

# I like you if you are like me: How the Italians' opinion on Twitter about migrants changed after the 2022 Russo-Ukrainian conflict

**Luca Palla**  
l.palla5@studenti.unipi.it  
Student ID: 533605

**Martina Sustrico**  
m.sustrico@studenti.unipi.it  
Student ID: 533252

**Giulio Cordova**  
g.cordova@studenti.unipi.it  
Student ID: 588294

## ABSTRACT

Immigration has been at center of the Italian public debate for the last 10 years, letting far-right parties foment fear among people, which eventually led to feelings of hate and racism. The Russian invasion of Ukraine let European white refugees flee, which differed from the stereotype of refugee in the Italian collective imagination. The goal of this study conducted with Network Science tools<sup>1</sup> is to determine whether the Russo-Ukrainian conflict of 2022 had an impact on the Italians opinion about refugees, i.e. it was investigated if the proximity of the conflict and the ethnic similarity of the refugees made Italians change their mind about refugees.

## KEYWORDS

Refugees, Russo-Ukrainian Conflict, Migrants, Social Network Analysis, Twitter, Opinion Dynamics, Community Discovery

## ACM Reference Format:

Luca Palla, Martina Sustrico, and Giulio Cordova. 2022. I like you if you are like me: How the Italians' opinion on Twitter about migrants changed after the 2022 Russo-Ukrainian conflict. In *Social Network Analysis '22*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

## <sup>1</sup>Project Repositories

Data Collection: [https://github.com/sna-unipi/sna-project-2022\\_cordova\\_sustrico\\_palla/tree/main/data\\_collection](https://github.com/sna-unipi/sna-project-2022_cordova_sustrico_palla/tree/main/data_collection)

Analytical Tasks: [https://github.com/sna-unipi/sna-project-2022\\_cordova\\_sustrico\\_palla/tree/main/network\\_analysis](https://github.com/sna-unipi/sna-project-2022_cordova_sustrico_palla/tree/main/network_analysis)

Report: [https://github.com/sna-unipi/sna-project-2022\\_cordova\\_sustrico\\_palla/tree/main/report](https://github.com/sna-unipi/sna-project-2022_cordova_sustrico_palla/tree/main/report)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. *SNA '22, 2021/22, University of Pisa, Italy*

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$0.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Due to its geographical position, Italy has historically been one the most used routes for undocumented migrants, which count both the ones migrating for economic reasons and asylum seekers. The Arab Spring in 2011 and the Syrian War in 2014 were factors that caused a significantly increase of the boat-dwellers' landing on the Italian shore[28].

As the number rose, Italy's far-right parties could frame the debate on migration in terms of "threats" and "security"[14], making the issue a key element on their political platforms, as reported in some example tweets[19][20] of Figure 1.



(a) Fratelli d'Italia's tweet calling for a "naval blockade" to "protect Italy's borders"

(b) A tweet by Lega on keeping Italy "secure" from migrants

Figure 1: Two example tweets of the electoral campaign of Summer 2022 by Fratelli d'Italia (Brothers of Italy) and Lega (League), that show the focus on "threat" and "security"

This resulted in echo-chambers that provided a distorted image of the real numbers about migrants and asylum seekers: according to a Ipsos study of 2019, the perceived presence of migrants on Italian territory is 31% compared to the real data 9%[13]. Furthermore, 63% of Italians admitted that they

were exposed to fake news on the theme Immigration, leading to believe that the majority of the crimes were committed by migrants (33 % of the interviewed[13]), and that immigration establish a threat for Italy (57% of the interviewed[13]).

As described in an article by Dylan Patrick McGinnis on the Yale Review of International Studies[18], the Matteo Salvini's (leader of Lega) - and now Giorgia Meloni's (leader of Fratelli d'Italia) - strategy to put Immigration policy to the far-right relies mostly on creating an 'us versus them' dichotomy to 'other' migrants and on blaming immigrants for Italy's poor economy.

According to these theses, the presence of a near conflict such as the Russian invasion of Ukraine, could let Italians empathize with the refugees; and that the similarity in terms of ethnicity could dismantle Salvini's 'us versus them' dichotomy.

In this paper it was therefore investigated the opinion dynamics of Italians about refugees, with a specific focus on the comparison and shift of opinions before and after the Russo-Ukrainian conflict of 2022.

## 2 DATA COLLECTION

For the goal of this paper, it was decided to create a new dataset. In fact, it was impossible to find a dataset which would fit our necessities. As described in this section, it was chosen to create two different datasets:

- one used to build and analyze the graph;
- one used to train a classifier for the labelling of the tweets, regarding the Opinion Dynamics task.

### Selected Data Sources

For the scraping of the data, it was chosen to use Twitter. Twitter works as a public square for the public debate, in which politicians, journalists and everybody else can intervene for the creation of a discussion. In particular, it was used the library snsScraper[15], which provides a simple scraping query in order to build a dataset based on the twitter data and metadata.

### Crawling Methodology and Assumptions

*Graph Creation.* Regarding the graph creation, it was decided to scrape all the tweets containing the hashtags #rifugiato, #rifugiati, #profugo, #profughi, #migrante, #migranti, in a temporal window from the first September 2020 until the first September 2022. As for the attributes, the following ones were scraped:

- "User": username of the account writing the tweet;
- "Language": language of the tweet;
- "Date\_Tweet": date on which the tweet was written;
- "Tweet\_Id": number of identification of the tweet;
- "Tweet": actual content of the tweet;

- "Hashtags": list of hashtag in the tweet;
- "Conversation\_Id": number of identification of the conversation (all tweet replying to the same tweet have the same conversation id);
- "In\_reply\_To": account to which the tweet answer. If the tweet is the start of the conversation, this value is set to "User".

The first scraping produced 71735 tweets, for a total of 13575 unique Users and 69580 unique Conversations. These data had to be cleaned regarding the users subject to spam, i.e. the duplicated tweet from the same user were removed. Also all the non-Italian written tweets were removed from the dataset.

Starting from the conversation id, a second scraping was performed, with the goal of reconstructing the whole conversation in which there is at least one tweet containing a keyword (i.e. a tweet that was already scraped). After the same spam elimination, it was decided to remove also the tweets of the people who were answering to themselves. After these cleaning procedures, in the dataset were left a total of 48262 unique Users.

With these data, the graph was created. The users were chosen as nodes, while the replies between each user count as directed weighted link (based on the number of interactions between the users). Using the graph visualizer Gephi[2], it was observed that some source tweets of people acting as hubs were missing in the network, so another last scraping was necessary. Lastly the Users without any information, i.e. the users who were in the column "In\_reply\_To", but not the in the column "User", were eliminated.

As a part of Data Understanding, it could be interesting to show how the number of tweets in our dataset change during the years, depicted in Figure 2.

