Supplementary Information for "Bots increase exposure to negative and inflammatory content in online social systems"

Massimo Stella[1], Emilio Ferrara[2] and Manlio De Domenico[1]

1. Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy
2. USC Information Sciences Institute, 4676 Admiralty Way 1001, Marina del Rey, CA 90292 USA

Corresponding authors: mdedomenico@fbk.eu; emiliofe@usc.edu

This is the Supplementary Information for the main text.

1. Bot detection techniques

One of the most daunting tasks in social media analysis is determining whether a user account is controlled by a human or a software (i.e., a bot). A great deal of research aims to address this issue, including our own efforts [1,2] and others' [3-6].

In this realm, *Botometer*[6] (formerly BotOrNot) represents, as of today, the only openly accessible solution [7] . It consists of an Application Programming Interface (API) developed in Python which allows to programmatically interact with the underlying machine learning system. *Botometer* has been proven quite accurate in detecting social bots [2,7].

However, the public interface of *Botometer* has two limitations that prevented us to use it in this project: the framework relies on the Twitter API to collect recent data about the accounts to inspect. The Twitter API imposes very strict query rate limits, therefore making it impossible to analyze more than a few thousand accounts with the public *Botometer* Python API. In this study, our goal is to detect bots in a very large population of over 2 million users, requiring an ad hoc large-scale bot detection solution. The second limitation is once again derived by the Twitter API: when *Botometer* inspects an account that has been either suspended, protected, quarantined, or deleted, the Twitter API does not provide any details about it, rendering *Botometer* unable to make any determination. Since this study will show that a significant portion of bot accounts involved in *MacronLeaks* has been either suspended, quarantined, or deleted shortly after Election Day (May 7, 2017), *Botometer* would not represent a suitable tool to analyze them.

www.pnas.org/cgi/doi/10.1073/pnas.1803470115

2. Our bot detection approach

For the reasons mentioned above, we decided to implement a simple yet accurate bot-detection algorithm reflecting the following requirements:

- The algorithm is accurate yet scalable and can be used to classify the over 2 million users present in our dataset in a reliable yet timely manner; this will address the scalability issues of *Botometer*.
- The algorithm can use historical tweets and account metadata collected and available in our dataset to determine bots and humans, without the need to query the Twitter API for recent data; this will address the limits imposed by the Twitter API and allow us to analyze all users, not only the active ones at the time of inspection, but also the suspended, protected, quarantined, and deleted accounts.
- The algorithm builds on top of the insights and lessons learned from the development of *Botometer*.

Botometer's underlying machine-learning framework generates a set of over one thousand features, spanning content and network structure, temporal activity, user profile data, and sentiment analysis. These indicators are aggregated and analyzed to determine the likelihood that the inspected account is a bot. Feature analysis revealed that the two most important classes of feature to detect social bots are the metadata and usage statistics associated with a user account [2,7]. We previously illustrated [8] that the following indicators provide the strongest signals to separate humans from, in particular, political bots: *(i)* whether the public Twitter profile looks like the default one or it is customized (it requires human efforts to customize a profile, therefore bots are more likely to exhibit the default setting); *(ii)* absence of geographical metadata (humans often use smartphones and the Twitter iPhone/Android App, which records the physical location of the mobile device as digital footprint); and, *(iii)* activity statistics such as total number of tweets and frequency of posting (bots exhibit incessant activity and excessive amounts of tweets), proportion of retweets over original tweets (bots retweet contents much more frequently than generating new tweets), proportion of followers over followees (bots usually have less followers and more followees), the number of times a user has been added to a public list – human users are often considered more influential [9] and their content more "contagious" [10-13].

3. Detection algorithm and features

Considering these insights, we used the following user metadata and activity features to create a simple yet effective bot detection algorithm:[7]

1) "statuses_count": number of tweets posted by the given user;
2) "followers_count": number of followers of the given user;
3) "friends_count": number of followees (friends) of the given user;
4) "favourites_count": number of favorited tweets of the given user;
5) "listed_count": number of times the given user has been added to a list;

6) "default_profile": binary field that indicates whether the user profile has the default setting or not;
7) "geo_enabled": binary field that indicates whether the geo-coordinates of the user are available;
8) "profile_use_background_image": binary field that indicates whether the user profile has the default image or a custom one;
9) "verified": binary field that indicates whether the account has been verified by Twitter; verified accounts are considered to belong without doubt to humans.
10) "protected": binary field that indicates whether the account has been set as protected.

