM.

Assortativity

Assortativity can be described as a quantitative measure of homophily, the node tendency to connect with similar

Figure 5: Betweenness centrality

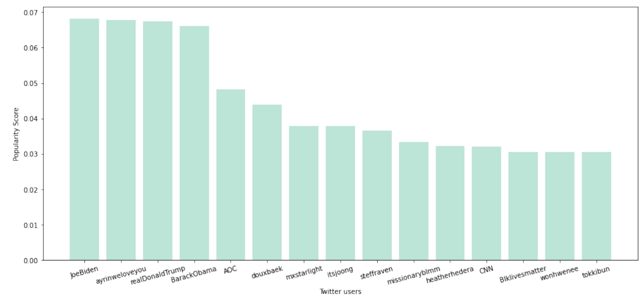
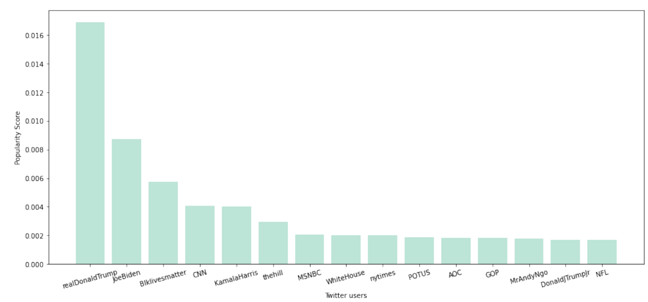


Figure 6: Closeness centrality



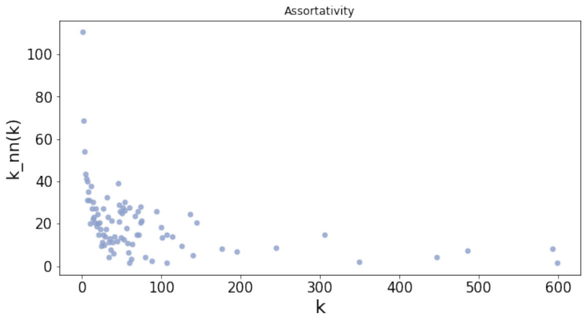
nodes. Additionally, assortativity may be considered as a correlation measure between nodes in a graph and it can be calculated using quantitative network attributes, such as the node degree: an assortative network has pairwise links between nodes with the same degree (so, hubs are connected with other hubs), while in a disassortative network the nodes with smaller degree tend to connect to other nodes with an higher degree.
For this project, the assortativity was calculated using the *NetworkX* implementation described in [8], that defines assortativity as the Pearson correlation coefficient of the degree calculated between pairs of linked nodes.
The network turned out to be slightly disassortative, since $R = -0.1311$: this can be explained by looking at Figure 3, where it can be seen that the RW network has an heavy-tailed degree distribution, characterized by a small number of hubs and, in contrast, a massive number of nodes featuring low degree.
The correlation between node was also visualized through the scatter plot shown in Figure 7.

4 TASK 1: COMMUNITY DISCOVERY

The primary task in order to identify – and subsequently analyse – clusters of users was the *community discovery*. We decided to extract the communities both in a static and temporal setting, since static communities are essential to find

[8]See for details: https://www.nbcnews.com/politics/meet-the-press/data-shows-how-bad-nfl-s-racial-equality-problem-among-n1288709
[9]See: https://twitter.com/bbcradio4/status/1342745749440897024

Figure 7: Nodes correlation through assortativity



an answer to the *open question* that is presented in Section 8, while their temporal evolution is interesting to gain some insights on the relationships and events occurred during the one-year timespan.

The community discovery task has been carried out running and evaluating the results of five different algorithms among the ones implemented in the CDlib Python library[10]: the ones chosen and tested were ***K-clique, Louvain, Demon, Angel, Label Propagation***. Then, all the algorithms were optimized for this network through hyperparameter tuning, conducted by using the the RandomSearch implemented in CDlib. The parameters returned by the RandomSearch can be found in Table 5.

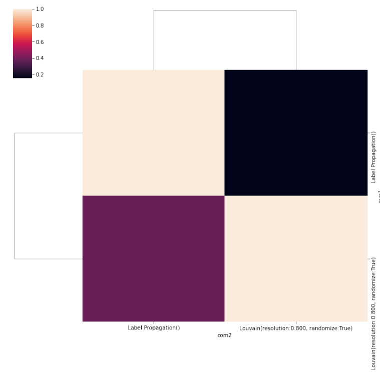
Table 5: Best parameters for CD algorithms

K-clique	*k = 6*
Label Propagation	/
Louvain	weight = weight resolution = 0.5 randomize = False
Demon	*epsilon = 0.5* *min_com_size = 4*
Angel	threshold = 0.1 min_community_size = 4

Hence, the results were analysed through both internal and external evaluation measures. The first ones are useful to assess the intrinsic quality of the communities identified by each algorithm, while the others are crucial to compare the results obtained running different algorithms (e.g. identify if there are similarities between clusters of users). The first internal evaluation measure used was the modularity, which describes the density of the identified partitions using a numeric value between -0.5 and $+1$. The highest results

[10]Documentation: https://cdlib.readthedocs.io/en/latest/

Figure 8: NMI similarity matrix for *Louvain* and *Label Propagation* communities



were obtained, at least in term of modularity, through *Label Propagation* and *Louvain*, where the value is close to +1. Furthermore, both the algorithms covered all the node in the graph, in contrast to what happens with the other three CD-algorithms. If we consider also the node coverage, *K-clique* is the worst-performing algorithm, since it identifies only three communities, moreover only in a small fraction of the original graph.

The drawback of both Louvain and Label Propagation is the high number of identified communities, a lot of them with just few users inside, and a very low *average internal degree*. This is balanced by very good results in the *conductance*, an evaluation measure that conveys the idea that edges should always link nodes inside the same community, and in the *cut ratio*, a measure that defines the possible number of links leaving the community.

The external evaluation instead, was carried out through ***Normalized Mutual Information***, that allowed to compare the similarity between the partitions indentified by pair of algorithms. In particular, it was applied on the output of *Louvain* and *Label Propagation*, since they have a full node coverage: the obtained score, 0.79, confirms that there is a partial overlapping between the communities identified by the two approaches.

