COMMUNICATIONS PHYSICS

ARTICLE

https://doi.org/10.1038/s42005-020-0340-4 OPEN

The role of bot squads in the political propaganda on Twitter

Guido Caldarelli [1,2,3,4], Rocco De Nicola [1,5], Fabio Del Vigna [1], Marinella Petrocchi [1,6✉] & Fabio Saracco [1]

Nowadays, Social Media are a privileged channel for news spreading, information exchange, and fact checking. Unexpectedly for many users, automated accounts, known as social bots, contribute more and more to this process of information diffusion. Using Twitter as a benchmark, we consider the traffic exchanged, over one month of observation, on the migration flux from Northern Africa to Italy. We measure the significant traffic of tweets only, by implementing an entropy-based null model that discounts the activity of users and the virality of tweets. Results show that social bots play a central role in the exchange of significant content. Indeed, not only the strongest hubs have a number of bots among their followers higher than expected, but furthermore a group of them, that can be assigned to the same political tendency, share a common set of bots as followers. The retweeting activity of such automated accounts amplifies the hubs' messages.

[1] IMT Scuola Alti Studi Lucca, Piazza S. Francesco 19, 55100 Lucca, Italy. [2] European Centre for Living Technology, Università di Venezia Ca' Foscari, S. Marco 2940, 30124 Venice, Italy. [3] Catchy srl, Talent Garden Poste Italiane Via Giuseppe Andreoli 9, 00195 Rome, Italy. [4] Istituto dei Sistemi Complessi CNR, Dip. Fisica, Università Sapienza, P.le Aldo Moro 2, 00185 Roma, Italy. [5] CINI—Cybersecurity National Laboratory, Rome, Italy. [6] Istituto di Informatica e Telematica, CNR, Pisa, Italy. ✉email: marinella.petrocchi@iit.cnr.it

Global Digital Report, in 2018, said that 'more than 3 billion people around the world now use social media each month' (https://digitalreport.wearesocial.com). Even traditional newspapers and news agencies moved to social networks, to cope with this societal change.

Since a decade microblogging platforms, like Twitter, have become prominent sources of information[1], catching breaking news and anticipating more traditional media like radio and television[2]. Helped by the simple activity consisting of creating a text of 140 (now 280) characters, on Twitter we assist to the proliferation of social accounts governed—completely or in part—by pieces of software that automatically create, share, and like contents on the platform. Such software, also known as social bots—or simply bots—can be programmed to automatically post information about news of any kind and even to provide help during emergencies. As amplifiers of messages, bots can simply be considered as a mere technological instrument. Unfortunately, the online ecosystem is constantly threatened by malicious automated accounts, recently deemed responsible for tampering with online discussions about major political elections in western countries, including the 2016 US presidential elections, and the UK Brexit referendum[3–6]. Worryingly, propensity to fall into disinformation has been recently measured in US by Ruths in ref. [7] and recent work by Shao et al. demonstrates that automated accounts are particularly efficient in spreading low credibility content and amplifying their visibility[8]. Also, Stella et al. show how such accounts target influential people, bombarding them with hateful contents[9] and influencing the public opinion during noteworthy events[10], while Grinberg et al. observe how they even interact with genuine users according to their political affiliations[11]. In 2017, a study by the Indiana University[12] highlighted that, on average, social bots account for 9% to 15% of total active Twitter users. This notable percentage maintains over time since bots evolve: in a large-scale experiment, Cresci at al. in ref. [13] proved that neither Twitter admins, nor tech-savvy social media users, nor cutting-edge applications were able to tell apart evolving bots and legitimate users.

Academicians make their best efforts to fight the never ending plague of malicious bots populating social networks. The literature offers a plethora of successful approaches, based, e.g., on profile-[14,15], network-[16–18] and posting-characteristics[19–21] of the accounts. In particular, the supervised approach proposed by Cresci et al. in ref. [14] tested a series of classification rules proposed by bloggers, and features sets by Academia, on a reference dataset of genuine and fake accounts, leading to the implementation of a classifier, which significantly reduces the cost for data gathering.

Actually, the studies regarding detection of automated accounts rarely analyse their effective contribution in the social networks panorama. Indeed, while messages exchanged on social platforms contain a great amount of data, just a fraction of them carries crucial information for the description of the system, while the rest contributes to random noise. Thus, detecting the relevant (i.e., those not compatible with users' random activity) communication and interaction patterns is of utmost importance in order to understand which accounts, including bots, contribute to the effective dissemination of messages. In this sense, it is necessary to compare the properties of the real network with a proper null model.

Entropy-based null-models are a natural choice, since they are general and, being based on Shannon entropy, unbiased by construction. In a nutshell, starting from the real network, their definition relies on three steps: (1) the definition of an ensemble of graphs; (2) the definition of the entropy for this ensemble and its maximization up to some (local or global) constraints[22]; (3) the maximization of the likelihood of the real network[23,24]. Entropy-based null-models have been successfully used in the last years for the analysis of complex networks[25,26]. The fields of application are the most varied, from reconstructing a network from partial information[27], to detecting early signals of structural changes[28,29], to assessing the systemic risk of a financial system[30,31]. Recently, this approach has been applied by Becatti et al.[32] to the Twitter traffic during the 2018 Italian election campaign. The study was able to infer political standings directly from data. Moreover, the analysis of the exchanged messages showed a signal of communication between opposite political forces during the election campaign, which anticipated an unexpected post-elections political agreement.

In the present paper, we merge the application of the lightweight classifier for bot detection proposed by Cresci et al. in ref. [14] with the analysis of complex networks via entropy-based null-models. Once we have cleaned the system from the random noise via the application of the null-model, we study the effects of social bots in retweeting a significant amount of messages on Twitter. The analysis is applied to a tweet corpus about migration in the Mediterranean Sea from North Africa to Italy.

This study has two main results: firstly, after cleaning the system from the random activity of users, we detect the main hubs of the network, i.e., the most effective accounts in significantly propagating their messages. We observe that those accounts have a number of bots among their followers (in the cleaned network) higher than average. Secondly, the strongest hubs in the network share a relatively high number of bots as followers, which most probably aim at further increasing the visibility of the hubs' messages via following and retweeting. Hereafter, we will refer to groups of bots that follow and retweet the same group of hubs with the term bot squads. To the best of our knowledge, the existence of formations of bots shared by a group of human-operated accounts has never been reported in the literature before.

Results

User polarization. On Twitter, users are strongly clustered in communities sharing similar ideas (evidences of this, and discussions of its implications, can be found in many papers, see, e.g., refs. [33–41]). Our assumption is that, if two users interact with the same followers and followees, they probably share similar viewpoints, including those regarding politics. We thus build clusters of politically homogeneous groups by starting from those accounts for which we have the greatest information available. We exploit the fact that Twitter offers the possibility (upon request of the account owner) to obtain an official certification of account's authenticity. The procedure is mostly adopted by VIPs, official political parties, newspapers, radios and TV channels, to reduce interferences of fake users. Accounts that pass the procedure are tagged as verified and on the official portal have a blue circle, with a white tick at the center, close to their name. Verified users have been proved to be a solid starting point for accurate analyses. In fact, not only they lead to valuable information about the number of bots that follow them (see work by Varol et al. in ref. [42]), but also, following the communication patterns of a set of verified accounts, it is possible to get a very large set of trusted, i.e., not bots, accounts[43].

To infer the political orientation of a user from the available data, we focus on the bipartite network of verified (on one layer) and unverified (on the other layer) accounts, as in Fig. 1. A link between two users belonging to the different sets is present if one of them retweeted at least once the other user. In our representation, the network is undirected: we do not consider who retweeted who, but only the mere presence of at least one retweet.

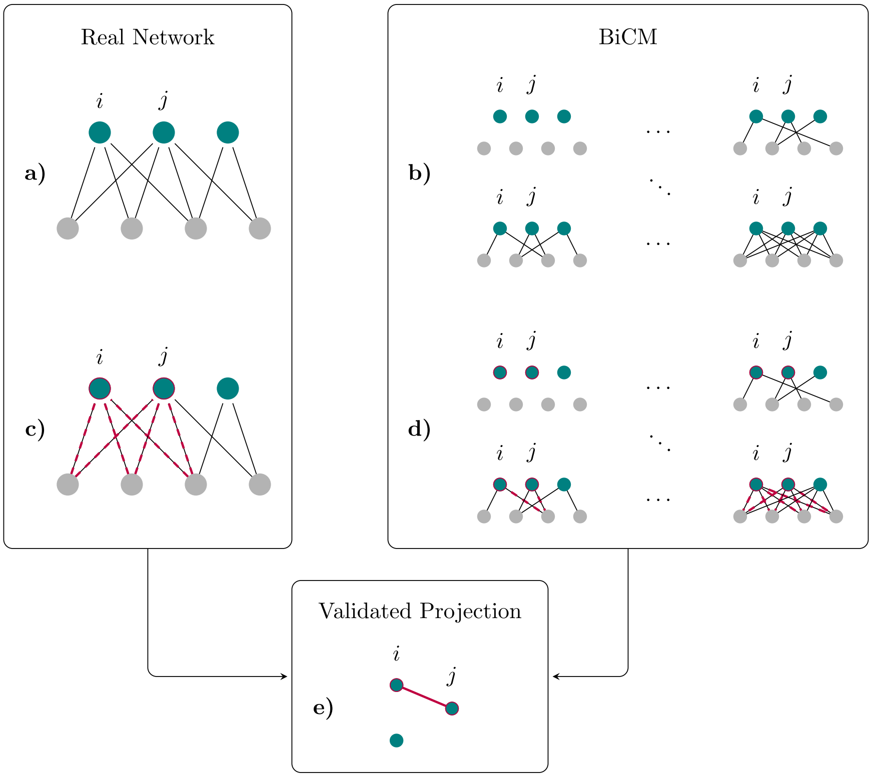

Fig. 1 Schematic representation of the projection procedure for bipartite undirected networks. a an example of a real bipartite network. For the actual application, the two layers represent verified (turquoise) and unverified (grey) users and a link between nodes of different layers is present if one of the two users retweeted the other one, at least one time. **b** the Bipartite Configuration Model (BiCM) ensemble is defined. The ensemble includes all possible link realisations, once the number of nodes per layers has been fixed. **c** we focus our attention on nodes *i* and *j*, i.e., two verified users, and count the number of common neighbours (in magenta both the nodes and the links to their common neighbours). Subsequently, (**d**) we compare this measure on the real network with the one on the ensemble: If this overlap is statistically significant with respect to the BiCM, (**e**) we have a link connecting the two verified users in the projected network.