Notice that the machine learning classifier is based on static features relative to the user profile, i.e. presence of a profile picture, number of followers, etc., whereas the main results reported in our analysis and relative to the Twitter core network of endorsements are based on the dynamical interactions among agents, i.e. the actions relative to retweeting and sharing content among users. Given the short time-scale of our data collection and study (less than two weeks), the features used for determining the identity of bot users are expected not to influence the dynamical sharing of endorsements, so that the results provided in the Twitter core network and also the content of our semantic network analysis are not expected to be critically dependent on the machine learning classification outcome.

As for machine learning models, we tested a variety of algorithms readily available in the Python toolbox named *scikit-learn* [14].

In line with the considerations above, we considered the ability of the algorithms to deal with large datasets, which excluded some computationally more demanding algorithms (e.g., Support Vector Machines) and we benchmarked the following methods: Logistic Regression, Decision Trees, various ensemble methods (Random Forests, AdaBoost, ExtraTrees, etc.), *K*-nearest neighbors, Stochastic Gradient Descent, and finally two-layer neural networks.

For performance evaluation, we used two standard metrics commonly adopted in machine learning research, namely *accuracy* and *AUC-ROC* (Area Under the Receiver Operating Characteristic curve) [15]. Both scores range between zero and one, the larger the better, with one indicating perfect classification.

We set up a traditional supervised learning task, constituted of three phases, namely models' training, validation (a.k.a. performance evaluation), and finally, classification of the users in the Twitter Catalan election dataset.

The first step required us to train each model with labeled examples of the two classes of users to detect (i.e., humans and bots). To this purpose, we used two datasets containing over five thousand of positive (bots) and negative (humans) examples of Twitter users in each category. The former training dataset is associated with *Botometer* [2,7]; the latter one is a labeled dataset provided by Cresci and collaborators [16].

For performance evaluation, to calculate the accuracy and AUC-ROC scores of all models, we used the approach of *10-fold cross-validation*. This procedure splits the training data into ten equally-sized sets of data-points (preserving the balance of positive and negative data-points of the original dataset): one of these folds is hold out for validation (i.e., performance evaluation) and the remainder are used for training the models (the procedure is iterated 10 times, each holding out a different fold, and then averaging the accuracy and AUC-ROC scores obtained across the ten rounds of cross validation).

All models achieved very good performance, above 80% in both accuracy and AUC-ROC scores. The top three models in terms of performance were Random Forests (93% accuracy, 92% AUC-ROC), AdaBoost (92% accuracy and AUC-ROC), and Logistic Regression (92% accuracy, 89% AUC-ROC). The latter also was over one order of magnitude faster than nearly any other model (only Stochastic Gradient Descent was comparable in terms of speed but significantly worse in terms of performance).

Logistic Regression has the benefit of being very fast as well as very interpretable, and for such a reason was used as the reference model for the analysis discussed in the manuscript. The other two models provide results that are qualitatively identical to using Logistic Regression. When we break down the in detail the performance on the accuracy of classification of respectively human accounts versus bot accounts, Logistic Regression provides the following performance:

Class	Precision	Recall
Human	0.98	0.88
Bot	0.92	0.99

In other words, the classifier is very precise at detecting human accounts (98%), less so with bot accounts (92%), but in turn is capable of correctly detecting nearly all existing bots, as suggested by 99% recall rate, whereas it is somewhat less able to correctly recall all human users (88%).

For all the above reasons, we decided to use Logistic Regression as reference model for bot detection purposes in the rest of this study. We re-trained a full Logistic Regression model on the ten, simple metadata and activity features described above, using all the available labeled training data. Finally, we used it to classify all two million users in the Twitter Catalan election dataset. An in-depth analysis of our findings follows.