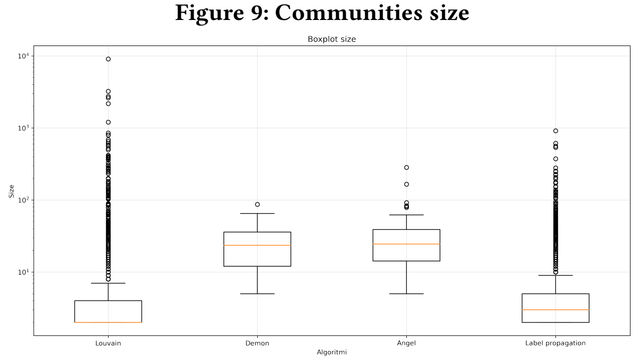
To compare the other algorithms, the ***Normalized F-1 score*** was used. All the obtained values are very low: the only exception is the comparison between *K-clique* and *Demon*, where it was registered a value equal to 0.42, so they have a similar, weak though, community division.

To better see the similarities, we plotted a *similarity matrix* for each pair of results. The most interesting, of course, is the similarity between *Louvain* and *Label Propagation*, followed by the one obtained through comparing the communities identified by *Demon* and *Angel*: their similarity may be explained by thinking to the nature of the algorithms, since they both exploit the *ego network* structures. The commu-

Table 6: Community discovery internal evaluation

	Community	Coverage	Average internal degree	Cut-ratio	Modularity	Conductance
K-clique	3	0.0007	7.20	0.0001	0.01	0.47
Label Propagation	8379	1.0	20.83	1.287	0.76	0.24
Louvain	3281	1.0	1.3468	1.38	0.90	0.00
Demon	159	0.063	5.4480	0.001	0.15	0.39
Angel	106	0.063	5.3974	7.5e-05	0.13	0.27

Figure 9: Communities size



nities were also compared by size through boxplots (see Figure 9): this was useful to find out that most of the *Louvain* communities are small, in contrast to the ones identified by *Demon* and *Angel*, where the median line in the respective box is significantly higher and there are lesser communities which size is considered an outlier, so it can be stated that their identified communities are more balanced.

5 TASK 2: DYNAMIC COMMUNITY DISCOVERY

In addition to the *static community discovery*, that allowed to identify the main *BLM* groups, we addressed also to its dynamic counterpart, which tries to recognize the fast changes in social relationships using snapshots of the network at different times. These snapshots captures the network topology across different timestamps and, by looking at the differences between them, it is possible to reconstruct the the possible evolution.
The main problem of this approach is tied to the inevitable loss of information about the evolution of the edges that are formed between nodes if the timespan between each snapshot is too wide, that can also be described, in other words, as not relieving interesting changes – and insights accordingly – in the behaviours of the users in a finer-grained perspective.
Given that the dynamic approach to community discovery needs an edgelist enriched with an additional information, namely a temporal indication, the initial division of the scraped tweet into four different *.csv* files (one per quarter) turned out to be useful: this allowed to assign the respective timestamp to each tweet in a much more easier way.

Table 7: Parameter tuning for Dynamic Community Discovery

Louvain	*weight = "weight"* *resolution =* *randomize = true*
Demon	*epsilon = 0.1* *min_com_size = 3*
Label Propagation	-

Preprocessing: agglomerative network

The first obstacle in this approach to community discovery, was the preprocessing of the network, since some algorithms, like *Demon* failed to work if applied on the whole network. This led to the creation of an *agglomerative network*, namely a network where the nodes and the edges are not removed once they're added at a timestamp t; it is possible for new nodes to appear in the network, though.
To avoid long execution times, we sampled only 25.000 random nodes from the agglomerative network.

Algorithm choice

After the definition of the agglomerative network, it was possible to proceed with the *Dynamic Community Discovery* task. There are many approaches to this task, namely *Temporal trade-off*, *Cross-Time* and *Instant Optimal*: the last one is the approach used for the specific purpose of this report.
First of all, we looped onto the different timestamps and, for each one of them, we applied the algorithms already ran in the previous section – with the only exception of *K-clique* – in order to execute *a)* a Random Search to find out the best parameters for each algorithm – if necessary – and *b)* identify the stability of the partitions through the four timestamps.
The best parameters obtained *via* RandomSearch are displayed in Table 7. Then, the algorithms were compared through the respective *NF-1* scores, that in this case define,

Table 8: NF-1 score

	Q1-Q2	Q2-Q3	Q3-Q4
Louvain	0.04	0.15	0.05
Demon	0.08	0.19	0.15
Label propagation	0.12	0.51	0.68

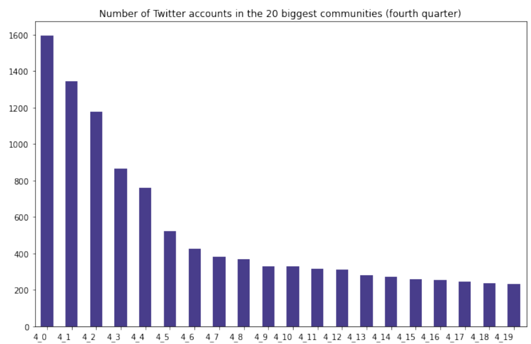


Figure 10: Communities size in Q4

for each pair of quarters, the diachronic stability of the snapshots. The results, shown in Table 7, prove a strong stability of the *Label Propagation*, that identified 203, 1328, 1933 and 2321 communities for each of the four timestamps, and less stability but a more balanced number of communities from *Louvain* that identified 110, 369, 249 and 76 communities. *Demon*, instead, was not able to identify any community in the first quarter unless the minimum number of nodes in a community was set to 1. This is possibly due to the fact that we random sampled the nodes for this task, which could have led to isolated nodes.

Final discussion

First of all, the identified communities were analysed in size: it was then found out that their sizes follow a Zipfian distribution, that is strictly evident in the last quarter (see Figure 10), where the first three communities are made up of, at least, 1000 nodes, while the other are evidently smaller, with a significant number of micro-communities, with just a few users inside.

The communities were then analysed by observing the lifecycle of one community using the *polytree* created by CDlib (see Figure 11): the nodes of the graph represent a community, tagged with the *t_id* and the *community_id* given by the *Temporal Clustering* object.
The polytree, for visualization's sake, was plot excluding all the other nodes with the exception of the ones involved in the evolution of the community tagged as *1* in the first timestamp. The figure clearly shows a ***merge*** event, in orange, where the community *1_0* (45 users) merged with 1_6 (30 users), 1_16 (14 users) and 1_68 (2 users) in order to create a new community, the *2_11* (132 users) during the quarter distinguished by the murder of George Floyd by Derek Chauvain.