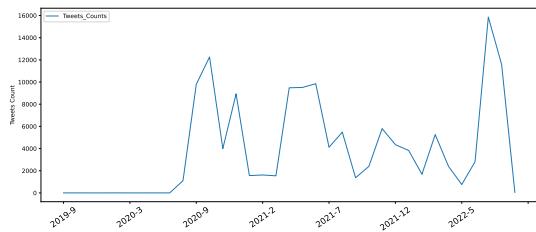


Figure 2: Number of tweets vs time

One can notice how the counts of tweets have not an uniform distribution through time. This means that the topic of immigration is not always discussed, but rather has spars peaks correlated with news or source tweets by influencers. For example, the last peak is easily identifiable with the Italian electoral campaign of 2022, as immigration was one of the most discussed topics.

*Dataset for classification in the opinion dynamics.* In addition to the dataset used for the graph creation, it was chosen to create another one, in order to help with the opinion dynamics task. It was decided to scrape all the tweets from September 2018 to September 2022 written in Italian, containing the hashtags #portichiusi (closed harbors) and #restiamoumani (stay human). The first was used by Matteo Salvini and other far right-wing people for an anti-immigration campaign[6], while the second was build by leftists and pro-immigration exponents as a response to #portichiusi. The hashtag #restiamoumani is also often associated with the international one #RefugeesWelcome. On these tweets was applied a similar cleaning procedure to the first scraping. It was also noticed that out of the 131877 scraped tweets, only 808 regards the refugees from Ukraine.

In this new dataset, the label 1 was associated to the tweets containing only the hashtag #restiamoumani, counting them as pro-refugees tweets. To the tweets containing only the hashtag #portichiusi, the label 0 was associated, i.e. counting them as anti-refugees tweets.

### Nodes labeling

The obtained dataset was used in order to train a classification model to label the original tweets; to do so, various different classification algorithms were tested. This dataset is slightly balanced towards a pro-refugees class, since the 1 label counts 67054 entries and the 0 label only 43542. When training the model, both oversampling and undersampling techniques were applied to improve the accuracy. As a summary of the analysis the calibration plots reported in Figure 3. Logistic Regression Model gave the best performance, both in terms of accuracy (0.78) and in terms of probability calibration.

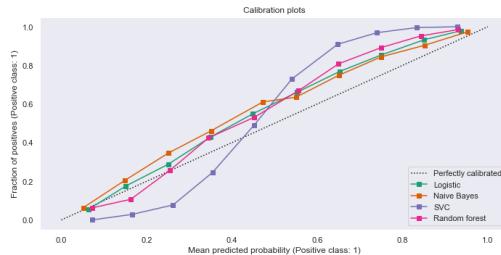


Figure 3: Calibration Plots

Starting from binary tweet classification, the variable "user label" was created by assigning a continuous value to all users based on the number of positive tweets they wrote (labeled 1) out of the total number of tweets. Basically, the closer the obtained mean is to 1, the more positive the user's opinion is (pro-refugee) and vice versa.

### 3 NETWORK CHARACTERIZATION

As described in the previous section, the network from Real World (RW) is composed by the User acting as nodes and tweet replies as weighted links. In order to create this graph, it was decided to use the library *NetworkX*[10]. Visualizing the graph in *Gephi*, it was noticed the abundant presence of isolated nodes, which were considered not interesting for the analysis. Also, some algorithms used to characterize the network are computationally expensive. For this reason, it was decided to work only with the giant component and to render the graph undirected and unweighted for this task. At the end of the cleaning one counts 46978 nodes and 88029 links. Our RW network has a density of  $7.98 \cdot 10^{-5}$ , an average Clustering Coefficient of 0.078 and an average shortest path of 3.39. The nodes acting as hubs are 93, while the Degree and Clustering Coefficient distributions can be seen in Figures 4 and 5.

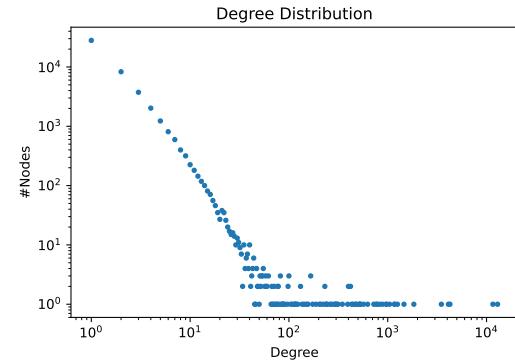


Figure 4: Degree Distribution of RW Network

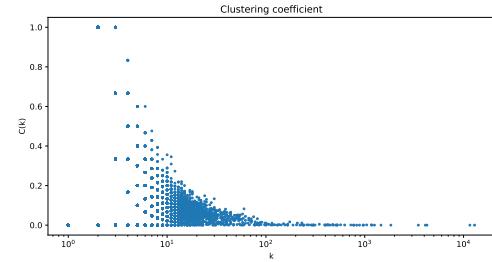


Figure 5: Clustering Coefficient Distribution of RW Network with respect to the degree  $k$

The Centrality Analysis was performed using different measures, in particular: Degree, Betweenness, Closeness, Harmonic, PageRank and Eigenvector. The Degree, PageRank and Eigenvector Centrality produces similar results. Between these three it was chosen to show the Degree Centrality, depicted in Figure 6. As one can see, the most important

nodes for this measure are Matteo Salvini (Lega) and Giorgia Meloni (Fratelli d'Italia). It follows a list of public figures, like other politicians (Giuseppe Conte, Enrico Letta, Luigi Di Maio) and journalists (Corrado Formigli, Roberto Saviano).

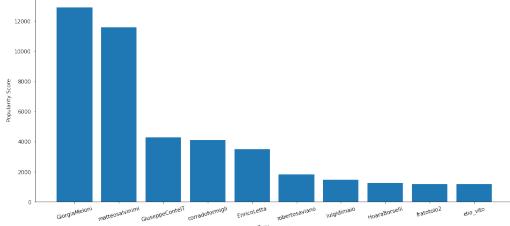


Figure 6: Degree Centrality

More interesting is the bar plot of the Harmonic Centrality, that can be seen in Figure 7. In this plot, one can see that private users can be almost as central as public figures.

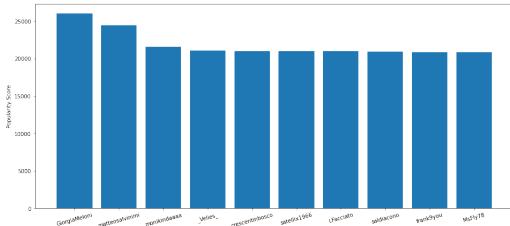


Figure 7: Closeness Centrality

To measure the homophily of the graph, it was used the assortativity measure  $R$  implemented by the library *NetworkX*, which uses the Pearson correlation coefficient between pairs of linked nodes' degrees. For this reason, the assortativity may also be considered a correlation measure between nodes.