3.1 Human-bot interactions

To characterize the nature of the observed interactions between human and bots, and within the two groups, we investigate the targets of such intensive social activities. Figure S1A summarizes the structure of human-bot interactions. While humans interact mostly with other humans, 19% of overall interactions are directed from bots to humans mainly through Retweets (74%) and Mentions (25%), as shown in Fig. S1(B-D). This highlights the potential influence over human users exerted by bots and their controllers.

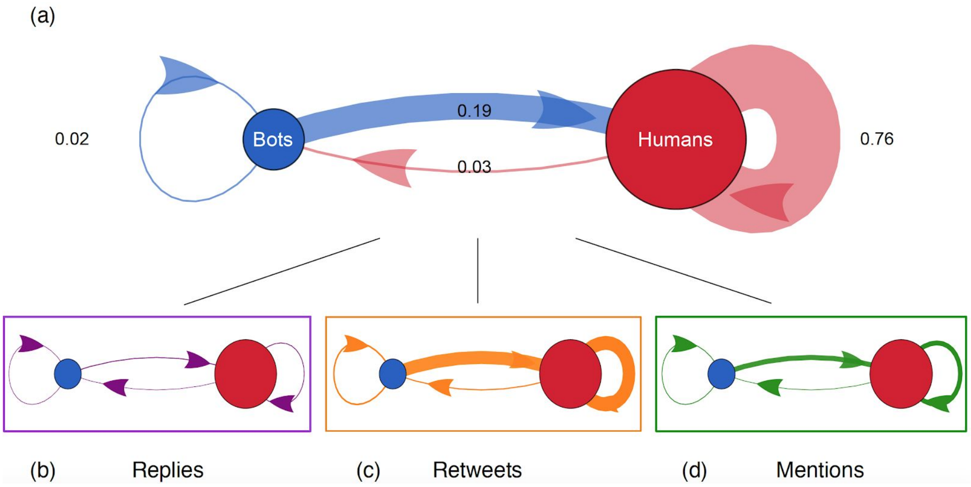


Figure S1. **Twitter interactions among humans and bots**. (a): Flowchart of human-bot Twitter interactions across the whole time window. 19% of the considered interactions are from bots to humans. (b-d): Bow-tie charts for Replies (b), Retweets (c) and Mentions (d) where edge thickness is normalized against the total volume of Tweets displayed in (a). The thicker edges in (c) indicate that most of the actions are Retweets. Retweets from bots to humans and among humans are the most frequent actions in our dataset.

4. Sentiment Analysis

Sentiment analysis aims at mapping words to the sentiment they express. Sentiment can be either categorical (e.g. negative, neutral or positive) or rather numerical (e.g. a number expressing the sentiment intensity). Lexicon-based approaches to detecting sentiment scores are based on sentiment lexicons, dictionary of emotions where words are attributed a given sentiment strength. The main advantage of lexicon-based tools for sentiment analysis is that they do not need supervised learning, i.e. training a model using labeled data. However this comes at the cost of finding large-scale lexicons of high quality.
In order to attribute a sentiment score to every tweet interaction in the considered dataset we used the VADER Sentiment package [17], a lexicon and rule-based sentiment analysis library considering also lexical features of tweets such as emoticons and acronyms (e.g. lol). VADER attributes scores of sentiment intensities between -1 (extremely negative sentiment) and +1 (extremely positive sentiment), with 0 values representing sentiment neutrality. For the English language, VADER uses a human-rated dictionary of word emotions for 7513 lexical items annotated through Amazon Mechanical Turk. Lexical features of individual words are summed up and normalised over the text length. VADER keeps into account also lexical features such as word capitalization (e.g. AMAZING), punctuation (e.g. amazing!), degree modifiers (sort of amazing), polarity shifts due to connectors (e.g. amazing but totally useless) and polarity negation (e.g. not amazing).

When tested against a human-rated dataset of 4200 annotated English tweets [17], the package correctly identified the polarization of tweets with a 3-class classification accuracy of 96%, a value higher than the average accuracy of 84% humans displayed in the same task.