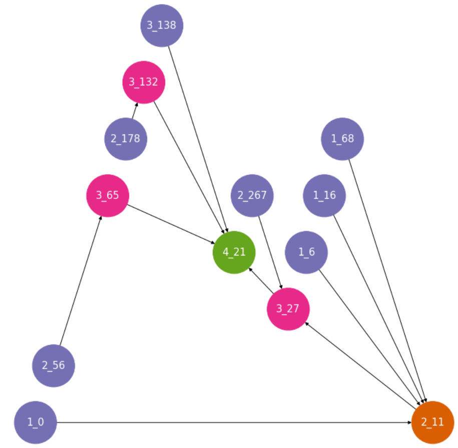


Figure 11: *Community 1_0* polytree

In addition to this, in the second quarter, we can observe also the ***birth*** of three other small communities, namely 2_56, *2_178* and *2_267*, formed by 16, 3 and 2 people respectively.
In the third quarter it can be observed the ***merge*** of existing communities into a bigger one: this is the case of *3_27* (158 users), formed by the aggregation of micro-community 2_267 into community 2_11. The same happens with 2_56, that merged into community 3_65 (16 users), and with community 3_132 (only 3 users).
The final result of this lifecycle is only one community, *4_21*, made up of 203 users.

6 TASK 3: OPINION DYNAMICS

The next step was to shape the way the opinions about Black Lives Matter spread in the network, using Opinion Dynamics models implemented with the NDlib Python library[11].
For the purpose of this analysis, it was decided to implement three different models: *Voter model*, *Majority Rule* and *Sznajd*. The results were then compared with the ones obtained *via* the application of these algorithms on two synthetic models, the *Barabasi-Albert* and the *Watts-Strogatz* graphs[12].

Voter model

Voter is a simple mathematical model born to simulate the competition between species, and then it was adapted to

[11] NDlib documentation; https://ndlib.readthedocs.io/en/latest/overview.html
[12] Both the synthetic networks were created using the same parameters as the ones described in Section 3

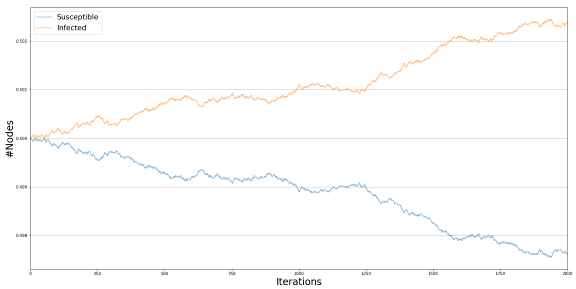


Figure 12: Voter model (RW)

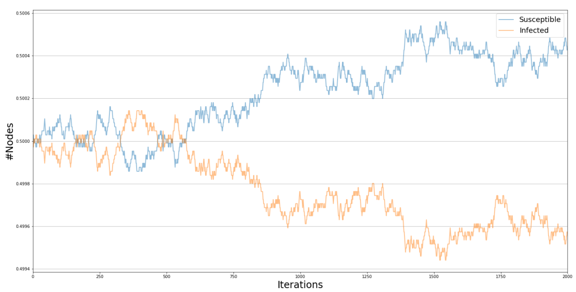


Figure 13: Voter model (BA)

Figure 14: Majority Rule (RW)

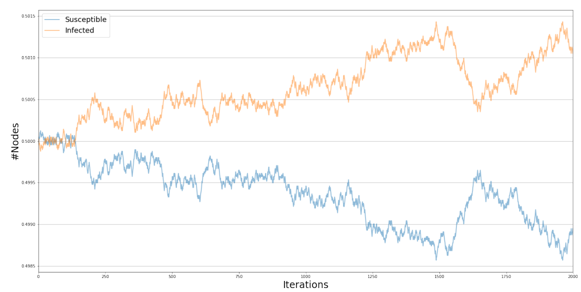


Figure 15: Majority Rule (BA)

simulate electoral competitions. In this case, the nodes of the network are the *voters*, and each one of them has an opionion described as a discrete variable, equal to ±1. During each of multiple iterations, the model chooses two neighbouring nodes, *i* and *j*. The node *i* takes the opinion of the node *j*.
For this work, we have chosen to run the algorithm for 2000 iterations and then we moved to the comparison with the synthetic models, in order to find out which one of them better approximates the behaviour of the *Voter* model on the RW network.
The *Voter* model in the RW shows a perfect split of the nodes with a positive and a negative opinion. Then, when the iteration number grows up, the percentage of positive opinions reach the 50.2%: complementarily, we can observe a simmetric decrease of the nodes with a negative opinion. This behaviour is not captured by any other application of *Voter* model on the other networks: for example, in Figure 13, there's a game of growth and decrease of nodes with a positive and negative opinion, then the model seems to be more stable in the remaining 1500 iterations, with a growing number of nodes with positive opinions.

Majority Rule model

The *Majority Rule* model assigns a discrete opinion, modeled as ±1, like the *Voter* model. The difference lies in what triggers the change of the opinion itself. *Majority Rule* implements a concept borrowed from sociology and psychology, that is *social inertia*: it defines the resistance to change of social groups, so it implies a degree of unwillingness w.t.r. opinion change. This is why, starting from a random group of nodes with their opinion, it assigns to every single node the opinion of the majority.
The model ran for 2000 iterations on all three networks – RW, BA and WS. The model behaviour on the RW and the WS is quite similar. The main difference between the two outputs is that the WS seems to model a way simpler diffusion process. They are both characterized by an phase of slight growth of the negative nodes, and then by a decrease of them followed by a long phase of decrease of users with a positive opinion. This last phenomenon starts clearly from the last 1000 iterations in the RW and after just 300 iterations in the WS network.
The algorithm acts differently when applied on a BA network: in this case, the exact opposite happens, since the decreasing node opinion during the final part is the positive one instead of the negative.

Sznajd

The *Sznajd* model is based on the *social impact* theory, which encapsules the principle that an idea shared by a group of individuals is more influent than the same idea but held by just one person.
Even in this framework, the opinion is modeled as ±1. The algorithm works by choosing a pair of nodes: if they share

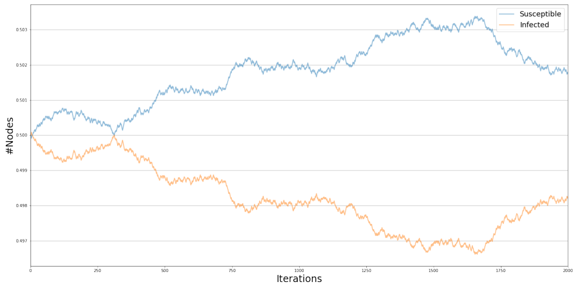


Figure 16: Majority Rule (WS)

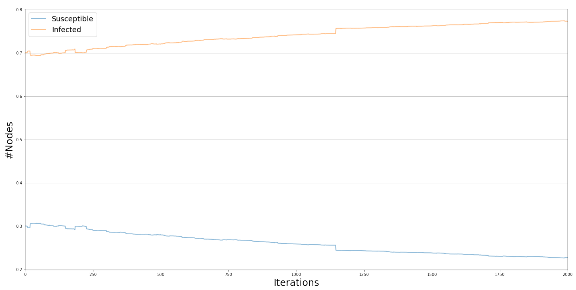


Figure 17: Sznajd model (RW)