It is worth noting that other types of interactions, such as replies, mentions and quoted tweets are present on Twitter. We are considering retweeting activity only, since it represents the preferred way through which users spread messages they agree with[34]. If we had used, for instance, replies, the network would have been much harder to interpret; replies can be used either to support the ideas of the original tweet, or to express disagreement towards them. A similar reasoning holds for the other possible interactions. By inserting mentions in a tweet, a user either invites the mentioned accounts to participate to the discussion, or points out that the tweet somehow affects them; however, it can be either in provocative or in constructive ways. Analogously, in the case of quoted tweets, the sender may intend to comment a tweet, but it might be either to support or to deplore it. Thus, since the intention of the mentions, quotes and replies can be of different kinds, we focus on retweetting, the type of interaction which is not amenable to multiple interpretations.

We project the bipartite network of verified and unverified users on the layer of the former. To do that, we consider the statistically significant amounts of interactions shared by pairs of verified accounts. The steps of such projection are sketched in Fig. 1. For every couple of verified users, we count the number of common unverified ones interacting with them and compare this number with its expected probability distribution according to an entropy-based null-model constraining the degree sequence of both layers of the bipartite network. If the *p*-value of the observation on the real network is statistically significant, we project a link on the layer of verified users. In this way, we can focus on that statistical significant group of verified users that is similarly perceived by the unverified users. The technical details about this validated projection can be found in the “Methods” section.

The presence of a strong community structure in the bipartite network composed of verified and unverified layers of users has already been observed by Adamic at al. in ref. [33]. Here, we repeat the analysis of Becatti et al.[32] and check the results on the layer of verified users, for which we have reliable information. Also in this case we find a strong community structure.

We would like to remark that we assign the terms ‘validated’ and ‘verified’ two different meanings: the former indicates a node that passes the filter of the projection while the latter refers to accounts that pass an authenticity check by Twitter.

There are other approaches to inferring the political orientation of users from data. Del Vicario et al.[37], for instance, consider several groups of Facebook pages and divide users depending on the frequency they interact with each group of pages. Instead, to decide whether a user is a Democrat or a Republican, Conover et al. in refs. [34–36] use network properties of retweets and a machine learning algorithm to determine the topic of a tweet. However, in both approaches, the communities are somehow decided a priori, either by data selection or by defining the groups. On the contrary, with our method, communities naturally arise from data.

Political orientation: in the following, we will often refer to Italian parties and representatives of the Italian government in office during the period in which the data were collected

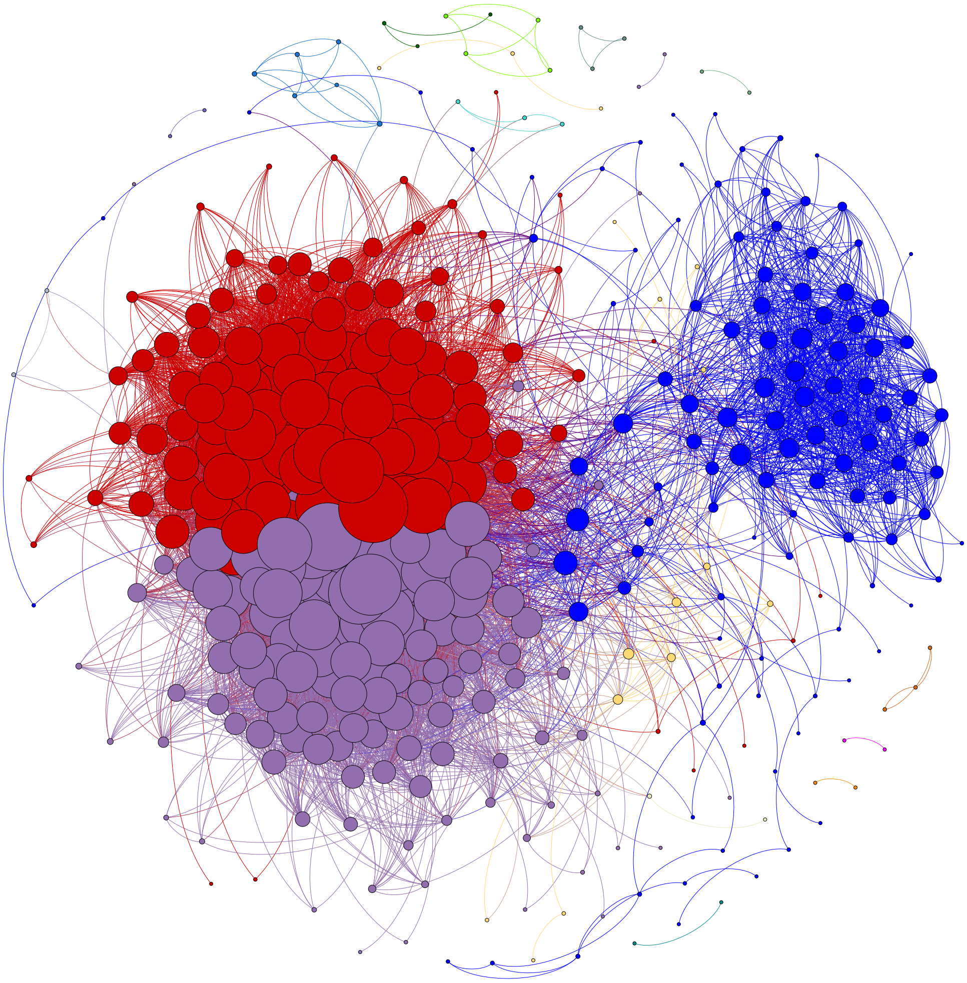

Fig. 2 The network resulting from the projection procedure. The network presents a strong community structure. The accounts tied to the Italian government (Lega and Movimento 5 Stelle) and other right wing parties are in Mediterranean blue. The accounts of the center-left wing parties (e.g., the Italian Democratic Party, PD) are in tomato red. The eggplant purple group includes media official accounts, several NGOs and left-wing politicians. Some official accounts related to the Catholic Church are in orange. In turquoise we represent some smaller groups involved in the debate, such as the Maltese Prime Minister Joseph Muscat and some of his ministers, and in basil green we represent a soccer commentators community.

(23 January– 22 February 2019). For the sake of clarity, Supplementary Note 1 briefly explains the high-level characteristics of such parties.

The bipartite network describing the retweets between verified and unverified users involves nearly one half of the unverified users in our dataset. Nevertheless, the network obtained by following the projection procedure described in the "Methods" section shows a strong community structure, see Fig. 2.

To quantify the presence of clusters, we use the Louvain algorithm[44], one of the most effective community detection algorithms. To avoid problems related to node ordering[45], on a network with N nodes, we apply the algorithm for N times, after reshuffling the order of nodes. Among the N partitions resulting by the application of the algorithm, we then select the configuration with the largest modularity[45]. Such configuration, reported in Fig. 2, displays three main communities: one tied to the current government (right-wing parties and Movimento 5 Stelle, the Mediterranean blue community), one tied to the Italian Democratic Party (PD), its representatives, and some representatives of smaller parties on the left of PD (the tomato red community), and one tied to several NGOs, politicians on the left of PD, and different online and offline news-papers (eggplant purple community). Smaller communities, including one with Joseph Muscat, Malta's Prime Minister, and part of his ministers, have been involved in the discussion for the aid of migrants and castaways. The composition of the communities in Fig. 2 is detailed in the Supplementary Note 1.

If we compare the emerged communities with those observed by Becatti et al.[32] during the 2018 electoral campaign, we notice several differences. The outcome of[32] was that the Movimento 5 stelle (hereafter M5S) group and their supporters were clearly distinguishable from the right wings, while, on issues concerned with the Mediterranean migration, they are not. Instead, on the same issues, the left wing representatives, inside and outside the Democratic Party, are much closer than during the electoral campaign.

Polarization of unverified accounts: verified accounts of politicians can be easily associated to a political party; then membership of unverified users can be guessed by considering their interactions with the communities of verified ones. To do this, we use the polarization index ρ_i defined by Bessi et al. in refs. [46,47]:

$$\rho_i = \frac{\max_{c \in \mathcal{C}} k_i^c}{k_i}, \tag{1}$$

where k_i is the degree of node i, k_i^c is the number of links towards the community c and $\mathcal{C}$ is the set of communities. The

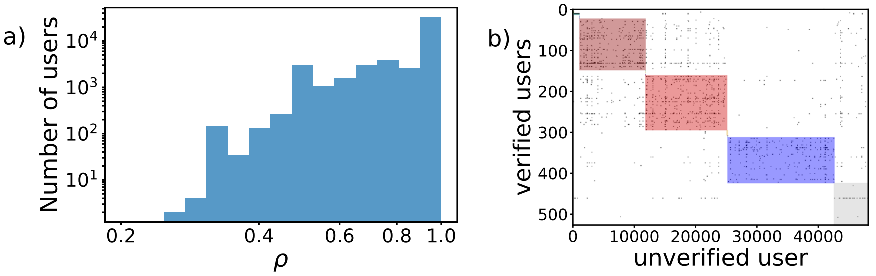

Fig. 3 Interactions between verified and unverified users and polarization index distribution among unverified users. a polarization index distribution among unverified users. The plot is in log-log scale, i.e., nodes with a polarization higher than 0.9 are more than 10 times the one of those with polarization between 0.8 and 0.9. **b** biadjacency matrix, describing the interactions between verified and unverified users. Nodes have been rearranged in order to highlight the community structure and colored according to their communities, i.e. eggplant purple for NGO's, media accounts and left wing politicians, tomato red for the Democratic Party community, orange for Catholic Church related accounts and Mediterranean blue for the pro-government users. In gray, users with lower values of polarization or not projected by the validated projection (see "Methods" and Supplementary Notes 2 and 3).

distribution of ρ_i is extremely peaked on values close to 1, see Fig. 3a. Given such a strong polarization, we can safely assign unverified users the polarization of the community they mostly interact with.

We also find a small amount (with respect to the size of verified–unverified network) of unverified nodes whose polarization is not strong enough to be uniquely assigned to a specific cluster: they are part of the grey group in Fig. 3b.

As noted at the beginning of the previous paragraph, this polarization procedure does not consider almost one half of the unverified users, since they do not interact, on the whole observation period, with a single verified user. This may be due to several reasons. Differently from Becatti et al.[32], where a corpus of tweets exchanged during the election campaign was analyzed, here we focus on a set of tweets concerned with a specific topic of the political propaganda. We conjecture that, in the former case, the amount of unverified accounts interacting with the verified ones was much higher because it was of interest of the verified accounts (mostly, candidates in the elections) to involve 'standard' users.