Our RW network is classified as slightly disassortative, because of  $R = -0.2$ . This can be explained by the behaviour of small nodes that tend to connect with a greater probability with hubs, and not with other small nodes. This effect can also be seen in Figure 8, which depicts the distribution of the average nearest neighbor degree of nodes with degree  $k$ .

### Comparison with Erdős-Rényi, Barabási-Albert, Watts-Strogatz Models

In order to compare the RW network with the theory models, three synthetic graphs were generated, in particular:

- Erdős-Rényi (ER) Graph[5]: in order to obtain a similar number of Nodes and Edges, it was set  $N=46978$  and  $p=0.0000787$ , for a total number of edges of  $E=86768$ ;
- Barabási-Albert (BA) Graph[1]: as above, it was set  $N=46978$  and  $m=2$ , making  $E=93952$ ;

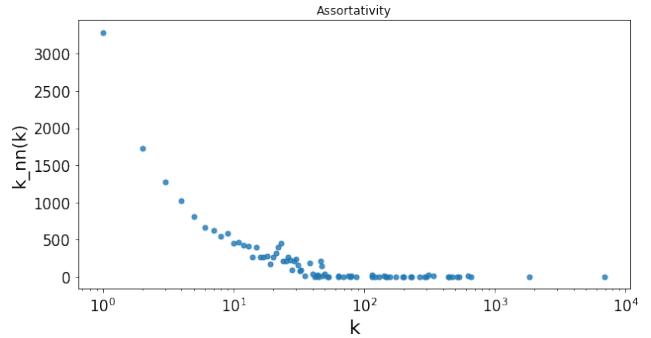


Figure 8: Assortativity Distribution with respect to the degree  $k$

- Watts-Strogatz (WS) Graph[29]:  $N=46978$ ,  $k=4$ ,  $p=0.8$ , leading to  $E=93956$ .

The other basic characteristics such as Average Degree (Degree) Connected Components (CC), Density, Average Clustering Coefficient (ACICo), Average Path Length (APL) of the simulated networks are summarized in Table 1.

Table 1: Characterization of RW and Simulated Networks. It shows number of Nodes, number of Edges, Average Degree (Degree), Density, Average Clustering Coefficient (ACICo) and Average Path Length (APL).

	Nodes	Edges	Degree	CC	Density	ACICo	APL
RW	46978	88029	3.75	1	$7.98 \cdot 10^{-5}$	0.078	3.39
ER	46978	86768	3.69	1214	$7.86 \cdot 10^{-5}$	$5.93 \cdot 10^{-5}$	1.66
BA	46978	93952	3.99	1	$8.51 \cdot 10^{-5}$	0.0014	9.21
WS	46978	93956	4.0	1	$8.51 \cdot 10^{-5}$	0.0044	2.38

Comparing the degree distribution in Figure 9, one can notice how the RW and the BA model follow a similar shape. This behaviour is explained by the nature of the BA model itself, since it reproduces preferential attachment and hubs, which play a very important role in the Real World Networks. The degree distribution of ER and WS are both Poissonian and tend to go to zero quickly, as well depicted in Figure 9.

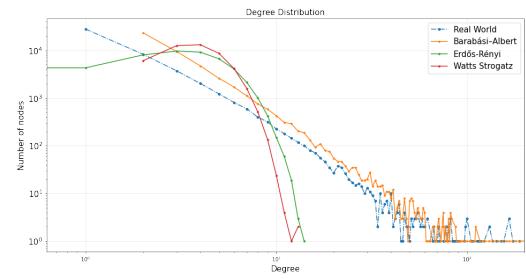


Figure 9: Degree Distribution Comparison

Regarding the path analysis, the algorithms are very expensive in terms of time of execution. For this reason, it was decided to build a sampled graph of the ER network of 15000 nodes and focus only on the giant component. In order to speed up the analysis, the results may then result biased by this choice.

#### 4 TASK 1: COMMUNITY DISCOVERY

Through discussion or searching information, people are influenced by their neighbours and by the environment around them. For that reason, studying how users were distributed among communities during the two years of observation is useful to understand the evolution of the network. Defining what a community is and how it born is not a trivial task. This section will show different ways to detect them based on different strategies. Two different kinds of community discovery methods were run:

- the first one refers to a static view of graph where the network is not supposed to be influenced by the time dimension;
- the second one is instead considered dynamic for the opposite reason. Few snapshots will be used in order to achieve this goal.

In both cases, the graph links were considered undirected and weighted because most of the algorithms do not support directed links.

##### Community evaluation's formulas

During the evaluation phase internal and external evaluation was made using CDLib modules. In order to uniquely identify formulas used in this process, a brief description is given:

- **Internal evaluation:** fitness functions allow to summarize the characteristics of a computed set of communities.
  - **Conductance:** it is the fraction of total edge volume that points outside the community. Lower values are preferable than higher.

$$f(S) = \frac{c_S}{2m_S + c_S} \quad (1)$$

- **Newman-Girvan Modularity:**[9] it is the difference of the fraction of intra community edges of a partition with the expected number of such edges if distributed according to a null model. In other terms, modularity compares the real network structure with a corresponding one where nodes are connected without any preference about their neighbours. The aim is to maximize it.

$$Q(S) = \frac{1}{m} \sum_{c \in S} \left( m_c - \frac{(2m_c + l_c)^2}{4m} \right) \quad (2)$$

- **External evaluation:** it is useful to compare different graph partition to assess their resemblance.

- **Adjusted Mutual Information:** it is an adjustment of the Mutual Information. It accounts for the fact that the MI is generally higher for two clustering with a larger number of clusters, regardless of whether there is actually more information shared.

$$AMI(U, V) = \frac{MI(U, V) - E(MI(U, V))}{\max(H(U), H, V) - E(MI(U, V))} \quad (3)$$

where U and V are clusters;

- **Overlapping Normalized Mutual Information LFK:** it is an extension of the Normalized Mutual Information (NMI) score to cope with overlapping partitions. Higher values are preferable than lower ones. It is a variant implemented by Lancichinetti [17]

- **Normalized F1 Score:** it computes a score of the optimal algorithms matches among the partitions in input. It works on overlapping/non-overlapping complete/partial coverage partitions. Higher values are preferable than lower ones[25].