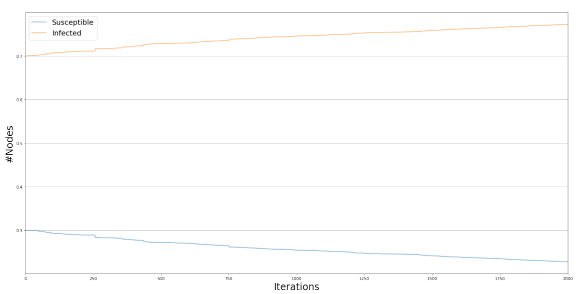


Figure 18: Sznajd model (BA)

the same opinion, their neighbourhood accepts this opinion. The model gave similar results: iteration after iteration, in all the three networks is registered a growth of the nodes that accepts the positive opinion (from 70% up to 80%) and a consequent decrease of the nodes with a negative one.

7 TASK 4: UNSUPERVISED LINK PREDICTION

In this section, we address another network science task, the so-called *link prediction*, useful to understand if it is possible to infere the links between nodes in the network. In our case, the link prediction problem may be formulated as follows: *which Twitter users are more likely to reply to a tweet?*
The main disadvantage of the link prediction approaches – that may both be supervised or not supervised – is the high computational complexity, equal to $O(|V^2|)$, where V is the number of nodes in the whole network. Another issue is

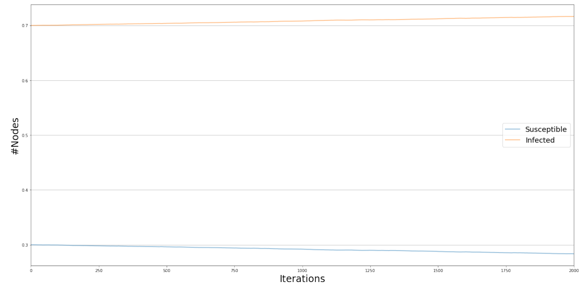


Figure 19: Sznajd model (WS)

that real world networks tend to be sparse, which may lead the algorithms to return wrong possible links as output. To overcome the first problem, we created a smaller graph, composed only by those nodes whom degree is equal or higher than 30.
At first, this analysis was conducted using simple unsupervised approach, but then, because of the poor results, we also tried to address the problem in a supervised manner, using different Machine Learning algorithms.
The following paragraphs will address the unsupervised link prediction using approaches that rely on node topology at first – and in particular on node neighbourhood – namely ***Common Neighbors***, ***Jaccard***, ***Adamic Adar***, and then one approach that use node rank, namely the ***SimRank***.
A summary of all the results is shown in Table 9.

Common Neighbors

The *Common Neighbors* algorithm assumes that linked pair of nodes, which are also connected to an high number of linked nodes, are likely to be linked in turn. This idea is formalized as follows

$$CN(u,v) = |N(u) \bigcap N(v)|$$

where N_u and N_v refers to the neighbourhood of u and v. The performance of this algorithm is debatable, as shown by the AUC-ROC in figure 20

Jaccard

This algorithm, a topology-based one like the *CN*, uses the Jaccard similarity formalized as follows to identify the number of common neighbours:

$$Jaccard(u,v) = \frac{N_u \bigcap N_v}{N_u \bigcup N_v}$$

Due to its similarity to the algorithm mentioned above, the results were quite similar and, unfortunately, it was also the worst predictor among all the topology-based ones.

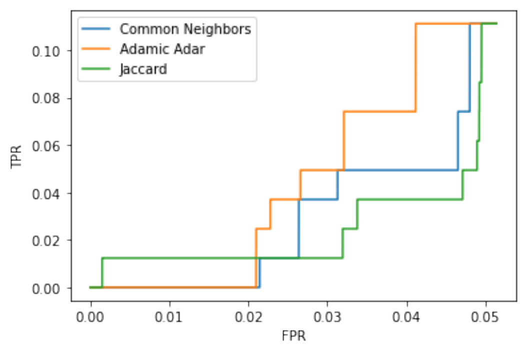
Adamic Adar

The *Adamic-Adar* algorithm is included in the topology-based set of algorithms. The main difference with *Common Neighbors* and *Jaccard*, is that it takes into account the node degree: so, what it happens is that nodes with lower degree have more impact, while hubs are less important for the link prediction. The following definition vehicolates the idea that nodes with an higher degree may occurr an higher number of times in the same neighbourhood than nodes with a lower degree, so they contribute less.

$$AA(u,v) = \sum_{z \in N_u \bigcap N_v} \frac{1}{\log k_z}$$

This method, like the others of the same category, shows a lot of difficulty in finding new possible edges.

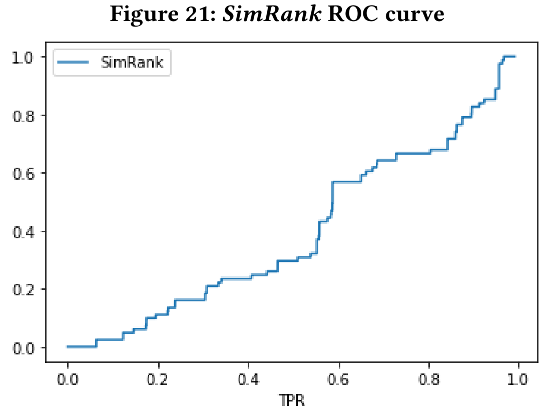
Figure 20: ROC curve (CN, JCC, AA)



SimRank

The SimRank (Similarity Rank) is a rank approach to link prediction. It computes the likelihood to see a new link between nodes (or users in this case) by analysing the similarity of their respective neighbourhoods. This approach returned better results, still low, though, as it can be seen in Figure 21.

Figure 21: *SimRank* ROC curve



Supervised Link Prediction

The supervised link prediction, that can be roughly described as the implementation of a machine learning model (i.e. for classification or regression) that predicts the potential new links, was conducted by using the traditional *scikit-learn* library. The first initial step was the the data preprocessing though, that was carried out using two different libraries: *StellarGraph*[13], that allowed to split the nodes into a training and a test set, and *node2vec*[14], that extracted the embeddings necessary to apply the machine learning algorithms using a biased second order random walk.

The nodes were split into training and test set following StellarGraph documentation[15] and the process returned 8586 nodes for the *training*, 2862 nodes for the *validation set* and, last but not least, 12722 nodes for the *test*. All of them were then embedded into vectors that were used for the parameter tuning of the models, the training and the test.