In order to know more about unverified users not directly interacting with verified ones, we use what we call a contagion of polarization, namely a label propagation procedure in which the labels are those assigned in the previous steps. Even if unverified users do not retweet verified users, they may retweet other unverified ones and reveal their political orientation through those. As seeds for the propagation of the label, we use the tags obtained in the previous step, i.e., the polarization of verified and unverified users of the previously defined bipartite network. If the majority of the unverified accounts retweeted by the account we want to tag is polarized towards a certain group, we assign that account to the same group. The procedure is iterated and terminates as soon as it is not possible to assign a polarization to any unverified user. We thus assign to unverified users the prevailing polarization of the accounts they interact with and stop when there is no possibility to assign a polarization anymore, i.e., if there is no clear agreement among the neighbours of the considered node. In this way, after 10 rounds of such a procedure, we are able to increase the fraction of the users for which we determine a clear polarization by 27%. Even if this percentage looks small, we will see that the label propagation process is effective for the set of validated unverified accounts considered in the following sections. In this case, the increase of polarized users is almost 58%; additional details are reported in the Supplementary Note 4.

Supplementary Figure 1 illustrates the density of all users (left panel) and bots (right panel) in the three biggest communities, after the polarization by contagion procedure. Interestingly enough, assigning a polarization to bots turns out to be much harder than for genuine users. If we neglect the contribution of the grey bars, we can notice that the relative fractions of the different communities are more or less similar, but for a slight increase of the abundance of 'eggplant purple' bots.

The backbone of the content exchange on Twitter. In the analysis of a complex system, one of the main issues is to skim relevant information from noise. Of course, the definition of noise itself depends on the system. In the previous section, we obtained the political affiliation of verified users by projecting the information in the bipartite network describing the interactions between verified and unverified users. Now, we will apply the entropy-based null model, in its directed variant—Bipartite Directed Configuration Model (BiDCM)—proposed by van Lidth de Jeude et al.[48], to filter the total exchange of content in our dataset, after discounting the information regarding the activity of users and the virality of messages, as in ref. [32]. We sketch the procedure in Fig. 4.

We start by considering the directed bipartite network of users (on one layer) and tweets (on the other layer): an arrow from user u to tweet t indicates that u wrote t. Analogously, the arrow from t to u represents u retweeting t. Thus, the bipartite directed network is intended to describe the retweeting activity of users, considering the information about the virality of the messages. We then construct the BiDCM. In the present case, the constraints describe the node activities, i.e., the number of original tweets posted by every user, the number of retweets of every message and the number of retweets of every account. As for the case of accounts' polarization, we project the information contained in the directed bipartite network on one of the two layers, in order to obtain a monopartite directed network of accounts. The resulting network represents the significant flow of information among users. More in details, for each (ordered) pair of users (u, u') we consider the number of tweets written by u that are retweeted by u'. Subsequently, we assess the statistical significance of this retweeting activity by comparing the real value in the network under investigation with the theoretical distributions of the BiDCM. Otherwise stated, if the number of tweets written by u and retweeted by u' is greater than expected and it is not compatible with the theoretical distribution, a link from u to u' is projected. Thus, by comparing the real system with the null model, we can highlight all the contributions that cannot be explained only by the fixed constraints. Technical details can be found in the "Methods" section and in the Supplementary Notes 5 and 6.

Summing up, the filtering procedure returns a directed network in which the arrows go from the authors to the

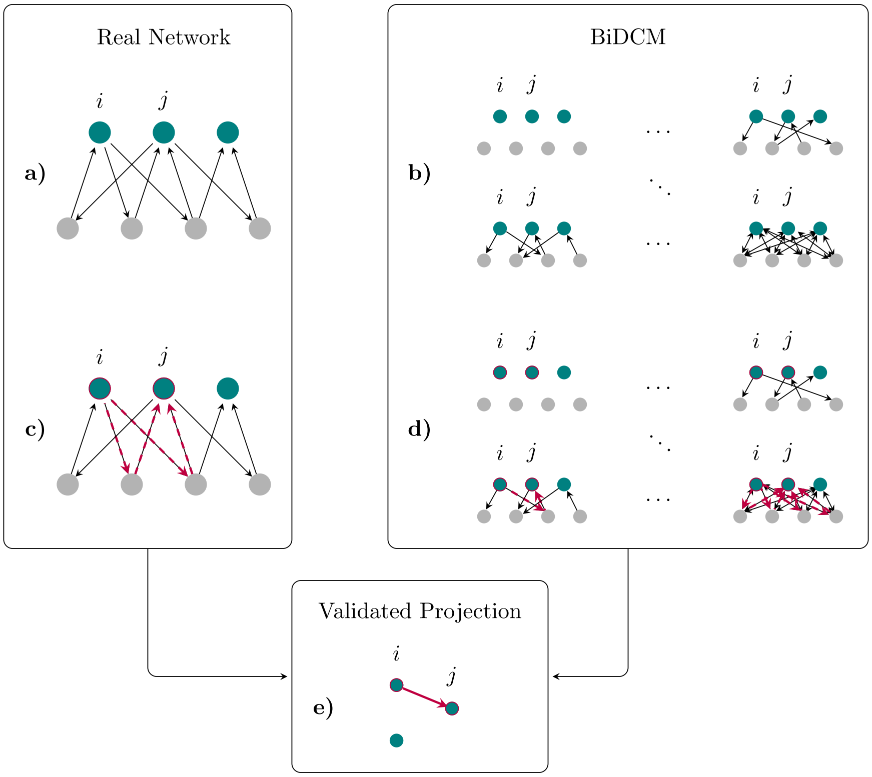

Fig. 4 Schematic representation of the projection procedure for bipartite directed network. **a** an example of a real directed bipartite network. For the actual application, the two layers represent Twitter accounts (turquoise) and posts (grey). A link from a turquoise node to a gray one represents that the post has been written by the user; a link in the opposite direction represents a retweet by the considered account. **b** the Bipartite Directed Configuration Model (BiDCM) ensemble is defined. The ensemble includes all the link realisations, once the number of nodes per layer has been fixed. **c** we focus our attention on nodes *i* and *j* and count the number of directed common neighbours (in magenta both the nodes and the links to their common neighbours), i.e., the number of posts written by *i* and retweeted by *j*. Subsequently, **d** we compare this measure on the real network with the one on the ensemble: if this overlap is statistically significant with respect to the BiDCM, (**e**) we have a link from *i* to *j* in the projected network.

retweeters and it reduces the number of nodes to 14,883 users and of links to 34,302. The connectance, i.e. the density of links, of the network is $\rho \simeq 3\times10^{-5}$. This network is hereafter referred as directed validated network, or simply, validated network. Figure 5 shows the structure of the validated network in terms of communities.

On top of this, we have analyzed the presence of automated accounts, by using the bot detection method described in the “Methods” Section. The incidence of bots in the validated network is about 2.5%, against almost the 7% of nodes in the original network. The number of loops, i.e., users that retweet (significantly with respect to their activity) their own tweets, is around 1.2% of the total amount of links of the validated network, thus relatively high. This effect reverberates also on the number of validated nodes, that significantly retweet themselves (slightly <3%). For the subsequent analyses, we discard the contribution of loops, since we are interested in analysing the source of the shared contents on Twitter.

Hubs and bots: as mentioned in the previous section, the validated links go from the authors to the retweeters. The effectiveness of an author can be derived by its ability to reach a high number of relevant nodes: this principle is finely implemented in the Hubs-Authorities algorithm, originally introduced by Kleinberg in ref. [49] to rate web pages. In the original version, the paradigm assigns two scores for each web page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. In the scenario currently under investigation, hubs and authorities are Twitter accounts. The authorities are sort of sink of the content exchange. In the following, we will focus on hubs, because they represent the driving force of the discussion and are relatively popular users, and even if they are not verified by Twitter, we often have reliable information about their accounts.

Table 1 shows the values for the top 20 nodes, in term of hub scores. The first account is the one of Mr. Matteo Salvini. The second and the third ones refer to two journalists of a news website supported by Casa Pound, a neo-fascist Italian party. The fourth is the account of Ms. Giorgia Meloni, former ally, during the 2018 Italian electoral campaign, of Lega, the party of Mr. Salvini. We remind the reader that Mr. Salvini and Ms. Meloni are the leaders of the two major conservative Italian parties. Their account names have not been anonymed in Table 1 since they are verified accounts. The two leaders have similar opinions on how to deal with the Mediterranean migration. The fifth and sixth accounts are, respectively, a journalist of ‘Il Fatto Quotidiano’ (a newspaper close to M5S) and an unverified user with opinions in line with the ones of the two above mentioned politicians. Notably, all the accounts in Table 1 belong to the blue community. The first account with a different membership (‘TgLa7’, a popular newscast by a private TV channel, whose account belongs to the purple community) ranks 176th in the hub score ranking.

The case of the Italian chapter of a NGO assisting migrants in the Mediterranean Sea is worthy of note: while it has the fifth highest value of out-degree ($k_{\rm out}=1104$), it has an extra low hub score (4×10^{-4}), ranking 452nd. This is impressive since, in several occasions, the Italian government (in the figure of Salvini,

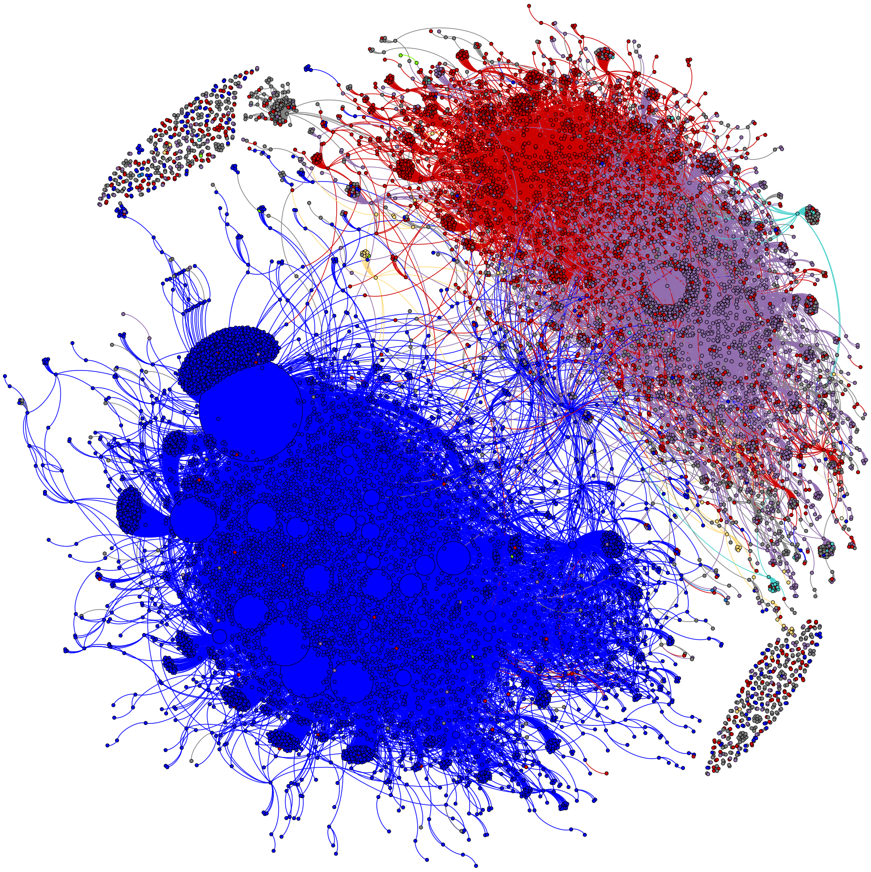

Fig. 5 The directed validated projection of the retweet activity network. Nodes are eggplant purple for NGO's, media accounts and left wing politicians, tomato red for the Democratic Party community and Mediterranean blue for the pro-government users; other colours identify smaller communities. An arrow between a source node and a target node is present if the target is a significant retweeter of the source. The dimension of each node is proportional to its hub score: the biggest node (in blue) is the account of Matteo Salvini, i.e. the leader of a major wing party and the Minister of Internal Affairs at the time of the data collection.

the Minister of Internal Affair) and the NGO have been opponents on the issue of disembarking the migrants rescued during the NGO activities.