In the last few years, VADER has been successfully applied for many tasks such as detecting smoking and drinking abstinence [18], identifying protest activity [20], and quantifying the effect of multiple fake accounts in online communities [21]. It is important to underline that VADER performs considerably better in short-length, domain-generic texts (e.g. tweets) rather than in longer domain-specific texts (e.g. user reviews) [21].

We chose the VADER tool for performing sentiment analysis not only for its high accuracy on random tweets but also for its potential in performing sentiment analysis across multiple languages. We preferred not to rely on automatic translations of tweets in languages different from English, as they might alter the meaning of sentences. We rather modified the VADER suite for it to natively support sentiment analysis of texts also in Spanish and Catalan. To this aim, we enriched VADER with sentiment lexica for Spanish and Catalan from the ML-SentiCon datasets [22]. The Catalan (Spanish) lexicon included 7,816 (7,377) lexical items with annotated sentiment polarities renormalized between -1 and 1. We also modified polarity negation, connectors and degree modifiers in the VADER code in order for it to fully account for lexical rules also in Catalan and Spanish.

The above enrichment of the VADER tool enabled us to compute the sentiment of English, Catalan and Spanish within one consistent framework for sentiment analysis.

5. Sentiment scores in Humans and Bots

In order to explore the sentiment dynamics within the considered dataset, we analyzed the probability distribution of finding a social interaction with a given sentiment score between -1 (extremely negative sentiment) and +1 (extremely positive sentiment). We focused on a 2-class classification of the data, so that tweets classified as neutral (i.e. with 0 sentiment score) from the sentiment analysis were not considered. Since retweets represent roughly 3 out of 4 interactions in the dataset and retweets carry a clear interpretation in terms of social endorsement, in our sentiment analysis we focused on sentiment polarity over retweets. Supplementary Figure S2 shows the distribution of sentiment scores obtained across the whole time window (from September 22 until October 3 2017).

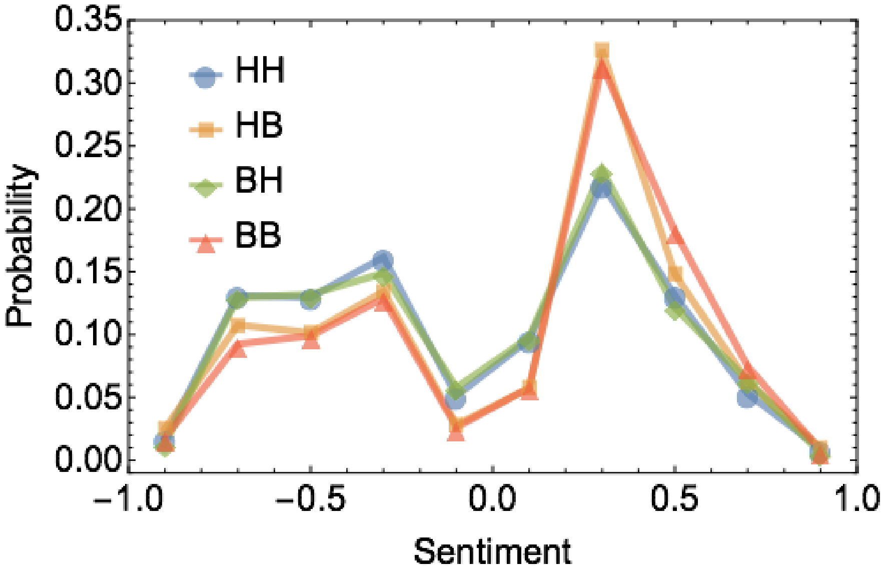


Figure S2: Above: Probability distribution of finding an interaction with a given sentiment score between -1 (extremely negative sentiment) and +1 (extremely positive sentiment) among humans (HH), from humans to bots (HB), from bots to humans (BH) and among bots (BB). Interactions directed towards humans tend to have a marked more negative sentiment compared to interactions directed towards bots.