The results, as it can be seen in Table 10 are clearly better than the ones obtained with the unsupervised approaches. The *logistic regressor* is the weaker model, since it has an average score in precision and a AUC-ROC score around – that is for the AUC-ROC of a dummy predictor. Moreover, it was trained a traditional *Random Forest Classifier*. The algorithm was fine-tuned using a *GridSearch* featuring a *5-fold cross-validation* and then the model was trained setting ***entropy*** as purity measure, a *maximum depth* equal to 20 and a *number of estimators* equal to 200.

The results on the test are the highest obtained through these "traditional" supervised approaches: the *fine-tuned* RF classifier achieved a 79% in accuracy and precision, so it can be considered the best result among all the three classifiers. The SVM – that was fine-tuned and trained using a non-linear kernel (rbf) – instead, returned results below the ones of the RF, except for the recall.

8 TASK 5: OPEN QUESTION

In this last part of the report, we tried to cross Network Science and Natural Language Processing (NLP), in order to better understand which is the *leitmotif* behind each community and hopefully gain some interesting semantic insights about the way they interact with the others about this specific and wide topic.

At the beginning we focused on exploiting the borderline, "grey communities", those ones composed by people coming from different socio-cultural backgrounds that did not took a strong position about the debate (e.g. people that supported the anti-racist protests, but condemned violence itself). This

[13]StellarGraph documentation: https://stellargraph.readthedocs.io/en/stable/index.html
[14]node2Vec documentation: https://snap.stanford.edu/node2Vec/
[15]Link prediction with Node2Vec: https://stellargraph.readthedocs.io/en/stable/demos/link-prediction/node2vec-link-prediction.html

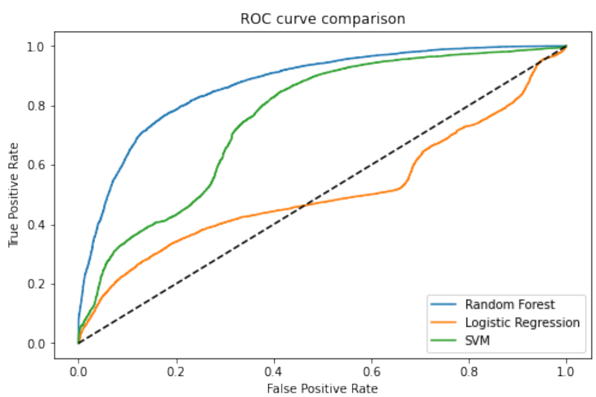
	Common Neighbors	Jaccard	Adamic Adar	SimRank
AUC-ROC	0.001	0.002	0.001	0.389

Table 9: Unsupervised predictors AUC-ROC

Table 10: Supervised predictors evaluation

	Logistic Regression	Random Forest	SVM
Accuracy	0.57	0.79	0.71
Precision	0.62	0.79	0.67
Recall	0.37	0.80	0.85
F-1 score	0.46	0.79	0.79
AUC-ROC	0.51	0.87	0.74

Figure 22: ROC curve of the supervised predictors



choice was driven by the major interests they have from a sociological point of view.

Preprocessing

The first problem was the choice of the communities: after a long evaluation between *pros* and *cons*, we opted to consider only the biggest communities detected by the Louvain algorithm. This decision was driven by the high number of communities and the stability of the network w.r.t. internal and external measures. In order to contain the main problem of the Louvain communities, which is tied to their high number and often the small number of users inside them, we filtered the communities and took only the comunities having at least 40 users inside them.
In addition to this, it was also decided to remove the tweets considered "uninformative", i.e. tweets that did not spread so much for various reasons and probably had a low *Impression* value also for Twitter[16]. Therefore, the tweets that did not meet the following, arbitrarily-chosen conditions, were not considered:

- Total number of likes >= 5
- Total number of retweets >= 3

These two thresholds were discussed a lot, since setting them higher than 5 and 3 respectively, removed too many records, so we decided that those values were a good compromise between informative and uninformative tweets. A sample of the communities after this filtering process is shown in Table 11.

Table 11: Community extension before and after the removal of "uninformative" tweets

Community	Before filtering	Filtered
community_0	4506	1506
community_1	3818	1377
community_2	2532	1059
community_3	2189	475
community_5	1706	569

Topic Modeling

The topic modeling task, in this unsupervised context, was carried out using using BERT, and in particular, a topic-modeling specific version of BERT called BERTopic, presented in [5] and implemented through Python in the *bertopic* library[17].
Using this kind of approach based on *transformers*, the *raw text* of the tweets was cleaned but not deeply preprocessed using the traditional NLP pipeline (i.e. sentence splitting, tokenization...), since it is not strictly required by this language representation model. This cleaning part was necessary to normalize tweets and set all of them to lowercase, remove URLs, emails, tags for audio and video, numbers and special characters, such as "&". At the very beginning, we also removed all the mentions and hashtags from the tweets, but later we decided to keep them also in the cleaned text, since they were highly informative to interpret the results.

[16]Impression is defined as *times a user is served a Tweet in timeline or search results*, see https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard
[17]GitHub repository: https://github.com/MaartenGr/BERTopic

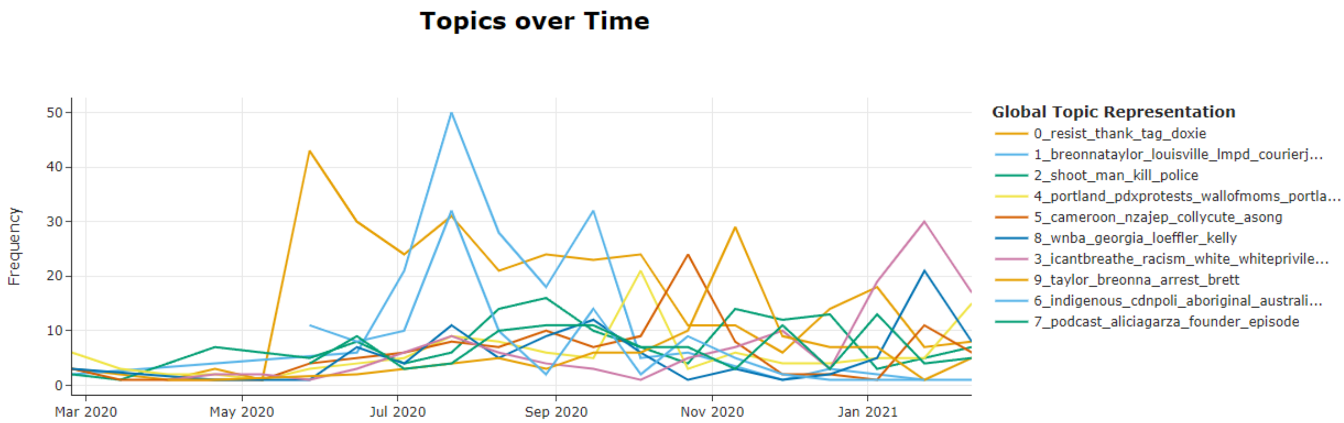


Figure 23: Topics extracted from the one-year span of *blm* tweets

Topics extraction. From the earliest results, it was possible to assess the lack of racist ideas in all the topics: this may be caused by the policies[18] on hate speech and violence followed and still applied by Twitter to limit the spreading of contents that trigger violence or hate. This first gained insight, forced us to redefine the *Open Question* that evolved into a deeper analysis of the topics that characterize each community, since not all the discussions are strictly about Black Lives Matter.