- The last measure implemented was **Jaccard** computed as:

$$J(U, V) = \frac{|U \cup V|}{|U \cap V|} \quad (4)$$

##### Static community discovery

The algorithms used in this section – using CDLib[7] modules - regards three different ways of constructing communities:

- Crisp communities: nodes belong to only one community (Label propagation[21], Louvain[3], Leiden[27] was implemented)
- Overlapping communities: nodes can belong to different communities simultaneously (Demon[4], Angel[22], K-clique were implemented)
- Fuzzy communities: each node has assigned a probability to belong in a community. (Principled clustering was implemented)

During the hyperparameter tuning phase a grid search was run for those algorithms that have a specific metric to maximize or minimize. For the others a script was implemented in order to compare conductance and *Newman-Girvan modularity* and choose the best trade-off between them. Table 2 shows the result of the most informative metrics used to evaluate communities obtained during the analysis.

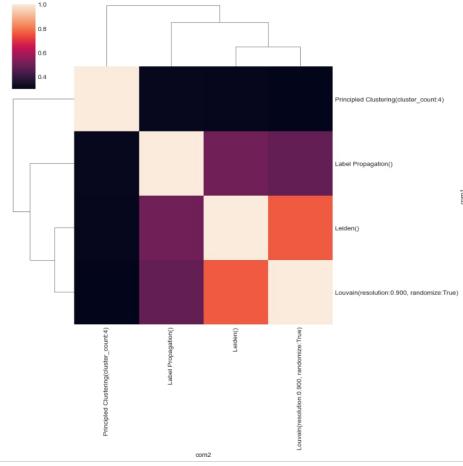
Based on the table, Louvain, Leiden and principled clustering performed better than other in both the metrics.

Interesting for the analysis is a comparison between the different communities returned, in order to assess if they are matching or not. The way in which the comparison was done

**Table 2: Summary of parameters, results and performance of the algorithms used for community discovery**

	Parameter(s)	#Communities	Node coverage	Conductance	NG Modularity
<b>Label propagation</b>	none	1781	1	0.8461	0.4550
<b>Louvain</b>	resolution= 0.9	31	1	0.375	0.5874
<b>Leiden</b>	none	40	1	0.3761	0.5952
<b>Demon</b>	$\epsilon = 0.1$ , min_com_size= 3	18	0.1694	0.9854	0.05
<b>Angel</b>	t= 0.1, min_com_size= 6	3	0.1054	0.8788	0.2507
<b>K-Clique</b>	K= 3	912	0.1926	0.9237	0.0141
<b>Principled Clustering</b>	num_communities= 4	4	1	0.2972	0.4491

takes in consideration the adjusted mutual information for communities that do not overlap and the overlapping normalized mutual information for communities that overlap. This analysis was done knowing that the results may be biased because some algorithms optimize different metrics. For this reason, it was difficult to compare the algorithms in terms of efficiency. The most significant similarities - excluding the one between Louvain and Leiden which was predictable - were between the non overlapping communities. In Figure 10 is visualized the comparison of them using an heatmap.

**Figure 10: Heatmap of different Community Discovery algorithms used**

This static analysis has the advantage of being easier to implement but has a lot of limitation as the impossibility to detect community events. In real networks, communities change or evolve at every timestamp and identifying their transformation is helpful also to understand how users change their opinions, friendship etc. To achieve this goal, a dynamic community discovery is implemented.

### Dynamic Community Discovery

For the task of dynamic community discovery, it was decided to divide the dataset in semesters and then the giant

component was extracted for each of them. In order to be coherent during the split, all the tweets related to a conversation started in the given semester are also considered inserted in that semester. The number of tweets in each semester are respectively 17440, 17387, 10090, 18205 denoting a uniform distribution, except for the third period. Being aware of the results reached in the static analysis, the algorithms selected for this task were Leiden and Principled Clustering. The idea was to treat the different semester as a static snapshot and perform on it the selected community discovery's algorithm. The purpose was to evaluate whether there were matches between the created temporal clusters and investigate the communities life events occurred in the meanwhile. As for the static analysis an hyperparameter tuning was needed for Principled Clustering in order to choose the best trade-off between conductance and modularity. In Table 3 are returned the parameter configuration and the results given by the comparison between communities of adjacent snapshots in terms of stability using Normalized F1 score.

**Table 3: Results and performance of Leiden and Principled Clustering (PC) on the four different snapshots**

Snapshot	# Communities		Normalized F1 Score	
	Leiden	PC	Leiden	PC
Sep 2020 – Feb 2021	60	8	-	-
Mar 2021 – Aug 2021	38	9	0.001914	0.007343
Sep 2021 – Feb 2022	46	8	0.000493	0.016666
Mar 2022 – Aug 2022	35	9	0.001925	0.010277

As it is evident from the data, principled clustering performs better in terms of stability although results are below the expectation. Low stability could be explained by making a deeper analysis on communities' composition and on snapshot composition. Using Jaccard as a metric, it was investigated how many users that were present in a snapshot were also present in the following one, and how many users that were present in the first snapshot were also available in the last one. Data showed that on average only 13% of the users are shared between snapshots, while the others are totally

new ones. This inhomogeneous conformation could create a great difficulty in detecting stable communities. With the aim to investigate communities life events a Sankey Plot was constructed, regarding a sample of matching communities returned by principled clustering. According to Wikipedia, Sankey diagrams are a type of flow diagram in which the width of the arrows is proportional to the flow rate[16]. As one can see in Figure 11 different events are present like born and merges.



**Figure 11: Sankey plot of the communities life through time**

With the aim to improve the stability of clusters it was chosen Tiles [24]. Tiles is based on an online method which does not partition the full temporal graph, but tries to build and maintain communities in an online fashion following the rising and vanishing of new nodes and edges. In this way, communities' changes were evaluated each time using a few numbers of new users giving a smooth evolution of clusters. As before, four snapshots were used to create the dynamic network. The results achieved are given Table 4.