Over the whole time window, interestingly human-directed interactions and bot-directed interactions display different patterns: human-to-human and bot-to-human interactions display a smaller degree of polarity compared to human-to-bot and bot-to-bot interactions, which display a dominant peak over positive sentiment scores. This finding suggests that bot-directed and human-directed interactions differ in their nature.

This motivated additional analysis in terms of sentiment across the human/bot groups.

7. Statistical testing of sentiment across Human/Bot Groups

In Figure 3 of the main text, we attributed positive/neutral/negative sentiment to the interactions between humans and bots across Group 1 and Group 2 according to a statistical analysis of sentiment. We considered all the interactions across two factions (e.g. interactions from humans of Group 1 to humans of Group 2). We tested the relative distribution of scores through a nonparametric testing, with the aim of detecting in which cases the median sentiment score was compatible with neutrality (i.e. sentiment score equal to 0), in which cases it was positive and in which negative. In order not to make any strong assumption on the shape of the underlying distribution of sentiment scores, we used a Sign Test. We chose a confidence level of 95%. Results for the statistical testings are reported in Supplementary Table T1.

In order to quantify potential polarization of sentiment scores, we also computed the positive score index I^+ and the negative score index I^-. The positive (negative) score index is the average value E^+ of all interactions with positive (negative) sentiment multiplied by the frequency f^+ of positive (negative) interactions relative to the total number of interactions. In formulas: $I^+ = E^+ f^+$. While the average value E^+ expresses the magnitude of positiveness of the considered sentiment scores it does not consider the frequency of positive interactions. This motivated our definition for the index I^+.

Interactions	**Positive Scores**	**Negative Scores**	**Median Sentiment**	**p-value**	**Outcome**
H1 to H1	0.096	-0.161	-0.02	0.0003	Negative
H1 to H2	0.106	-0.089	0.05	0.0008	Positive
H1 to B1	0.156	-0.009	0.08	>0.05	Neutral
H1 to B2	/	/	/	/	/
H2 to H1	0.098	-0.132	0.002	>0.05	Neutral
H2 to H2	0.115	-0.083	0.069	10^{-5}	Positive
H2 to B1	/	/	/	/	/
H2 to B2	0.237	-0.005	0.212	>0.05	Neutral
B1 to H1	0.091	-0.202	-0.089	10^{-5}	Negative
B1 to H2	0.117	-0.091	0.069	>0.05	Neutral
B1 to B1	0.099	-0.195	-0.186	0.0001	Negative
B1 to B2	/	/	/	/	/
B2 to H1	0.052	-0.136	-0.121	0.035	Negative
B2 to H2	0.135	-0.075	0.112	10^{-5}	Positive
B2 to B1	/	/	/	/	/
B2 to B2	0.178	-0.003	0.196	0.0005	Positive

Supplementary Table T1: Probability distribution of finding an interaction with a given sentiment score between -1 (negative sentiment) and +1 (positive sentiment) among humans (HH), from humans to bots (HB), from bots to humans (BH) and among bots (BB). Interactions directed towards humans tend to have a marked more negative sentiment compared to interactions directed towards bots.