The first topic that was extracted by BERT was a general support to the movement, often focused on the main event that turned the spotlight on *BLM* on 2020, that is the George Floyd's death by Derek Chauvin's hands; this particular detail of a black person killed by a police officer is the basis also to other topics, in particular to the murder of Breonna Taylor – killed on March, 2020 because she was suspected of being involved in a drug trade – and the murder of the 12-years old guy Tamir Rice, killed in 2014 while he was playing with a toy gun. In addition to this, the Twitter users involved into these discussions often required the police defund.

The controversies about law enforcements are reflected on another hot topic, that is death sentences, often for black people, whose guilt is not always clearly defined and acts as a trigger for outrage and controversies. During this timespan, Twitter users discussed about Julius Jones, saved from the death sentence four hours before his execution: this event has been linked by users to the case of Howard Guidry, that is waiting on the death row for more than 20 year. The other two similar cases discussed on Twitter are the case of Pervis Payne, accused of murder too, and the one of Brandon Bernard, executed in December 2020 for the same reason.

According to what has been extracted with the *topic modeling*, the support to the movement became concrete by the means of demonstrations spread throughout the territory. The most significant were those ones of Portland, in which stood out the *Wall of Mom* group[19], and *Abolish ICE*, a political movement that demands the abolition of *U.S. Immigration and Customs Enforcement*. Other protests that were discussed a lot on Twitter were those ones in New York, i.e. in *Times Square*, and the block of the *Dan Ryan Expressway* by protesters.

Another major event that took place in New York and was also reported on Twitter was the creation of a mural in favor of the *BLM* in front of Trump Tower. It is interesting to notice how this last point it is strictly intertwined with another cluster of topics, solely political: there are various references to the competition between the two candidates in the US presidential elections, Joe Biden – and accordingly with him, Kamala Harris – and Donald Trump, with an evident support for the former by *BLM* supporters. In addition to exclusively American events, there are also other events – and scandals – involving foreign political figures: we find both the name of Boris Johnson – probably due to some statements against the protest methods of the *BLM* – and Nicolás Maduro, for a photo portraying him with the co-founder of the *BLM* movement, Opal Tometi.

Another relevant phenomenon is linked to the sport field, often because of the unclear positions assumed by sports club that sometimes clash with the firmer positions taken by individual players. A clear example is the case of the football player Colin Kaepernick, an activist against police violence and racism since 2016, that pushed the NFL to take clear anti-racist positions through his protests. Two other cases

[18] *Hate speech* Twitter policies: https://business.twitter.com/en/help/ads-policies/ads-content-policies/hate-content.html

[19] A group primarly formed by women and, in particular, mothers that demonstrated for Floyd

are the one of the WNBA women's football team, which members wore T-shirts in support of the *BLM* to silently protest against the senator – as well as club president – Kelly Loeffler for her anti-BLM stances.
Finally, the last collateral topic that emerged through *topic modeling* is tied with the history of colonialism and, consequently, of slavery: first of all surfaced the issue of the demolition of statues, in particular the one of Robert Geffrye in London – since it was considered a symbol of slavery – and the civil war, namely *Anglophone Crisis*, that takes place in Cameroon, internally divided into two factions – the Anglophone territories and the Southern Cameroons – with the first one fighting to achieve independence from the Camerooninan government.

An in-depth analysis of the four main communities

Willing to carry out an extensive study of the relationships established between the different groups of users, at first our work focused on the study of the five main communities, at least from a merely quantitative point of view. This approach was interrupted after the first application of BERTopic, which identified only generalist topics, such as a general disagreement against Trump or a support for *BLM* protest, that were not so interesting since they i) were also present in the smaller communities and ii) were probably overshadowing other more interesting topics. For this reason, we rediscussed again the task itself, in order to find another approach to the problem: as a result, it was finally decided to focus and to analyse on a finer-grain perspective those communities that are quantitatively smaller, but that are more interesting from a sociological perspective for being more informative and characterizing.
Furthermore, the BERTopic labeled some clusters of topics as *noise* – identified with the tag -1: with a more deeper insight, these are nothing more than themes that constitute a "core" of topics in support to BLM, without being truly specific (i.e. *#icantbreathe*). Similarly, the first topics identified by BERT in different communities are characterized by the same level of generality, that progressively disappeares until we reach clusters of topics that are a niche, but at the same time are more influential from a sociological point of view, since they'reable to kindle again those debates settled over time.

Topic Modeling on Community 2

This community is mainly focused on the comparison between Biden and Trump, with a clear preference for the first. This is evident by the hashtags extracted by BERT both general – like *#GDAmerica*[20] – and clearly sided – this is the case of the hashtag *#FBR*, acronym for *Follow Back Resistance*, used as a hallmark by a group of anti-Trump users on Twitter in order to identify themselves on this social network and be able to create a chain of mutual follow and a more clustered "Twitter community" inside a "Louvain community".
There are also other generalist themes that are instead focused on issues related but at the same time ancillary to the "main" *BLM* topic – from criticism and protests against police brutality, to the attack on the so-called *white privilege* that is intrinsic in this kind of violence. In addition to the more general topics just mentioned – i.e., the election, protests against Trump's policies on various fronts – in this community may be discerned also some other particular events. The first one is the installation of a yellow neon sign with "Black Lives Matter" written on it in Emeryville, in California, conceived by the engineer and artist Mauricio Bustos and his art collective "Crisis Labs" with the aim of spread the message to as many people as possible. The second one, concerns instead the statements of the Georgian Democratic candidate for the U.S. Senate, Rev. Raphael Warnock, opponent of the Republican Senator Kelly Loeffler: in fact, Rev. Warnock declared towards the end of October 2020 that racism is the pre-existing condition of America.
In the end, this community also includes the tweets posted by the satirical and goliardic account *TrumpDeathToll*, whose *bio* claims that this account contains the "daily count of deaths caused by Donald J. Trump's inept response to SARs-CoV-2 in the United States.".