Remarkably, we observe a non zero overlap among the bots in the list of the validated followers of human-operated users. To the best of our knowledge, this is the first time that such a phenomenon is detected. In our opinion, the use of bot squads, retweeting the messages of two or more strong hubs, aims at increasing the visibility of their tweets. We have detected two main groups of such accounts, the other being composed by a maximum of two common bots. The first one includes 22 genuine accounts (9 of which are in the top 10 hubs), sharing 22 bots. In this set, some users share a relatively high fraction of bots; there is one right wing account that shares all its automated followers with both Meloni and Salvini, see Fig. 6. In Figs. 7 and 8, we represent two subgraphs of the validated network in Fig. 5. The subgraph in Fig. 7 shows the first group of genuine accounts sharing bots and all their bot followers. Such accounts belong almost exclusively to the blue community. The hub scores, represented as the dimensions of the nodes, are nearly homogeneous among the hubs. This does not happen in the subgraph referring to the second group (see Fig. 8): beside the presence of a strong hub, the hub score distribution is much skewer than for the previous group. Moreover, in absolute terms, the hub scores are much smaller than in the previous case, since the strongest hub is the aforementioned account of 'TgLa7' newscast. The accounts in the subgraph of Fig. 8 belong almost exclusively to the purple community.

Figure 9a shows that the main activity of the bots in the first bot squad is retweeting. As expected, they mostly retweets human-operated accounts connected to them (Fig. 9b). The same cannot be said for mentions that may be used either to provoke or to involve the target. Accounts from different political side are mentioned by bot squads; in fact, the bot accounts with more than 30 mentions point to members of the blue community as well as the official account of the Democratic Party ('pdnetwork'). It is worth noting that other 'not-sided' verified accounts, as the one of the President of the Republic ('Quirinale') and the one of the President of the Chamber of the Deputies ('Roberto_Fico') are mentioned there and that, in most cases, the messages containing those mentions are sort of invites, for the institutional figures, to manage immigrants and migration (Fig. 9c).

The most striking outcome of our content analysis, however, concerns the sources cited by the bots in the blue squads: 89% of their original tweets (i.e, not replies, nor retweets or quoted tweets), contains a URL and 97% of those URLs refers to www.voxnews.info, a website blacklisted as a source of political disinformation by two popular fact checking websites, namely www.butac.it and www.bufale.net. Additional details about our study on the bots' squads can be found in the Supplementary Note 7.

Table 1 Screen names of the hubs in the validated network, their hub score, their out-degree k^{out}, the fraction of bots in their out-neighbours (indicated as $\frac{|bot_i|}{k_i^{out}}$ and the ratio between this value and the average over the entire network (indicated as $\frac{|bot_i|}{k_i^{out}} / \frac{|bot|}{N_{validatedusers}}$). For the sake of privacy, the screen names of the unverified accounts have been anonymized. Matteo Salvini and Giorgia Meloni, the leaders of the two major right wing parties, have not being anonymised, since they are verified accounts.

Screen name	Hub_{score}	k_i^{out}	$\frac{|bot_i|}{k_i^{out}}$	$\frac{|bot_i|}{k_i^{out}} / \frac{|bot|}{N_{validatedusers}}$
'matteosalvinimi'	1.000	3473	0.023	1.058
'hub_1'	0.490	1270	0.003	0.141
'hub_2'	0.465	1199	0.004	0.187
'GiorgiaMeloni'	0.427	1303	0.032	1.444
'hub_4'	0.395	1040	0.005	0.215
'hub_5'	0.326	809	0.011	0.498
'hub_6'	0.300	775	0.009	0.404
'hub_7'	0.290	574	0.002	0.078
'hub_8'	0.282	583	0.0	0.0
'hub_9'	0.271	646	0.003	0.139
'hub_10'	0.200	395	0.005	0.227
'hub_11'	0.189	368	0.0	0.0
'hub_12'	0.186	401	0.005	0.224
'hub_13'	0.166	341	0.009	0.394
'hub_14'	0.152	268	0.0	0.0
'hub_15'	0.133	245	0.012	0.549
'hub_16'	0.128	222	0.0	0.0
'hub_17'	0.126	299	0.013	0.600
'hub_18'	0.112	190	0.0	0.0
'hub_19'	0.106	279	0.011	0.482

Discussion

The 2018 Eurobarometer report on news consumption presents a clear increasing trend of popularity of online sources with respect to traditional ones[50]. Albeit this widespread favour, online media are not trusted as their offline counterparts: in a survey conducted in autumn 2017, 59% of respondents said they trusted radio content, while only 20% said they trusted information available on online social networks. Even beside the perception of common users, the presence of fake content has indeed been revealed in several research work, both at level of news per se, as done by, e.g., Quattrociocchi et al. in ref. [39] and Shao et al. in ref. [8], and of fake accounts contributing to spreading them (see, for example, the overview on the rise of social bots by Ferrara et al.[51]).

Twitter is one of the most studied social media, due to the openness of its data through the available public APIs. Also, it is strongly used by professionals for news distribution: a 2017 survey by AGCOM, the Italian guaranteeing agency for communications[52], showed that Italian journalists appear on Twitter much more frequently than common users. Therefore, Twitter has been used for many analyses of communication in the political propaganda, see, e.g., refs. [4,8–10,32,38,53–60]. Obviously, a major issue when performing such kind of analyses is the reliability of the results, which is closely connected to the reliability of the users in the game: in such sense, a rich stream of research is devoted to finding means for detecting automated accounts—even anticipating their future evolution, as done by Cresci et al. via genetic algorithms in ref. [61] —and their interactions with human-operated accounts[9,10,62].

Remarkably, all the previous analyses rarely tackle the effect of random noise, which is indeed of utmost importance when studying complex systems. In[63], Jaynes showed how Statistical Physics could be derived from Information Theory from an entropy maximization principle. Following Jaynes work, in recent years the same approach has been extended to complex networks[22–26], to provide an unbiased benchmark for the analysis, by filtering out random noise. Such a framework proved to be extremely ductile and adaptable to the analysis of different phenomena, in trade networks[24,28,29,64–66], financial networks[27,30,31] or online social networks[32,67]. In the present study, we jointly use a bot detection techniques and an entropy-based null-model for the analysis of the content exchange on Twitter in the Italian discussions on migrant flows from Northern Africa. The analysed corpus has been extremely useful for highlighting the mechanisms used for disseminating information in political debates.

To get the political affiliation of users, we focused on the bipartite network in which the two layers represent verified and unverified users, respectively, and the (undirected) links label the correspond to the retweetting interactions between the two classes. The main idea is to infer the inclination of users towards a political point of view from (a proxy of) their contacts: users which share a big number of followers and followees probably have similar opinions. The bipartite network is then projected on the layer of verified users using an entropy-based null-model, by following the procedure introduced in Saracco et al. in ref. [29].

Verified users have been clustered into three main groups, see Fig. 2: one group includes Italian government representatives, the right wing and the Movimento 5 Stelle party; a second group includes the Italian Democratic party; a third one includes NGOs, online and offline media, journalists and some VIPs (like actors, singers, movie directors). Confirming results presented in other studies[38,39,68,69], the polarization of unverified users is particularly strong: they interact quite exclusively with accounts of a single community, see Fig. 3.

Starting from verified users, using a label propagation algorithm we iteratively assign a group membership to unverified users, by considering the political inclination of the majority of all their followers and followees. This procedure reduces the number of unpolarized accounts of more than 35%. Oddly, the ratio of bot accounts that remain unpolarized after the 'political contagion' is higher than the analogous for all users, see Supplementary Fig. 1. However, we have seen that users, be they human-operated or bots, who take part in a significant and effective way in the discussion, are mostly polarized.

Finally, we extract the non trivial content exchange by adopting the validated projection developed by Becatti et al in ref. [32] in order to detect the significant flow of messages among users, while discounting the virality of messages, the retweeting activity of users and their productivity in writing tweets.

The network represented in Fig. 5 is extremely informative for different reasons. The validated network contains only 14,883 validated users out of the 127,275 users in the dataset. This highlights the fact that just a minority of all users effectively contributes to the online propaganda on the migration flow. Interestingly, the incidence of bots on the validated network is almost one third of the analogous measure on the entire dataset, signaling that the number of bots whose retweets are non compatible with a random activity is a minority. Since one of the targets of a social bot is to increase audience of the online content of a specific (group of) user(s), such a reduction shows that, in our scenario, the number of bots affecting significantly the political discussion is limited.