8. Network clustering methodology

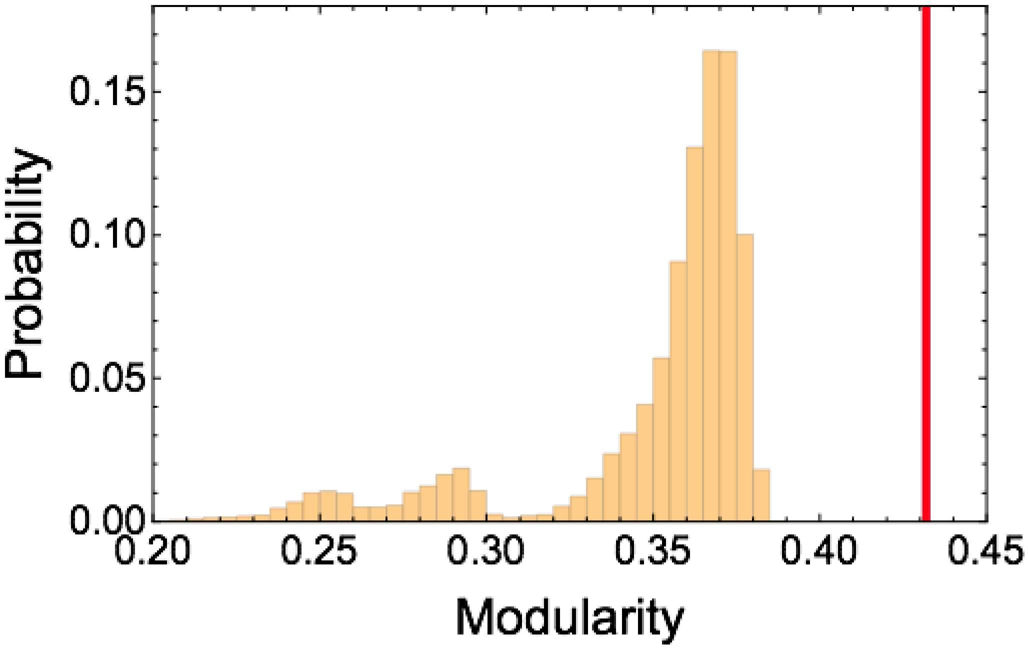
A minimum cut of a graph corresponds to the maximum graph modularity when nodes are partitioned in two clusters. Modularity is a widely used technique in community detection for quantifying how distinct a given graph partitioning is compared to a random allocation of links across partitions [23].

Notice that Fiedler partitioning is a valid heuristic for solving the min-cut problem on unweighted, undirected graphs while the TCN is a directed network. Hence, in order to test the distinctness of the Fiedler partitioning, we developed the following methodology:

1. Compute the Fiedler spectral graph partitioning on an undirected version of the TCN, by ignoring edge directionality.
2. Compute the modularity M_0 of the resulting partitioning on the original TCN (i.e. with directed links) through the *Mathematica* software.
3. Iterate 20000 times the following:
    a. Reshuffle 5% of nodes across the original groups, i.e. select at random 5% of nodes from Group 1 and 5% of nodes from Group 2 and switch their partitioning labels (from Group 1 to Group 2 and vice versa). The reshuffling creates a perturbed partitioning of the original Fiedler partition.
    b. Compute and store the modularity of the perturbed partition on the directed TCN.
4. With the obtained ensemble of perturbed modularities on the directed TCN, perform a statistical testing for quantifying the significance of the observed modularity M_0 for the TCN.

As evident from Supplementary Figure S3, there is a large gap between the perturbed partitions and the observed one for the TCN. The p-value of observing the modularity M_0 from the ensemble of perturbed modularity is smaller than 10^{-5}.

This indicates an important testing for the Fiedler partitioning: even though the Fiedler partition was obtained on the undirected version of the TCN, the relative modularity is still optimal compared to small random modifications of the partitioning.



Supplementary Figure S3: Empirical probability distribution of modularities for the same empirical network of directed Twitter interactions but with perturbed partitions of users into two groups. The red thick line indicates the modularity of the empirical network with the partitioning into groups coming from the algebraic connectivity (modularity *M=0.43*). The empirical distribution of modularities has been obtained over 20000 perturbed partitions.

9. Global comparison of hashtag co-occurrence networks through consistency

In order to compare the nature of social interactions in Group 1 and Group 2, we perform a consistency analysis, quantifying how similarly the same hashtag is associated across the two groups. In semantic network analysis, consistency measures how consistently a given word is used in semantic network of co-occurrences [24]. Consistency is defined as a the cosine similarity of the neighbourhoods n_1 and n_2 for word w_i in networks $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, in formulas:

$$c_i = \frac{\sum_j A_{ij}^{(1)} A_{ij}^{(2)}}{\sqrt{k_i^{(1)} k_i^{(2)}}}$$

A word w_i co-occurring with the same set of words in both the co-occurrence networks is fully consistent ($c_i = 1$). A word co-occurring with two completely disjoint sets of words in the two co-occurrence networks has zero consistence. Consistency is a local measure, quantifying how co-occurrence networks coincide at the local scale of neighbourhoods.