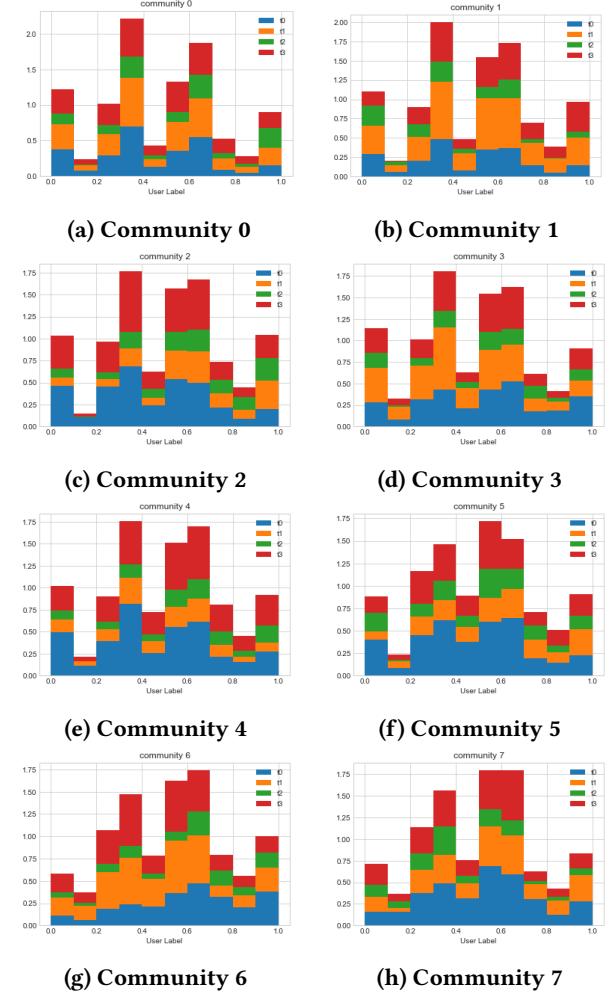
**Table 4: Summary of Tiles**

Snapshot	#communities	Stability trend
Sep 2020 – Feb 2021	587	-
Mar 2021 – Aug 2021	1040	0.1445
Sep 2021 – Feb 2022	1254	0.2171
Mar 2022 – Aug 2022	1627	0.2106

As expected, Tiles produces communities that are more stable than other algorithms; the main problem is represented by the huge number of communities created that make their representation and analysis very complex. For the next task, when communities discovered are involved, they are considered as those created by the principled clustering algorithms despite being aware of problems related to non-deterministic ones. This decision was principally taken due to the possibility of better managing them when a qualitative analysis on users were needed.

Lastly, it could be interesting to study how the user label distributed in each community through time. As shown in Figure 12, one can see that a community does not really represent a definite opinion. This behaviour is similar to all the distributions, and it is confirmed by the high variance and similar mean of each distribution. This is not surprising, since the communities are built by topological assumption

and not by taking in consideration the user label. This may be a hint that a community represents a topic, i.e. users with different opinions interact with each others over a defined topic. Moreover, it can be noticed in Figure 12 that the user label distribution appears to be stable through time, i.e. the changes of the bins are small.



**Figure 12: Distribution of user label for the communities in the all timestamps**

## 5 TASK 2: OPINION DYNAMICS

In this section some Opinion Dynamics models have been explored, in order to shape the way the opinions about refugees spread in the network. For this purpose, three discrete models were implemented from *NDlib* python library[23]: *Voter Model*[11], *Majority Rule*[8] and *Sznajd Model*[26]. These models were ran on the strongly connected component of the graph, with 46978 nodes. Visualizing the network on *Gephi*, it was noticed that 50% of the users had an opinion

between 0.5 and 1 and the remaining 50% had an opinion between 0 and 0.5; the results were then compared with those of two synthetic models (the *Barabási-Albert* and the *Watts-Strogatz* networks), by applying the same algorithms and parameters. Although opinions in the network are expressed continuously, for this task it was decided to run only discrete models of opinion dynamics in order to have simpler results to read and interpret. A more detailed analysis of how real user opinion changes over time is presented in the open-question section.

### Voter Model

The *Voter* model is one of the simplest models of opinion dynamics. It is possible to imagine that each node in the graph has a discrete opinion ( $\pm 1$ ). At each timestamp the opinion of a random node  $i$  is changed (or confirmed) with the opinion of a neighboring node  $j$  -also taken randomly. In our context, the model was run for 2000 iterations on the RW graph, starting from a perfectly balanced situation between infected and susceptible nodes (fraction 0.5). The results were then compared with those obtained from the same model with the same parameters applied to the synthetic graphs.

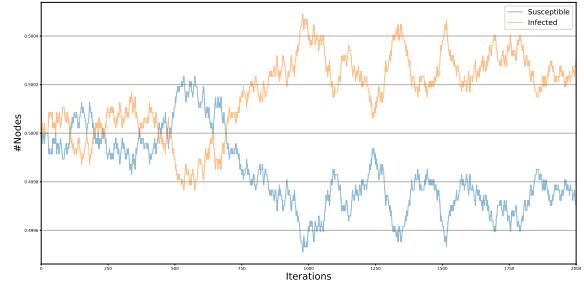
The *Voter* model in the RW shows a fluctuating initial condition, until, as the iterations grow up, fluctuations tend to decrease. The same exact behavior is not observed in either of the two synthetic graphs, although the BA graph approaches the RW network trend, even if with the opposite opinion, as shown in Figure 13.

### Majority Rule model

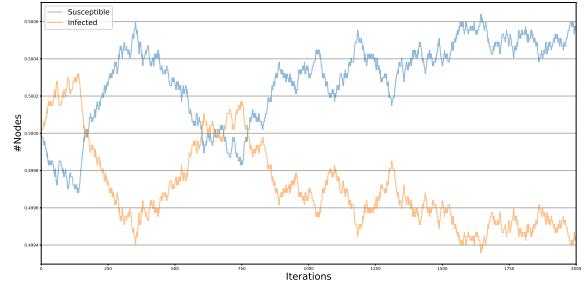
The *Majority Rule* model is one of the discrete opinion's model, in which each agent has a discrete opinion  $\pm 1$ . The difference with *Voter* lies on *social inertia* concept: i.e., an individual's hostility to change its social group. The model, starting from a random group of nodes with their opinion, assigns each node the majority opinion. Again, the model ran for 2000 iterations starting from a balanced division of opinions. In Figure 14 it is possible to see the comparison between Real World network and Barabási-Albert ones. The trend of opinion diffusion observed in the RW network is characterized by a perpetual exchange of opinions between nodes, throughout the execution of the model. A completely different situation is observed in the two synthetic networks, where the bias of positive opinion is sharp from the outset.

### Sznajd

In this context each agent has an opinion  $\pm 1$ . At each timestamp, a pair of neighbouring agents is selected and, if their opinion coincides, all their neighbours take that opinion. Given the operation of the model, there is an obvious immediate prevalence of negative opinion in the RW network, as depicted in Figure 15a. The situation is different regarding

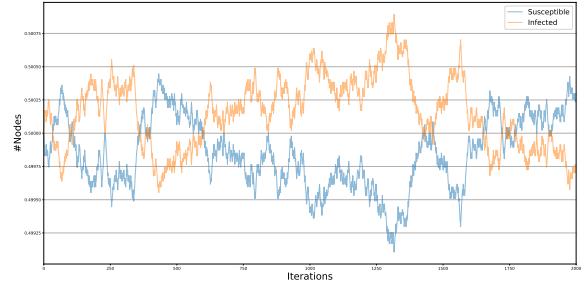


(a) Real World Network

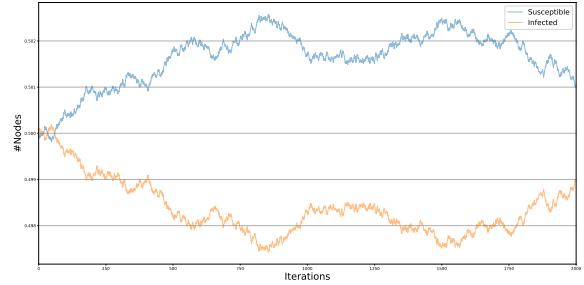


(b) Barabási-Albert Network

Figure 13: Voter Model



(a) Real World Network



(b) Barabási-Albert Network

Figure 14: Majority Rule Model

the BA network, in which at the end positive nodes prevail, as shown in Figure 15b. To conclude, regarding WS network, the distribution of opinions remains very balanced during the

first 1000 iterations, after which an increase in the negative class can be observed in Figure 15c.