The accounts in the validated network are much more polarized than the whole set of users in the original network, see Fig. 10. In fact, in the original network, the overall fraction of unpolarized accounts represents more than 40% of all the accounts and more than 50% of all the automated ones. Instead, when considering the validated network, the same ratio is around 10% for the former and around 5% for the latter. Otherwise

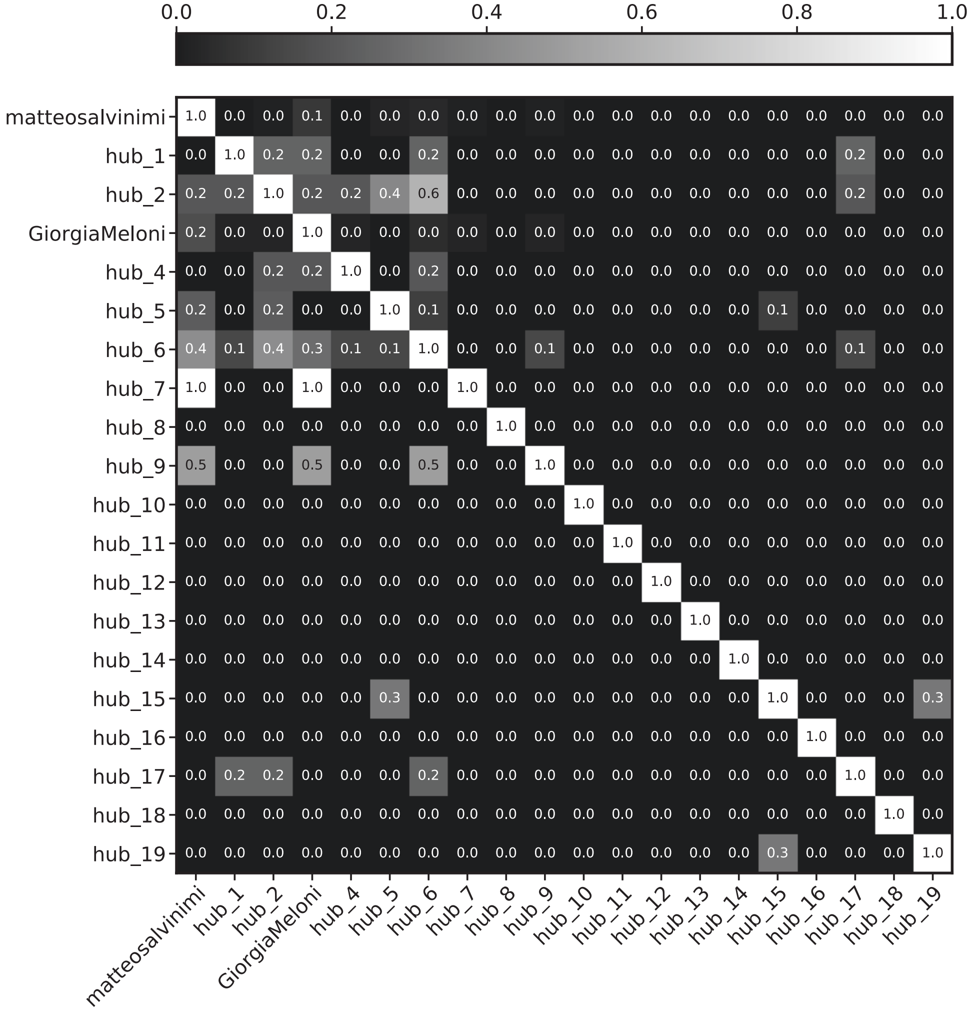

Fig. 6 The relative overlap matrix among list of bots following the top 20 hubs. The generic matrix entry represents $\frac{|\mathrm{bot}_i \cap \mathrm{bot}_j|}{|\mathrm{bot}_i|}$, i.e. the percentage of shared bots between users *i* and *j* over the number of bots following node *i*. There are 12 accounts sharing a relatively high number of bots.

stated, the polarized bots pass the validation process more easily than their unpolarized counterparts and their contribution in spreading messages is more significant.

All the accounts that are mostly effective in delivering their messages (i.e., the Hubs, by following the paradigm defined by Kleinberg in ref. [49]) refer to the blue area in Fig. 5, where we can find representatives of the Italian government in charge at the time of data collection, and the right wing. The first account referring to a community different from the blue one is the official account of the newscast 'TgLa7', at position 176th in the hub ranking.

Regarding the contribution of bots to the visibility of the various accounts, the fraction of bots that significantly retweet the content of two right wing political leaders (Mr. Salvini and Ms. Meloni) is greater than the incidence of bots in the whole validated network. Interestingly enough, other hubs show a smaller presence of bots among their followers, even if their hub score is not that different from the two political leaders.

Finally, we found that some hubs do share their bots: Fig. 6 describes the normalized overlap between the list of bots of each pair of users in the list of the top 20 hubs. As mentioned before, those accounts are from the right wing political area. To the best of our knowledge, this is the first time that such a behaviour is reported: in analyses tackling the same problem, i.e., the interactions among human-operated accounts and bots[9,10,12], only star-like sub-graphs were observed, with a big number of bots among the followers of a (presumably) human-operated user. We have found bunches of bots attached to bunches of hubs, and although we cannot make any claim about the mind and the strategy behind this organization, we find the result noteworthy, and consider interesting to look for the phenomenon in other datasets.

We underline that the considered shared bots are particularly effective, since they are validated by the entropy-based projection. Actually, the group of right wing bots, each supporting more than a human-operated account, is not the only one in the set, but it is the greatest: if we consider the subgraphs of human-operated accounts sharing their bots, see Figs. 7 and 8, the former has 172 nodes against 58 of the latter. Moreover the first subgraph is by far more efficient; indeed, in the second one the greatest hub score ranks 176th.

It is well known that bots aim at increasing popularity of users by retweeting their messages (see, e.g., work by Cresci et al.[13], that reveals how bots, e.g., retweet in a coordinated fashion celebrities' accounts). The projection procedures followed in this paper reveal such a coordinated activity in a pretty robust way. In fact, we argue that the emergence of statistically significant communication patterns could hardly be hidden by an attacker, because the latter would have to employ more automated accounts to 'hide' their activities within the expectations of the probability distribution obtained by the BiDCM.

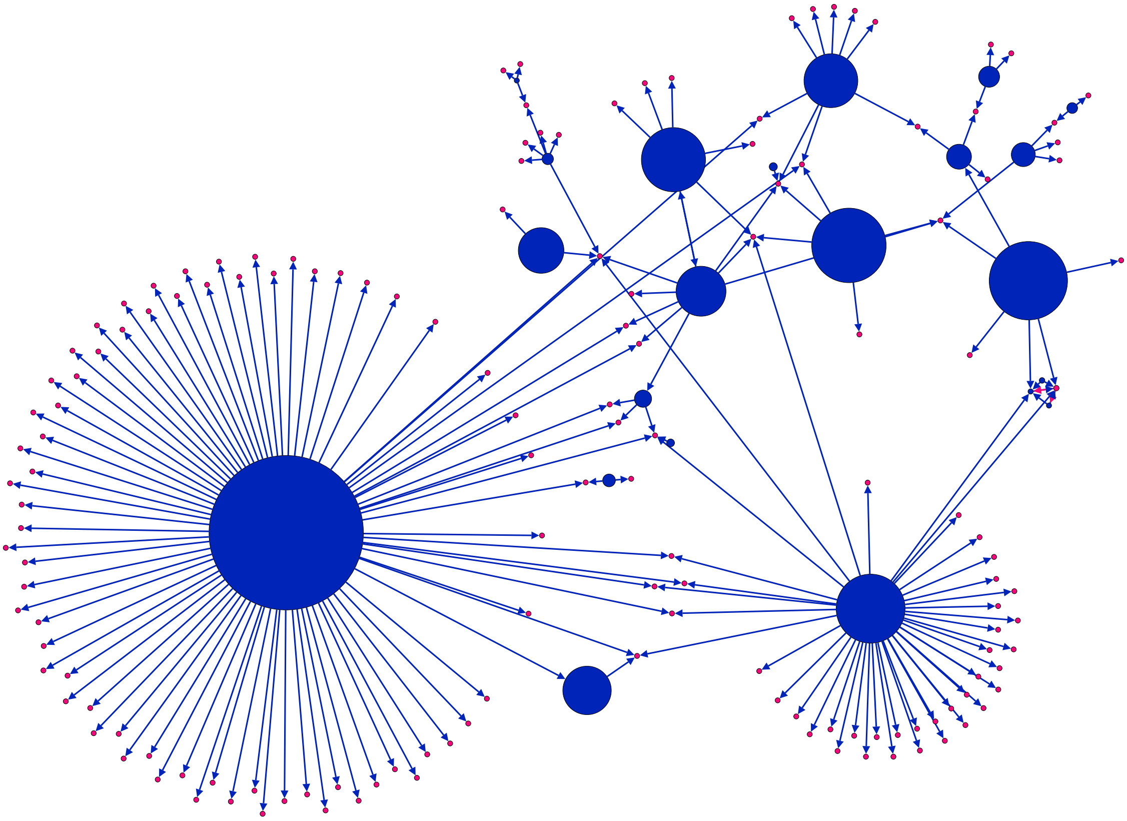

Fig. 7 Subgraph of the largest group of users sharing bots. The subgraph includes genuine accounts (in dark blue) and all the bots following them (in magenta). The dimension of the nodes is proportional to their hub score, but normalised on the subgraph. The biggest node represents the account of Mr. Salvini. In the picture, there are 22 bots shared by 22 humans. Among the latter, 9 accounts are among the top 10 hubs. The subgraph contains 172 nodes. Notably, the accounts belong almost exclusively to the blue, i.e., pro-government, community.

Table 2 Keywords used for collecting tweets about migrant flows from Northern Africa.

Keywords	
Immigrati	Immigrants
Migranti	Migrants
Ong	NGO
Scafisti	Boat drivers as human smugglers
Seawatch	A NGO operating in the Mediterranean Sea
Barconi	Barges/boats
Clandestini	Illegal immigrants
Guardia costiera libica	Lybian coast guard
Naufragio	Shipwreck
Sbarco	disembarkation

Keywords have been searched in Italian, the English correspondents are at their right.

To the best of our knowledge, our study is the first investigation that merges bot detection and entropy-based analysis of Twitter traffic. Moreover, the obtained results are in line with the previous work by Shao et al.[8], where the authors showed how bots massively support the spread of (low credibility) content. At the same time, the present investigation contributes in a different way, being not specifically focused on fake news, whereas[8] concentrates on the way fake news become viral. Interestingly enough, among the many studies about the 2016 US presidential election, Grinberg et al.[11] analyzed the proliferation of fake news on Twitter and determined both fake news spreaders and exposed users. The role of bots in effectively conveying a message—for the first time here highlighted even in a 'shared fashion'—and the spreading of fake news in online discussions of great importance, e.g., about elections and news reports, see refs. [8,11] calls for future studies, which include a deeper analysis of the exchanged messages.