Topic Modeling on Community 3

Besides discussions about US politics – and the Democratic Party – the users in this community talked a lot about *BLM* and the side effects of colonialism. A large portion of the users discussed about complex topics like the civil war in Cameroon – fought since October 2017 by Anglophone regions that try to gain independence from the current government – and the issue of the Tigray genocide in Ethiopia, which has been ongoing since the late 2020s.
A final interesting group included in this community discussed instead of the pro-BLM protests in Uganda and the arrests of some protesters[21].

Topic Modeling on Community 5

This community appears to be characterized by the prevalence, though not exclusively, of events and topics concerning the British area – including BBC articles related to the BLM. Among the topics not directly connected to the British context, the most characteristic ones are the protests in New York against law enforcement and the movements that

[20] *God bless America*

[21] The news on the ABC website: https://abcnews.go.com/International/americans-15-arrested-black-lives-matter-protest-uganda/story?id=71172060

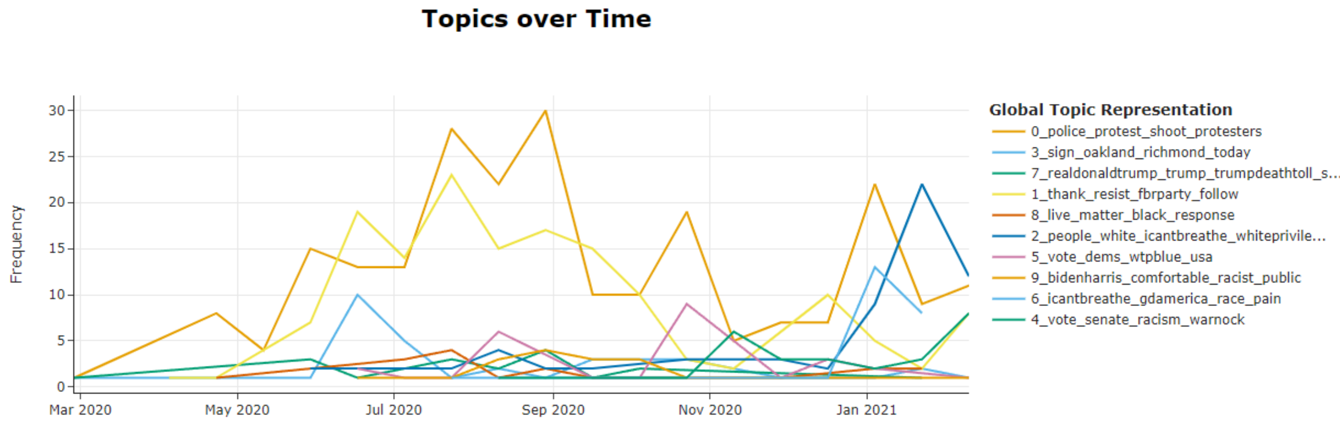


Figure 24: Topics extracted from *community 5*

demonstated in favor of the abolition of ICE (*Immigration and Customs Enforcement's*).
Another interesting topic, that on the timeseries of the topics given by BERTopic is set between the end of October, 2020 and the beginning of November, 2020, is the death of Walter Wallace Jr., a 27-year-old African-American man, killed by two police officers in Philadelphia. The man – that according to his family members suffered from mental illnesses – was killed by the police officers while he was having a mental health crisis, reason why his family members looked for the help of an ambulance with medical staff and police[22] This event led to fierce protests in Philadelphia, New York, Portland, accordingly to the words representative of this topic recognized by BERT.
Furthermore, it is interesting to note another identified topic, which is the request to fall the statue of Robert Geffrye placed at the entrance of the *Museum of the Home* in the Hackney district of London by the *Hackney Stand Up To Racism* (HSUTR) movement. The statue is placed there because Geffrye funded the hospices where the museum is located today. This topic is part of the major debate that has arisen thanks to Black Lives Matter regarding the fall of statues of colonialists and/or slavers. The fall of the statue of Edward Colston in Bristol set the theoretical and practical precedent behind the demands of the fall of the statue of Robery Geffrey, for which in July 2020 the museum asked for a consultation with the local inhabitants: almost 80% of the local population expressed their opposition to the presence of the statue at the entrance of the museum and therefore in favor of its removal[23]. The hesitancy of the museum in its removal, probably due also to economic issues([24]), eventually matured into the choice to move the statue in a less relevant place in the back of the building, with a plaque that explains, interprets and contextualizes the character and his life, allowing in this way also to confront the truth regarding the origin of the museum[25]. The trend of the topic mainly concerns the month of July 2020 and has a major peak in November 2020.

Topic Modeling on Community 8

This cluster contains a summary of the topics already presented in the other communities, such as the *AbolishICE* protests from *Community 5* or the death of Breonna Taylor, that emerged also during the analysis of the whole *corpus*. Exclusive topics of this community are, instead, the protests led by the *Wall of Moms* movement in Portland after the death of George Floyd, and the positions taken by the British British *Green Party* in favor of Black Lives Matter[26].

By observing the graph in Figure 25 it can be seen that the links between communities are denser between*community*

[22]Link to New York Times website about the event: https://www.nytimes.com/article/walter-wallace-jr-philadelphia.html; https://www.nbcphiladelphia.com/news/local/walter-wallace-jr-struggled-with-mental-health-issues-family-says/2575493/.

[23]Petition link: https://www.whatdotheyknow.com/request/683353/response/1648912/attach/html/2/Statue%20consultation%20report%20FINAL.pdf.html

[24]https://www.voice-online.co.uk/news/uk-news/2022/01/25/geffrye-must-fall-campaigners-step-up-fight-to-remove-slave-trader-statue/

[25]The Guardian website: https://www.theguardian.com/uk-news/2021/nov/18/museum-of-the-home-considering-moving-statue-slave-ship-owner-robert-geffrye

[26]Green Party website: https://www.greenparty.org.uk/news/2020/06/04/green-party-stands-with-the-black-lives-matter-movement-every-step-of-the-way/

Figure 25: Community 2 (orange), 3 (turquoise), 5 (green), 8 (pink) plot on Gephi.