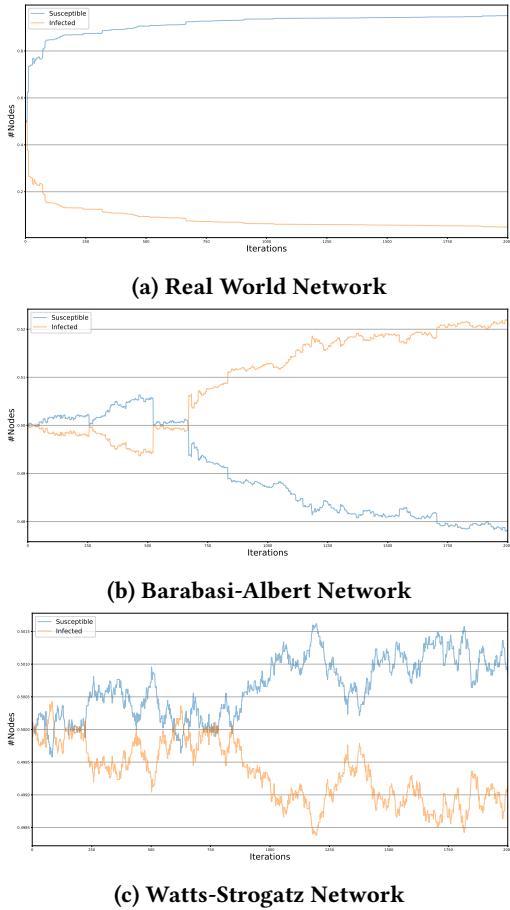


Figure 15: Sznajd Model

## 6 TASK 3: OPEN QUESTION

In this last task, we mostly used the results of the last two ones in order to investigate our main question, i.e. if the Russian invasion of Ukraine had an impact on the opinions of Italians about migrants and refugees.

As a preliminary study, it would be interesting to study if a tweet regarding Ukraine has a more positive sentiment compared to a tweet regarding migrants. In order to do so, the distribution of the user label (calculated in the nodes labeling subsection) was plotted in two cases: if the tweet contained the substring ucr or not.

As one can see, in Figure 16a the distribution is massively shifted towards the 1 class, i.e. people having a positive sentiment; while in Figure 16b the distribution tends to be more uniform compared to the other one. This behaviour can also be seen in the mean of the user label cases: in the

first case the mean is 0.72, while in the other 0.51. This latter result shows how the tweets regarding Ukraine tend to have a more positive sentiment compared to the others.

In order to look at the change of opinions by the users, it was decided to use the Sankey Plot for a visual representation. This is particularly useful in our situation, since one wants to characterize the amount of shifts in our graph between two or more temporal snapshots. For the creation of this kind of plot, the library *Plotly*[12] was used.

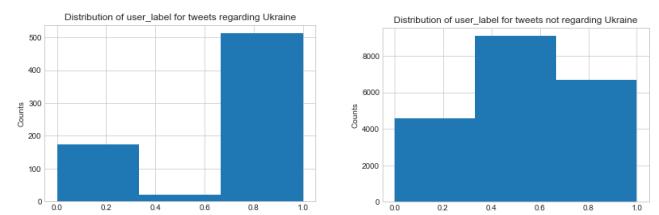
The first necessary step to compare the opinions before and after the Russo-Ukrainian conflict is to split the dataset in two frames, using the date of the tweet as a discriminator. In particular, it was chosen to use the date of the Russian invasion of Ukraine (24/2/2022) as split condition. In order to have more statistically relevant results, it was chosen to apply another condition, i.e. it was required that each user wrote at least two tweets in every temporal snapshot. The user label is then recomputed as the mean of each tweet label in each temporal frame. Furthermore, it was decided to bin the user label, creating four categories bins in which the user label is included in:

- 0-0.25: negative sentiment about migrants;
- 0.25-0.5: slightly negative sentiment about migrants;
- 0.5-0.75: slightly positive sentiment about migrants;
- 0.75-1: positive sentiment about migrants;

The Sankey diagram is finally produced by selecting the users that appear in both temporal snapshot and plotting their shift.

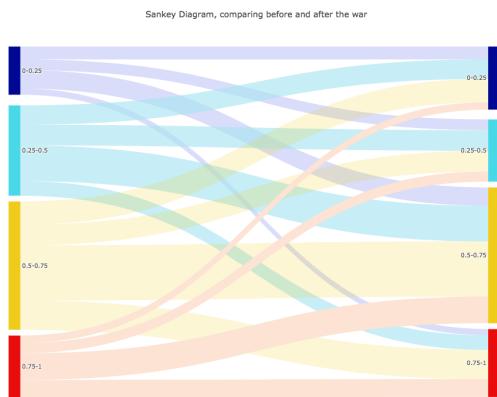
The result shown in Figure 17 highlights the natural behaviour of a user to preferentially stay in the same bin or to move in an adjacent bin, i.e. to have a moderate change of opinion instead of a drastic one.

The other effect one notices is that the bins "0-0.25" and "0.75-1" are bigger in the snapshot after the conflict. This radicalization might be seen as the consequence of the arrival of Ukrainian refugees, or could be an effect due to the proximity of Italian election. Figure 2 shows how a great number of tweets were written after May 2022, with a peak in July 2022, which indicates the start of the electoral campaign



(a) Tweets containing the substring "ucr" (b) Tweets not containing the substring "ucr"

Figure 16: Distribution of the user label in the two cases.



**Figure 17: Sankey Diagram of user label in the two temporal snapshot, before and after the 2022 Russo-Ukrainian conflict**

for the Italian election of September 2022. In this temporal frame one finds therefore a great number of tweets written by politicians with a well defined opinion on immigration, and as a cascade effect the user might have taken a defined position as well.

In addition to the flux, one could investigate the increase or decrease percentage of every bin count in the two temporal snapshot, in order to have a quantitative measure for the opinion change.