Methods

Data collection and processing. Our study is based on a large corpus of Twitter data, generated by collecting tweets about migrations, and focusing on the case of migrant flows from Northern Africa to Italy. For data collection, we developed a crawler based on the Twitter public Filter API, which provides real-time tweet delivery, filtered according to specified keywords. We selected a set of keywords compatible with recent chronicles. Table 2 lists the selected keywords. The filtering procedure was not case-sensitive. The keywords have been selected because they are commonly used in Italy when talking and writing about immigration flows from Northern Africa to the Italian coasts, including the dispute about the holder of jurisdiction for handling emergencies, involving European countries and NGOs (https://en.wikipedia.org/wiki/African_immigration_to_Europe).

We collected 1,082,029 tweets, posted by 127,275 unique account IDs, over a period of one month (from 23 January 2019 to 22 February 2019). By relying on the bot detection classifier developed by Cresci et al. in[14] and recapped in the following section, all the accounts have been classified either as human-operated or as bots. This classification led to 117,879 genuine accounts and 9,396 social bots. All the collected tweets were stored in Elasticsearch (https://www.elastic.co) for fast and efficient retrieval.

It may be worth noting that the period over which the data have been collected was characterized by a lively political debate in Italy about the landing of one ship operated by NGOs rescuing migrants fleeing from North Africa to Italy. On 16 August 2018, the Italian coastguard boat 'Diciotti' rescued almost 200 migrants off Lampedusa island and initially received a veto to land from the Italian government; it was allowed to do so only after 10 days. Mr. Matteo Salvini, at that time Minister of Internal Affairs, was afterwards investigated for kidnapping and abuse of office; the case was stopped on 19 February 2019, when the Italian Senate did not grant judges the possibility to prosecute him. Right before and after the Senate's decision there was an intense debate on social networks about migrants and NGOs, and about the role of Italian Government and of the European Union.

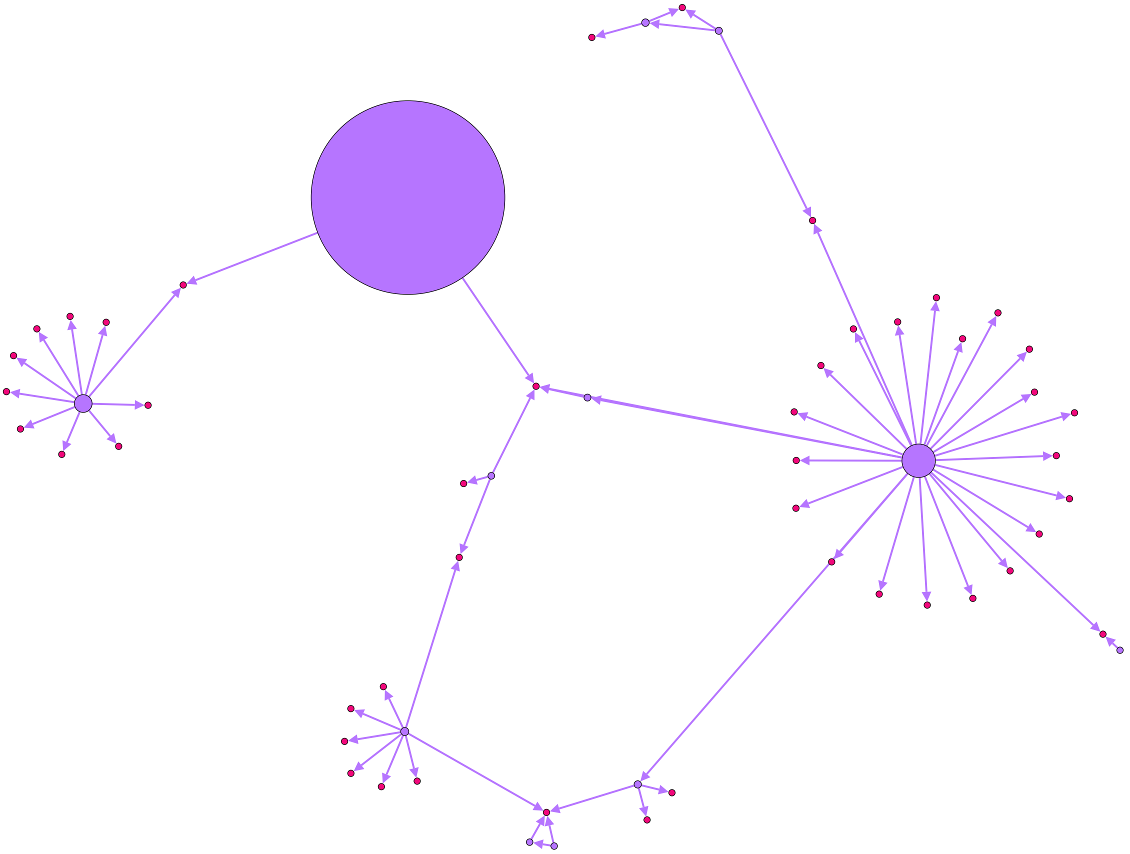

Fig. 8 Subgraph of the second largest group of users sharing bots. The subgraph includes human-operated accounts (in violet) and all the bots following them (in magenta). The dimensions of the nodes are proportional to the hub scores, but normalised on the subgraph. The impact on the validated network of the nodes in this subgraph is much more limited: the largest hub is the official account of 'TgLa7', a newscast ranking 176th in the hub scores. Considering even the non shared bots, the subgraph contains 58 nodes. Notably, accounts belong almost exclusively to the purple community.

Table 3 Features adopted for the fake account detector designed in ref. [14] and here re-constructed.

Features
Friends count
Followers count
Tweets count
$\frac{friends}{followers^2}$
Account age
Following rate (approximated as $\frac{friends}{age}$)
The account's profile has a name
The account's profile has an image
The account's profile has an address
The account's profile has a biography
The account's profile has a URL
The account belongs to a list
$2 \times followers \geq friends$
$\frac{friends}{followers} \simeq 100$
$\frac{friends}{followers} \geq 50$

Bot detection classifier. To assess the nature of the accounts in the dataset about migration from Northern Africa, we rely on a slightly modified version of the supervised classification model proposed in ref. [14].

The bot detector uses all the features of an account profile, and has been developed after testing known detection methodologies on a baseline dataset of fake and genuine accounts. The former were bought from three online markets, while the latter were certified as genuine by tech-savvy social media analysts. The authors tested Twitter accounts in the reference set against algorithms based on: (i) single classification rules proposed by Media and bloggers, and (ii) feature sets proposed in the literature for detecting spammers. The results of such preliminary analysis suggested that fake accounts detection needs specialized mechanisms. They classified rules and features according to the cost required for gathering the data needed to compute them and showed how the best performing features are also the most costly ones. Then, building on the cost of crawling analysis, they implemented a series of lightweight classifiers using less costly features, while still being able to correctly classify more than 95% of the accounts of the baseline dataset. They also validated performances of the classifiers over two other sets of human and fake accounts, disjoint from the original training dataset.

For the present paper, we reconstruct the model of the classifier of Cresci et al.[14] and test its performances with J48, the Weka (https://www.cs.waikato.ac.nz/ml/weka/) implementation of C4.5 algorithm, on the same training set, publicly available at http://mib.projects.iit.cnr.it/dataset.html, obtaining the same classification performance results. The used features are listed in Table 3.

As a final note, it seems fair to underline how the war against automated, malicious accounts is constantly in progress. In fact, social bot developers have become so smart that their creations are very similar to genuine accounts. An interesting line of research is that of 'adversarial social bot detection', where, instead of taking countermeasures only after having collected evidence of new bot mischief, detection techniques are proactive and able to anticipate attacks and next generations of bots. The first seeds for this research were put by Yang et al. back in 2011–13 who provided the first evidence of social bot evolutions[70]. While the first wave of social bots, populating OSNs until around 2011, were rather simplistic, the second wave featured characteristics that were quite advanced for the time. Differently from the previous ones, the social bots studied by Yang et al. were used to purchase or exchange followers between each other, in order to look more popular and credible. Bot evolution thus leads us to 2016, when Ferrara et al. documented a third generation of social bots[71]. Needless to say, Yang's classifier was no longer successful at detecting this third wave of social bots, as experimentally demonstrated in ref. [13] by Cresci et al.

After Yang et al.'s first adversarial work, many years passed before further studies were carried out. Indeed, only recently Cresci et al.[61] and Grimme et al.[72] proposed new adversarial studies in social bot detection. The continuous research around the theme of social bot detection leads to more and more accurate techniques, with increasingly reduced errors in classification.

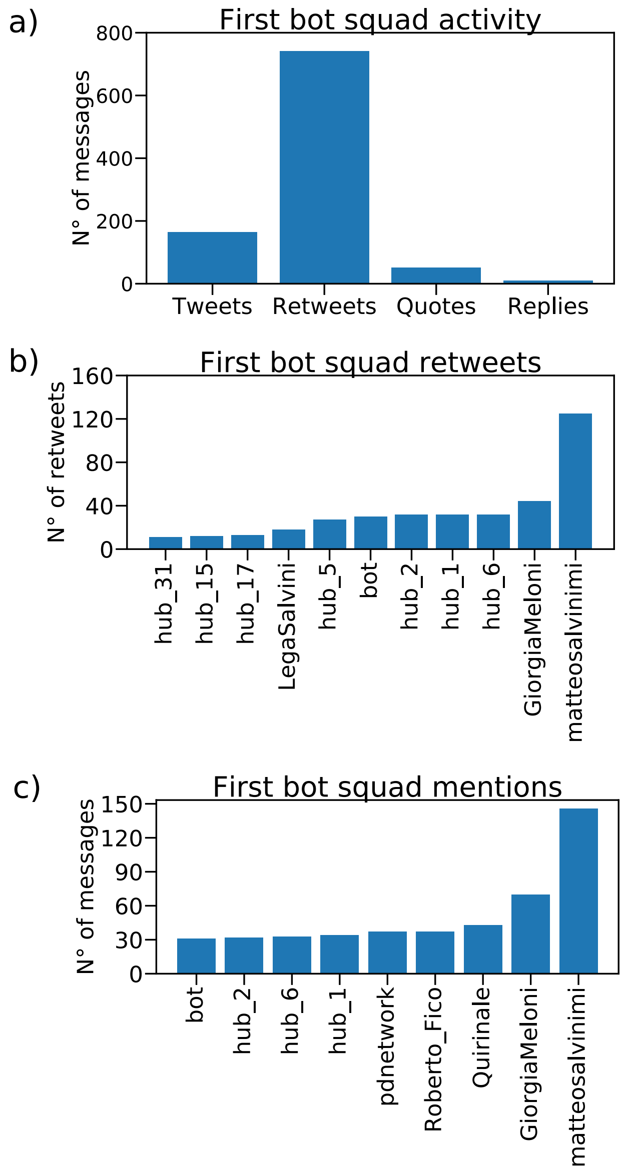

Fig. 9 Statistics of the largest bot squad. a The main activity of the bot squad is retweeting. **b** The target of the retweets of the bot squads are the strongest hubs of the validated directed network. **c** The situation is different for the mentions in the bots' original tweets: bots in this group mention the accounts of the right wing (Matteo Salvini, Giorgia Meloni and some other strong hubs of the validated network), accounts of the left wings (as the official account of the Democratic Party) and the official account of some super-partes institutional figures.