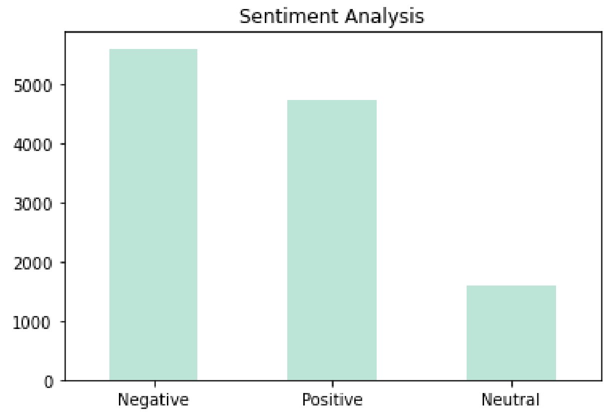


Figure 26: Sentiment analysis results on the tweets

3 and *community 5*: it is interesting to note that this corresponds also from a strictly semantic point of view, since the topics partially overlap – w.r.t. slavery and colonialism treated according two different perspectives. It is also interesting to notice what happens with *community 8*, that may be seen as bridge between communities, both topologically and semantically, since it is located in the middle of the other ones and the users inside discussed about topics that are in common with the other, bigger, communities.

Sentiment Analysis

The final step of the linguistic analysis was the *sentiment analysis*[27] part: the task was carried out using the VADER sentiment analyzer[28], that is trained and optimized for this specific kind of linguistic data; it was applied both on the whole dataset and the single communities analysed in the previous section.
The main problem of this dataset is the lack of tags that describes for each record the correct sentiment detected by an human being, so it is almost impossible to evaluate it correctly. It can be roughly said that most of the tweets, 5593 in total, are tagged as *negative* by VADER, followed by 4730 *positive* tweet and other 1585 tweets considered as *neutral*.
This results were then deepened by looking at the sentiment distribution in each community. Even in this case the pattern was the same, with a majority of negative-tagged tweets, followed by positive and neutral tweets: however, we noticed that there's a higher number of positive values in communities *8* and *2* than in the other ones.

9 FINAL DISCUSSION

In this work we carried out an analysis the of Black Lives Matter movement using as source of knowledge Twitter and the tools offered by *Network Science* and *Natural Language Processing*. We only focused on the events happened between 2020 and 2021, namely the murder of George Floyd and the Capitol Attack.
Firstly, we performed the explorative analysis of the network: among the most notable insights there are the the central nodes – according to different centrality measures – and the slight disassortativity of the network itself.

Then, in Section 4, we addressed to the *Community Discovery* problem with the aim to extract communities to solve the initial *Open Question*: at first, it was about exploiting the communities not clearly polarized towards racism/anti-racism, and how they interacted through tweets with the more polarized communities. In this part of the work, we tested various algorithms: their results were compared through internal and external evaluation metrics.
Later, the analyis was further deepened in order to observe the temporal evolution of those communities: as example, it was also analysed more in detail the evolution of one arbitrarily-chosen community.
Before moving to the *open question*, we attempted to put together the way opinions spread and changed over time, from positive ones to negative and vice-versa. In this respect, the most interesting results were obtained using a *Majority*

[27] *Sentiment analysis*, also known as *opinion mining*, is a subfield of NLP that tries to extract opinions and the positive, negative or neutral emotions that emerge from linguistic data
[28] GitHub repository: https://github.com/cjhutto/vaderSentiment

Rule model, since the dynamic opinion change followed was similar to the *Majority Rule* model applied on a *Barabási-Albert* network similar to the RW network.

Then, we tried different link prediction approaches in order to find out which nodes were more likely to have an edge connecting them. The best results were obtained using a supervised approach and traditional machine learning classifiers: the most interesting turned out to be the *Random Forest* classifier, which reached a 79% in accuracy, precision and F1-score and an 80% in recall.

The *open question*, as it was defined at the very beginning, was soon abbandoned since were not included explicitly racist tweets: their absence was justified by looking at the Twitter policies on *hate speech*. Consequently, we tried to gain other interesting insights from results obtained after the application of the *topic modeling* chosen algorithm, namely BERTopic, and the *sentiment analysis* implemented through VADER.

This led to two different insights: the first one, is that *community discovery* and network science in general, allowed to analyse large amounts of data; this allowed us to identify also events that, despite their secundary nature, still played a critical role in igniting some specific issues; they also revealed hidden sides of the same issue, with a different granularity level.

The second aspect, is that *BLM* had – and still has – a considerable impact in the legitimization of discussion and issues born as a result of the processes of decolonization that have characterized the second half of the 20th century. These consideration have converged in the intellectual field and labeled as *Post-colonial studies* but, were previously tied to a purely academic setting: *BLM* has clearly the merit, in our opinion, for having brought those ideas to the attention and sensitivity of a wider and general public.

As it regards the future developments of this work, it may be furtherly expanded by exploiting the differences (and the similarities) between a network composed by solely verified/not verified accounts. Another path to follow may be to use another dataset with additional information, i.e. labels that allow to evaluate the sentiment analysis results or geographical informations. There is also the possibility to develop this analysis by appling other algorithms, such as other classifiers to improve even more the results obtained in the supervised *Link Prediction.*

10 ACKNOWLEDGEMENTS

We wish to thank the debugging duck Pippo da Pisa and their nearest neighbours for their continuous help and support through this journey.

REFERENCES

[1] Albert-László Barabási and Márton Pósfai. 2016. *Network science.* Cambridge University Press, Cambridge. http://barabasi.com/networksciencebook/

[2] Michele Coscia. 2021. The Atlas for the Aspiring Network Scientist. *CoRR* abs/2101.00863 (2021). arXiv:2101.00863 https://arxiv.org/abs/2101.00863

[3] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. 2012. DEMON: A Local-First Discovery Method for Overlapping Communities. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing, China) *(KDD '12).* Association for Computing Machinery, New York, NY, USA, 615–623. https://doi.org/10.1145/2339530.2339630

[4] Felix Gaisbauer, Ariana Strandburg-Peshkin, and Helge Giese. 2022. Local Majority-with-inertia Rule Can Explain Global Consensus Dynamics in A Network Coordination Game. *Social Networks* 70 (July 2022), 218–227. https://doi.org/10.1016/j.socnet.2022.01.013

[5] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (2022). https://doi.org/10.48550/ARXIV.2203.05794

[6] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9, 6 (June 2014), e98679. https://doi.org/10.1371/journal.pone.0098679

[7] Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (May 2012), 1–167. https://doi.org/10.2200/s00416ed1v01y201204hlt016

[8] M. E. J. Newman. 2002. Assortative Mixing in Networks. *Phys. Rev. Lett.* 89 (Oct 2002), 208701. Issue 20. https://doi.org/10.1103/PhysRevLett.89.208701

[9] M. E. J. Newman. 2003. Mixing patterns in networks. *Physical Review E* 67, 2 (feb 2003). https://doi.org/10.1103/physreve.67.026126

[10] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. 2019. CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science* 4, 1 (July 2019). https://doi.org/10.1007/s41109-019-0165-9