Since the goal is to identify the change of opinion about migrants and refugees, we feared that these numbers might be biased from the presence of the tweets regarding Ukraine, which are positive biased as shown in Figure 16a. For this reason it was decided to repeat the analysis on the tweets not containing the substring "ucr".

It was also considered to repeat this analysis on two temporal snapshots before the war. The idea behind this reasoning is the following: if one compares two temporal snapshot which are not divided by an event so strong to make the people change their mind, one could get a very rough estimation of a natural statistical fluctuation of the counts of the bins. This measure is useful to compare the change in our Before-After study with some statistical fluctuations. As temporal frames for the last described comparison were chosen the first two snapshot created in the dynamic community discovery task.

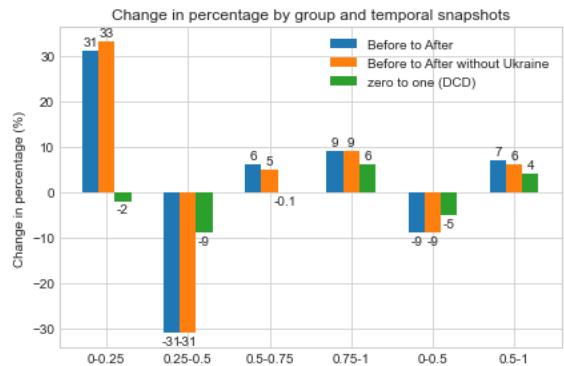
The results of this analysis is presented in Table 5. For a better visualization, a bar chart of these data is reported in Figure 18. Besides the change of each bin, it could be interesting to look at the sum of the classes in the bins "0-0.25", "0.25-0.5" and "0.5-0.75", "0.75-1", which are now bins indicating the negative and positive sentiment. As one can

**Table 5: Percentual change of bin counts.**

BtA indicates the change of user label from before to after the 2022 Russo-Ukrainian Conflict; BtA/Ukr idem but having removed the tweets containing the substring "ucr"; 1to2 indicates the change between the first and second temporal snapshot created in the dynamic community discovery task.

	0-0.5	0.5-1	0-0.25	0.25-0.5	0.5-0.75	0.75-1
BtA	-9%	+7%	+31%	-31%	+6%	+9%
BtA/Ukr	-9%	+6%	+33%	-31%	+5%	+9%
1to2	-5%	+4%	-2%	-9%	-0.1%	+6%

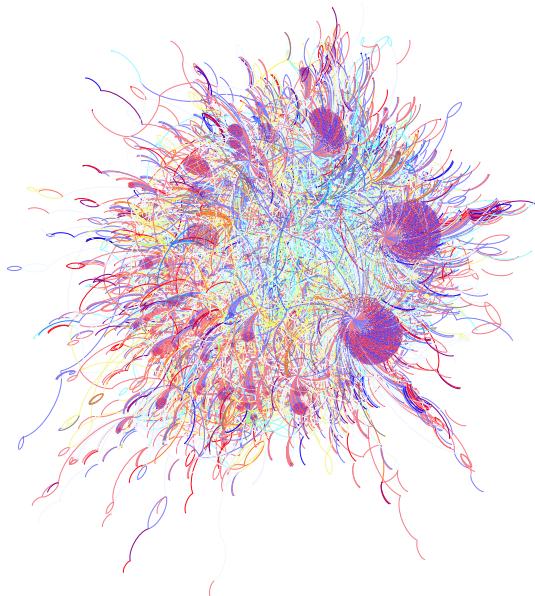
see, the rows BtA and BtA/Ukr are similar and there is no significantly change, although it was noticed the small effect of the slightly smaller increase of the positive 0.5-1 class, with the consequent increase of the negative sentiment class, as predicted. This effect is small and therefore negligible. Comparing the changes in 1to2, these are significantly smaller than the one in the row BtA. This may be a hint that the Russo-Ukrainian conflict (and also the Italian election) could have had an effect on the Italians' opinion about migrants.



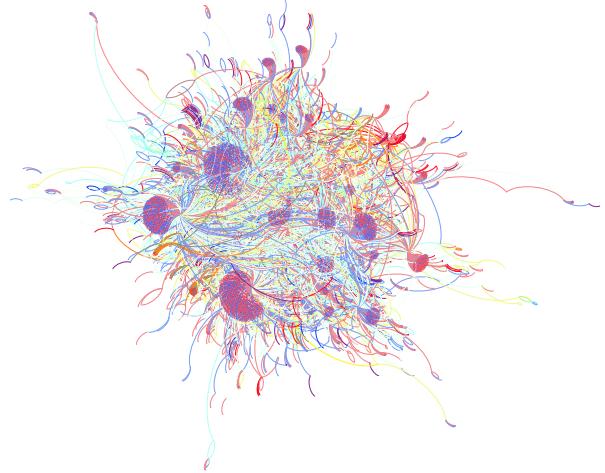
**Figure 18: Percentage change of bin counts per temporal snapshots succession**

The overview results of the graph can be visualized with the tool *Gephi*. One can visualize the distribution of the user label, which is represented by a colorization coherent with the Sankey plot of Figure 17. In particular, in Figure 19 there is a dominance of cold colors indicating a majority of negative sentiment, while in Figure 20 the graph is dominated by warm colors, indicating a positive sentiment. There is furthermore another effect identifiable by the graph visualization, which is the color distribution inside of hubs. These aggregation of nodes represent the conversation in which take part a great amount of users. As one can notice, the hubs are mainly constituted by blue or red users, which represent the most extreme categories (respectively 0-0.25 and 0.75-1). This effect indicates that in the most discussed topics the

most polarized users take the word, leaving less space to the people with a more undefined opinion.



**Figure 19: Graph before the Russo-Ukrainian conflict**



**Figure 20: Graph after the Russo-Ukrainian conflict**

## 7 DISCUSSION

In this project an analysis about Italians' opinion regarding refugees and migrants in general has been carried out. To do so, through the python library *snscraper*, data about tweets and users has been collected starting from September 2020 until September 2022. This specific temporal window has been chosen because the main goal of the analysis was to capture a change in opinions following the Russian-Ukrainian conflict. The data has been manipulated by tools and algorithms offered by *Network Science* and *Data Mining* knowledge in general. As a preliminary step, in order to obtain an indicator about the opinion of the user, it has been created a new attribute called *User\_Label* based on the mean-sentiment of tweets wrote by each User. Then, the topology of the obtained network has been explored in order to catch some useful information about which nodes were hubs or generally more influential in the discussion comparing different centrality measures.

As a first task, in Section 4, *Community Discovery* has been carried out, both in a static and dynamic way. In particular,

the latter allows to reach the most relevant considerations, due to the temporal dimension. In fact, time was a key variable for whole analysis. The contribution of this task was crucial during the conduct of the open question, allowing a more complete interpretation of the results.