Validated projection of the bipartite network and users polarization. Because of the official certification released by Twitter about the authenticity of an account, users can be divided into two sets, the verified and unverified ones. Becatti et al., in ref. [32], used this feature to infer the accounts' inclination towards a specific political area, directly from data. This has been possible by implementing the method of Saracco et al. proposed in ref. [66]. The underlying idea is that unverified users follow and interact with verified users sharing their political ideals. In this sense, if two verified users have a high number of common followers and followees, they probably have a similar political affiliation. The a posteriori analysis of the results of the validated projection confirms the previous hypothesis. Due to the Twitter verification procedure, only the information provided by verified users is fact-checkable, thus our check is restricted to this class of users.

We have to pay attention to the contribution of remarkably active users. For example, if a verified user is extremely engaged in the political propaganda, such user may interact with a huge number of unverified ones and may thus share a great amount of contacts with almost all the other verified users, even those with an opposite political inclination. In this case, the contribution should be considered spurious, being just due to the popularity of the user. Analogously, the role of an unverified user that retweets all messages from her/his contacts should be discounted.

We obtain the political affiliation of the accounts by considering the undirected bipartite network of interactions (i.e., retweets) between verified and unverified users, aggregated over the whole period: we disregard the information about the direction of the retweets, since we are just interested in groups of users sharing content. The previous intuition leads us to compare the overlap of connections (literally, the number of common followers and followees) in the real network with the expectations of a null-model able to account for the degree sequence of both layers. In this way, we are able to discount the random noise due to the activity of users and get the statistically significant information from the data. The entropy-based BiCM, ref. [66] provides the correct benchmark for this analysis. While we describe more extensively the theoretical construction in the Supplementary Notes 2 and 3, here we outline the main intuitions behind the Bipartite Configuration Model and its monopartite validated projection.

The Bipartite Configuration Model: let us start from a real bipartite network and call the two layers L and Γ and their dimension respectively $N_{\rm L}$ and N_Γ; we label the nodes on those layers respectively with Latin and Greek indices. We represent the connection via the biadjacency matrix, i.e., the rectangular ($N_{\rm L} \times N_\Gamma$)-matrix $\mathbf{M}$ whose generic entry $m_{i\alpha}$ is 1 if there is a link connecting node $i \in {\rm L}$ and node $\alpha \in \Gamma$, and 0 otherwise. We then consider the ensemble $\mathcal{G}_{\rm Bi}$ of all possible graphs with the same number of nodes on the two layers as those in the real network. If we assign a (formal) probability per graph, we can maximize the (Shannon) entropy,

$$S = -\sum_{G_{\rm Bi}\in\mathcal{G}_{\rm Bi}} P(G_{\rm Bi})\ln P(G_{\rm Bi}),$$

constraining the average value of some quantities of interest on the entire ensemble. If, as it is the case of the present article, we impose the ensemble to have fixed average for the number of links per node (i.e., the degree), the probability per graph factorizes in independent probabilities per link:

$$p_{i\alpha} = \frac{x_i y_\alpha}{1 + x_i y_\alpha}, \tag{2}$$

where $_{pi\alpha}$ is the probability of finding a link between i and α and x_i and y_α (the fitnesses, as defined by Caldarelli et al. in ref. [73]) are quantities that encode the attitude of the nodes to form links[22]. At this level, the previous definition is formal, since we just imposed to fix the average (over the ensemble) of the degree sequence, but we did not decide its value. It can be shown (see Supplementary Notes 2 and 3) that maximizing the likelihood of the real network is equivalent to fixing the average of the degree sequence to the one measured on the real network (proofs by Garlaschelli and Squartini, see refs. [23,24]).

Monopartite validated projection: we can now highlight all those contributions that cannot be related to the degree sequence only, comparing the real network with the expectations of the BiCM. Following this line, Saracco et al.[66] proposed a validated projection on top of the BiCM. The main idea is to consider the common links of two nodes on the same layer and compare it with the theoretical distribution of the BiCM: if the real system shows a commonality of links that cannot be explained only by the activity of the users, then we project a link between the two nodes under analysis. Using the formalism of Saracco et al.[65], we call V-motif the overlap.

In formulas, by using the independence of probabilities per graph (2), the probability that both node i and j link the same node α is simply

$$p(V_\alpha^{ij}) = p_{i\alpha} p_{j\alpha},$$

where V_α^{ij} is the above mentioned V-motif among i, j and α. The total overlap between i and j is simply $V^{ij} = \sum_\alpha V_\alpha^{ij}$ and, according to the BiCM, is distributed as a Poisson-binomial, i.e., the extension of a binomial distribution in which all the events have a different probability (see Hong[74]). We can further associate p-values to the observed V-motifs, i.e., the probabilities of finding a number of V-motifs greater than or equal to the one measured on the real network. In order to state the statistical significance of several p-values at the same time, we relied on a multiple test hypothesis. The false discovery rate (FDR) is generally considered the most effective one since it permits to control the number of false negatives, without being too conservative (see Benjamini and Hockberg[75]). The result of the projection is a binary undirected monopartite network of nodes from the same layer, that are linked if their similarity cannot be explained only by their degree. We therefore apply the Louvain community detection algorithm (by Blondel et al.[44]). Since this method is known to be order dependent, as shown by Fortunato[45], we apply it several times after reshuffling the node order and take the maximum value of the modularity, i.e. the algorithm objective function.

The interested reader can find in the Supplementary Note 8 an alternative approach to assign a political inclination to users and a comparison with the approach followed in this paper (see also Supplementary Fig. 2).

Extraction of the backbone of tweeting activity. As done when evaluating the statistical significance of the common links of two nodes on the same layer, also when studying content exchange, we are interested in the flow of information that cannot be explained by users' activity only. Differently from other studies[4,9,10,53], we take into account the virality of tweets. Methodologically, the

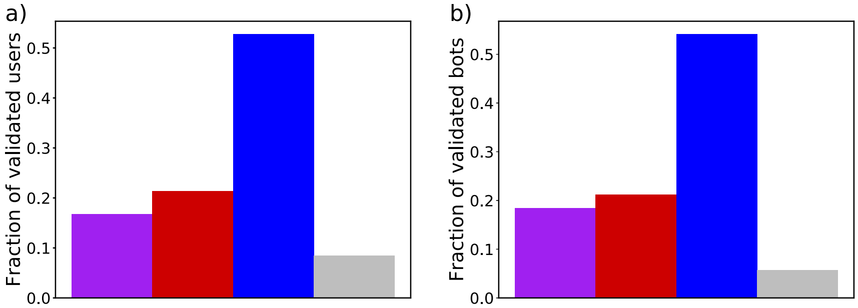

Fig. 10 Validated users affiliation, after the polarisation by contagion. a Fraction of users affiliated to the three main communities, in the validated network, after the contagion of polarization. Nodes are eggplant purple for NGO's, media accounts and left wing politicians, tomato red for the Democratic Party community and Mediterranean blue for the pro-government users. In grey, users which cannot be assigned uniquely to one community. **b** The same histogram for bots in the validated network. The fraction of validated accounts for which the polarization procedure is not able to produce a unique output is much smaller than the total network. Indeed, most of the unpolarized accounts do not significantly contribute to contents exchange in the validated network. We ran a Pearson Chi square test, comparing the polarization distribution of the total number of users and the one for the validated ones only. The difference is not statistically significant, having a p-value close to 1.

approach is similar to the one adopted for the extraction of the users political affiliation. The difference consists in (1) substituting the BiCM with its analogous directed version, the BiDCM (proposed by van Lidth de Jeude et al.[48]) and in (2) considering layers of different kind. While in the previous section layers represent verified and unverified users, here they represent tweets on one layer and users (both verified and unverified) on the other. The validated projection procedure returns a directed monopartite network of significant exchange of messages, in which the arrow goes from the message author to the retweeters. As mentioned in the Results section, the connectance of the network is $\rho \simeq 3\times 10^{-5}$. The effective FDR-threshold for p-values is $\mathrm{FDR_{th}} \simeq 3.0\times 10^{-7}$ for $\alpha = 0.01$; the analogous Bonferroni effective threshold is much stricter, $\mathrm{Bonferroni_{th}} \simeq 8.8\times 10^{-12}$. Additional details about the BiDCM procedure can be found in the Supplementary Notes 5 and 6.

Data availability

The datasets analyzed in this study are available under the limits of Twitter's Developer Terms at toffee.imtlucca.it.

Code availability

The authors are currently working on releasing the bot detector model and a Python package implementing the noise filtering in the bipartite directed network. When available, they will be published at toffee.imtlucca.it, together with instructions for their use.