Before moving to the last task, different *Opinion Dynamic* models have been run on the network (both in the real world and in synthetic ones) with the aim of observing the way opinions change over time, from positive to negative and vice-versa. Comparing the *voter model* on the real-world network with the Barabasi-Albert one, the greatest similarity in terms of spreading was found.

All the results obtained from the previous tasks were merged, in order to try to give an answer to the main paper's topic. Even though from the analysis of tweets containing the substring "ucr" emerges a prevalence of pro-refugees opinion, it is not exactly the same by taking the whole network. Moreover, the first two columns in Table 5 highlights a mild shift from a negative opinion to a positive one, when classifying an opinion between 0 and 0.5 as 0 and an opinion

between 0.5 and 1 as 1. However, looking at the more detailed opinions in four bins, it can be noticed that the stronger effect regards the increasing of the 0-0.25 bin (+31%). These effects are also visible in the Sankey diagram depicted in Figure 17, which also shows that the major contribution to this bin is coming from the 0.5-0.75 one, i.e. a non-negligible amount of people who had a slightly positive sentiment before the war found themselves having a strong negative feeling about migrants after the war. In conclusion, since the conflict is a fairly recent event, the data collected is not sufficient to derive statistically relevant and reliable results. We suggest, therefore, that the analysis be repeated when more data are available to work with.

## REFERENCES

- [1] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74 (Jan 2002), 47–97. Issue 1. <https://doi.org/10.1103/RevModPhys.74.47>
- [2] Mathieu Bastian, Sébastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. <http://wwwaaai.org/ocs/index.php/ICWSM/09/paper/view/154>
- [3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008. <https://doi.org/10.1088/1742-5468/2008/10/p10008>
- [4] 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>
- [5] P. Erdős and A. Rényi. 1959. On Random Graphs I. *Publicationes Mathematicae Debrecen* 6 (1959), 290.
- [6] Mariaconsiglia Flavia Fedele. 2018. portichiusi e cuoriaperti: il nuovo hashtag di Matteo Salvini. *Mar de Sargassi* (2018). <https://www.mardeisargassi.it/portichiusi-e-cuoriaperti-il-nuovo-motto-di-matteo-salvini/>
- [7] Rossetti G., Milli L., and Cazabet R. 2019. CDLIB: a python library to extract, compare and evaluate communities from complex networks. *Applied network science* 4 (2019). <https://doi.org/10.1007/s41109-019-0165-9>
- [8] S. Galam. 2002. Minority opinion spreading in random geometry. *The European Physical Journal B* 25, 4 (feb 2002), 403–406. <https://doi.org/10.1140/epjb/e20020045>
- [9] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 7821–7826. <https://doi.org/10.1073/pnas.122653799> arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.122653799>
- [10] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11 – 15.
- [11] Richard A. Holley and Thomas M. Liggett. 1975. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. *The Annals of Probability* 3, 4 (1975), 643 – 663. <https://doi.org/10.1214/aop/1176996306>
- [12] Plotly Technologies Inc. 2015. *Collaborative data science*. Montreal, QC. <https://plotly.com>
- [13] IPSOS. 2019. Ciak MigrAction: indagine sulla percezione del fenomeno migratorio in Italia. *We-World Onlus* (2019). <https://www.ipsos.com/it-it/ciak-migration-indagine-sulla-percezione-del-fenomeno-migratorio-italia>
- [14] Cecilia Emma Sottilotta Julian M. Campisi. 2022. Italy's Far-Right and the Migration Debate: Implications for Europe. *IAI Commentaries* 22, 41 (2022), 6.
- [15] JustAnotherArchivist. 2018. snscreper. <https://github.com/JustAnotherArchivist/snscreper>.
- [16] A B W KENNEDY and H R SANKEY. 1898. THE THERMAL EFFICIENCY OF STEAM ENGINES. REPORT OF THE COMMITTEE APPOINTED TO THE COUNCIL UPON THE SUBJECT OF THE DEFINITION OF A STANDARD OR STANDARDS OF THERMAL EFFICIENCY FOR STEAM ENGINES: WITH AN INTRODUCTORY NOTE. (INCLUDING APPENDICES AND PLATE AT BACK OF VOLUME). *Minutes of the Proceedings of the Institution of Civil Engineers* 134, 1898 (1898), 278–312. <https://doi.org/10.1680/imotp.1898.19100> arXiv:<https://doi.org/10.1680/imotp.1898.19100>
- [17] Andrea Lancichinetti, Santo Fortunato, and Já nos Kertész. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* 11, 3 (mar 2009), 033015. <https://doi.org/10.1088/1367-2630/11/3/033015>
- [18] Dylan Patrick McGinnis. 2021. Anti-Immigrant Populism in Italy: An Analysis of Matteo Salvini's Strategy to Push Italy's Immigration Policy to the Far Right. *The Yale Review of International Studies* (2021). <http://yris.yira.org/winter-issue/4659>
- [19] Giorgia Meloni. 2022. <https://twitter.com/giorgiameloni/status/1552212411662336001>
- [20] Lega Salvini Premier. 2022. <https://twitter.com/LegaSalvini/status/1569380071877627905>
- [21] Usha Nandini Raghavan, Ré ka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76, 3 (sep 2007). <https://doi.org/10.1103/physreve.76.036106>
- [22] Giulio Rossetti. 2019. Exorcising the Demon: Angel, Efficient Node-Centric Community Discovery. In *International Conference on Complex Networks and Their Applications*. Springer, Springer.
- [23] Giulio Rossetti, Letizia Milli, Salvatore Rinzivillo, Alina Sirbu, Fosca Giannotti, and Dino Pedreschi. 2018. NDlib: a Python Library to Model and Analyze Diffusion Processes Over Complex Networks. *CoRR* abs/1801.05854 (2018). arXiv:1801.05854 <http://arxiv.org/abs/1801.05854>
- [24] Giulio Rossetti, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2017. Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning* 106 (08 2017). <https://doi.org/10.1007/s10994-016-5582-8>
- [25] Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo. 2016. A novel approach to evaluate community detection algorithms on ground truth.
- [26] KATARZYNA SZNAJD-WERON and JÓ ZEF SZNAJD. 2000. OPINION EVOLUTION IN CLOSED COMMUNITY. *International Journal of Modern Physics C* 11, 06 (sep 2000), 1157–1165. <https://doi.org/10.1142/s0129183100000936>
- [27] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2018. From Louvain to Leiden: guaranteeing well-connected communities. *CoRR* abs/1810.08473 (2018). arXiv:1810.08473 <http://arxiv.org/abs/1810.08473>
- [28] Matteo Villa. 2021. Fact-checking: migrazioni 2021. *ISPI* (2021).
- [29] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 6684 (June 1998), 440–442. <https://doi.org/10.1038/30918>