Received: 21 June 2019; Accepted: 25 March 2020;
Published online: 11 May 2020

References

1. Kwak, H., Lee, C., Park, H. & Moon, S. What is Twitter, a social network or a news media? In *Proc. 19th International Conference on World Wide Web, WWW '10*, 591–600 (ACM, New York, 2010).
2. Hu, M. et al. Breaking news on Twitter. In *Proc. SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, 2751–2754 (ACM, New York, 2012).
3. Gangware, C. & Nemr, W. *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age* (Park Advisors, 2019).
4. Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **10**, 7 (2019).
5. Bastos, M. T. & Mercea, D. TheBrexit botnet and user-generated hyperpartisan news. *Soc. Sci. Comput. Rev.* **37**, 38–54 (2017).
6. Ferrara E. Manipulation and abuse on social media. *ACM SIGWEB Newsl.* **4**, 9 (2015).
7. Ruths, D. The misinformation machine. *Science* **363**, 348–348 (2019).
8. Shao, C. et al. The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787 (2018).
9. Stella, M., Cristoforetti, M. & De Domenico, M. Influence of augmented humans in online interactions during voting events. *PLoS ONE* **14**, 1–16 (2019).
10. Stella, M., Ferrara, E. & Domenico, M. D. Bots sustain and inflate striking opposition in online social systems. *Proc. Natl Acad. Sci. USA* **115**, 12535–12440 (2018).
11. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Political science: fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
12. Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In *Proc. 11th International Conference on Web and Social Media, ICWSM 2017,* 280–289 (AAAI, Montréal, 2017).
13. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proc. 26th International Conference on World Wide Web Companion WWW'17*, 963–972 (ACM, 2017).
14. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. Fame for sale: efficient detection of fake twitter followers. *Decis. Support Syst.* **80**, 56–71 (2015).
15. Satya, P. R. B., Lee, K., Lee, D., Tran, T. & Zhang, J. J. Uncovering fake likers in online social networks. In *Proc. 25th ACM International Conference on Information and Knowledge Management, CIKM 2016*, 2365–2370 (Association for Computing Machinery, Indianapolis, 2016).
16. Yuan, S., X., Wu, J., Li & A., Lu Spectrum-based deep neural networks for fraud detection. In *Proc. Conference on Information and knowledge, CIKM*. (ACM, 2017).
17. Wang, B., Gong, N. Z. & Fu, H. GANG: detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In Proc. *2017 IEEE International Conference on Data Mining, ICDM 2017,* 465–474 (IEEE, New Orleans, 2017).
18. Liu, S., Hooi, B. & Faloutsos, C. Holoscope: Topology-and-spike aware fraud detection. In *Proc. 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017,* 1539–1548 (Association for Computing Machinery, Singapore, 2017).
19. Giatsoglou, M. et al. ND-Sync: Detecting synchronized fraud activities. In *Advances in Knowledge Discovery and Data Mining* (eds. Cao, T. et al.) 201–214 (Springer International Publishing, Cham, 2015).
20. Chavoshi, N., Hamooni, H. & Mueen, A. DeBot: Twitter bot detection via warped correlation. In *Proc. 2016 IEEE 16th International Conference on Data Mining*, 817–822 (IEEE, 2016).
21. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Trans. Dependable Secur. Comput.* **15**, 561–576 (2018).
22. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 66117 (2004).
23. Garlaschelli, D. & Loffredo, M. I. Maximum likelihood: extracting unbiased information from complex networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **78**, 1–5 (2008).
24. Squartini, T. & Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.* **13**, 083001 (2011).
25. Cimini, G. et al. The statistical physics of real-world networks. *Nat. Rev. Phys.* **1**, 58–71 (2018).
26. Squartini, T. & Garlaschelli, D. *Maximum-entropy networks. Pattern detection, network reconstruction and graph combinatorics* (Springer International Publishing, 2017).

27. Squartini, T., Caldarelli, G., Cimini, G., Gabrielli, A. & Garlaschelli, D. Reconstruction methods for networks: the case of economic and financial systems. *Phys. Rep.* **757**, 1–47 (2018).
28. Squartini, T., van Lelyveld, I. & Garlaschelli, D. Early-warning signals of topological collapse in interbank networks. *Sci. Rep.* **3**, 3357 (2013).
29. Saracco, F., Di Clemente, R., Gabrielli, A. & Squartini, T. Detecting early signs of the 2007–2008 crisis in the world trade. *Sci. Rep.* **6**, 30286 (2016).
30. Gualdi, S., Cimini, G., Primicerio, K., Di Clemente, R. & Challet, D. Statistically validated network of portfolio overlaps and systemic risk. *Sci. Rep.* **6**, 39467 (2016).
31. Di Gangi, D., Lillo, F. & Pirino, D. Assessing systemic risk due to fire sales spillover through maximum entropy network reconstruction. *J. Econ. Dyn. Control* **94**, 117–141 (2018).
32. Becatti, C., Caldarelli, G., Lambiotte, R. & Saracco, F. Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections. *Palgrave Commun.* https://doi.org/10.1057/s41599-019-0300-3 (2019).
33. Adamic, L. A. & Glance, N. S. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proc. 3rd International Workshop on Link discovery, LinkKDD 2005,* 36–43 (Association for Computing Machinery, Chicago, 2005).
34. Conover, M., Ratkiewicz, J. & Francisco, M. Political polarization on twitter. In *Proc. Fifth International AAAI Conference on Weblogs and Social Media.* (2011).
35. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. & Menczer, F. Predicting the political alignment of twitter users. In *Proc. PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom)*, 192–199 (IEEE, 2011).
36. Conover, M. D., Gonçalves, B., Flammini, A. & Menczer, F. Partisan asymmetries in online political activity. *EPJ Data Sci.* **1**, 6 (2012).
37. Del Vicario, M. et al. Echo chambers: emotional contagion and group polarization on Facebook. *Sci. Rep.* (2016).
38. Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A. & Quattrociocchi, W. Mapping social dynamics on Facebook: the Brexit debate. *Soc. Networks* **50**, 6–16 (2017).
39. Quattrociocchi, W., Caldarelli, G. & Scala, A. Opinion dynamics on interacting networks: Media competition and social influence. *Sci. Rep.* **4**, 4938 (2014).
40. Zollo, F. et al. Debunking in a world of tribes. *PLoS ONE* **12**, 0181821 (2017).
41. Zollo, F. et al. Emotional dynamics in the age of misinformation. *PLoS ONE* **10**, 0138740 (2015).
42. Varol, O. & Uluturk, I. Journalists on Twitter: self-branding, audiences, and involvement of bots. *J. Comput. Soc. Sci.* **3**, 83–101 (2019).
43. Hentschel, M., Alonso, O., Counts, S. & Kandylas, V. Finding users we trust: Scaling up verified Twitter users using their communication patterns. In proc. *International AAAI Conference on Web and Social Media.* (The AAAI Press, 2014).
44. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10008**, 6 (2008).
45. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
46. Bessi, A. et al. Users polarization on Facebook and Youtube. *PLoS ONE* **11**, 0159641 (2016).
47. Bessi, A. et al. Homophily and polarization in the age of misinformation. *Eur. Phys. J. Spec. Top.* **225**, 2047–2059 (2016).
48. van Lidth de Jeude, J., Clemente, R. D., Caldarelli, G., Saracco, F. & Squartini, T. Reconstructing mesoscale network structures. *Complexity* **2019**, 5120581:1–5120581:13 (2019).
49. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM* **46**, 324133–324140 (1999).
50. Publication Office of the European Union. *Media use in the European Union.* (2017).
51. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
52. AGCOM. *Report on the consumption of information*. Technical Report. (Autorità per le Garanzie delle Comunicazioni, 2018).
53. Bovet, A., Morone, F. & Makse, H. A. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Sci. Rep.* **8**, 8673 (2018).
54. Bekafigo, M. A. & McBride, A. Who tweets about politics?: Political participation of Twitter users during the 2011 gubernatorial elections. *Soc. Sci. Comput. Rev.* **31**, 625–643 (2013).
55. Borondo, J., Morales, A. J., Losada, J. C. & Benito, R. M. Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study. *Chaos* **22**, 023138 (2012).
56. Vicario, M. D., Gaito, S., Quattrociocchi, W., Zignani, M. & Zollo, F. News consumption during the Italian referendum: A cross-platform analysis on Facebook and Twitter. In *Proc. IEEE Conference on Data Science and Advanced Analytics*, 648–657 (IEEE, 2017).
57. Di Grazia, J., McKelvey, K., Bollen, J. & Rojas, F. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE* **8**, e79449 (2013).
58. González-Bailón, S., Borge-Holthoefer, J., Rivero, A. & Moreno, Y. The dynamics of protest recruitment through an online network. *Sci. Rep.* **1**, 197 (2011).
59. González-Bailón, S., Borge-Holthoefer, J. & Moreno, Y. Broadcasters and hidden influentials in online protest diffusion. *Am. Behav. Sci.* **57**, 943–965 (2013).
60. Ferrara, E. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* **22**, https://doi.org/10.5210/fm.v22i8.8005 (2017).
61. Cresci, S., Petrocchi, M., Spognardi, A. & Tognazzi, S. Better Safe than Sorry: an adversarial approach to improve social bot detection. In *Proc. 11th International ACM Web Science Conference*, 47–56 (Association for Computing Machinery, 2019).
62. Balestrucci, A., De Nicola, R., Petrocchi, M. & Trubiani, C. Do you really follow them? automatic detection of credulous Twitter users. In *Proc. Intelligent Data Engineering and Automated Learning -IDEAL 2019 - 20th International Conference*, 402–410. (Springer International Publishing. Manchester, 2019).
63. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
64. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.* **16**, 043022 (2014).
65. Saracco, F., Di Clemente, R., Gabrielli, A. & Squartini, T. Randomizing bipartite networks: the case of the World Trade Web. *Sci. Rep.* **5**, 10595 (2015).
66. Saracco, F. et al. Inferring monopartite projections of bipartite networks: An entropy-based approach. *New J. Phys.* **19**, 16 (2017).
67. Becatti, C., Caldarelli, G. & Saracco, F. Entropy-based randomization of rating networks. *Phys. Rev. E* **99**, 022306 (2019).
68. Schmidt, A. L., Zollo, F., Scala, A., Betsch, C. & Quattrociocchi, W. Polarization of the vaccination debate on Facebook. *Vaccine* **36**, 3606–3612 (2018).
69. Nikolov, D., Oliveira, D. F., Flammini, A. & Menczer, F. Measuring online social bubbles. *PeerJ Comput. Sci.* **1**, e38 (2015).
70. Yang, C., Harkreader, R. & Gu, G. Empirical evaluation and new design for fighting evolvingTwitter spammers. *IEEE Trans. Inform. Forensics and Secur.* **8**, 1280–1293 (2013).
71. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
72. Grimme, C., Assenmacher, D. & Adam, L. Changing perspectives: Is it sufficient to detect social bots? In *Proc. International Conference on Social Computing and Social Media*, 445–461 (Springer, 2018).
73. Caldarelli, G., Capocci, A., De Los Rios, P. & Muñoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**, 258702 (2002).
74. Hong, Y. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.* **59**, 41–51 (2013).
75. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

Acknowledgements

F.D.V. and G.C. acknowledge support by EC TENDER SMART (Grant No. 2017/ 0090 LC-00704693). F.S. and G.C. acknowledge the EU project SoBigData (Grant No. 654024). R.D.N., G.C., M.P. and F.S. acknowledge support from the Project TOFFEe (TOols for Fighting FakEs) funded by IMT School for Advanced Studies Lucca. All the authors acknowledge enlightening discussions with Andrea Nicolai and Simona De Rosa from the Social Observatory for Disinformation and Social Media Analysis (SOMA) and with Filippo Menczer from the School of Informatics, Computing & Engineering of Indiana University.

Author contributions

F.D.V. contributed to the acquisition of data. F.D.V. and F.S. contributed to the analysis of data. M.P. and F.S. contributed to the conception and design of work, and to the paper draft. G.C., R.D.N., M.P. and F.S. contributed to the interpretation of data and to the paper revision. Each author approved the submitted version and agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42005-020-0340-4.

Correspondence and requests for materials should be addressed to M.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.