The co-occurrence networks of hashtags across the two factions display an average consistency of $c = 0.65 \pm 0.29$, where the average is performed across all hashtags

appearing in both the networks. As a comparison, null models with the same total number of co-occurrences across two factions but randomised network edges for each faction displays an average consistency of $c_{ran} = 0.03 \pm 0.01$. This indicates that the observed consistency is not to be attributed to the way we built the co-occurrence networks but it rather encapsulates meaningful correlations among hashtags. Also, the large standard deviation on the empirical consistency across factions indicates that there are large fractions of neighbourhoods displaying extreme values of consistency, either $c_i = 0$ or $c_i = 1$. In particular, the value $c_i = 0$ ($c_i = 1$) indicates the extreme case in which the same hashtag is associated in a totally different (equivalent) way by the two factions. Across the two human factions, the likelihood of having hashtags with $c_i = 0$ ($c_i = 1$) is $p_{c=0} = 0.132$ ($p_{c=1} = 0.237$).
These results suggest that even though the two factions display an overlap in the associated hashtags, there seems to be a relatively large variability in the structure of associations of these overlapping hashtags across factions. This motivates the usage of the structure of hashtag co-occurrence for the identification of factions.

10. Centrality analysis of hashtag co-occurrence networks identifies identities of groups

The consistency analysis finds that users from Group 1 and Group 2 associate differently roughly 35% of the associated hashtags, with large variability in terms of consistency of individual concepts.

In order to focus on the most evident differences in the way individual concepts are associated from the two groups, we used another approach, focusing on the centrality of lexical items. From the relevant literature of semantic networks, it has been recently shown that the most prominent concepts in networks of associations are those with high degree, i.e. with a larger number of associations, high strength, i.e. with stronger associations, and with high closeness centrality, i.e. at a shorter average distance from all other concepts. Words of higher degree in networks of associations among concepts have been found to be better predictors of cognitive patterns both at the level of individuals, e.g. in characterising the individual level of creativity [25], and also at the population level, e.g. in characterising early word acquisition in children [26] or cognitive degradation in patients with the Alzheimer's Disease [27] or aphasia [26]. We exploit this idea that the most prominent lexical items in our network of hashtags associations are those with the highest centrality, be it degree or strength or closeness.

We then rank hashtags according to their centrality in both the networks. We then consider those concepts being the most central in one network and the most peripheric in the other one, i.e. maximising the difference in absolute value of the ranks in each network when a given centrality is chosen. Figure 5(A) from the main text reports the most extreme examples: words that are the most central in one co-occurrence network and the least central in the other co-occurrence network. Differences in the ways the same hashtag is collectively associated by all users in one group are evident, quantitatively showing the different nature of the two groups found.

Notice that the co-occurrence analysis encapsulates more information than a simpler investigation of frequency of hashtags. As it happens also in semantic networks [26], degree positively correlates with hashtag frequency (Kendall Tau $\kappa = 0.44$ p-value$<10^{-3}$, for Group 1 and $\kappa = 0.43$ p-value$<10^{-3}$ for Group 2) so that the network structure of co-occurrences encapsulates also information of hashtag frequency in tweets.

11. Human coding of tweets

In order to explore more thoroughly the quantitative findings from our computational analyses of tweets in terms of sentiment and network structure, we also performed human coding of a total of 2,412 tweets. Human coding of social media news is a widely used technique in the computational social sciences for complementing sentiment analysis with a more thorough analysis of content beyond sentiment polarity and more focused towards the semantic content of text (see [28] for a review).

Human coding was performed by the article authors over the tweets of three categories of users: (i) human hubs in the social bulk (defined as being involved in more than 100 endorsements in the Twitter core network), (ii) human common users in the social bulk (defined as those users not being hubs in the core network), and (iii) bot accounts. In order to minimise selection biases, all the human hubs were included in the analysis, while human common users and bot accounts were sampled uniformly at random.

Coding focused in particular over:
(i) the tone of the exposition of individual tweets (e.g. personal/impersonal);
(ii) the main topic of the tweet (politics, police, ideology, logistics of voting, etc.);
(iii) the main stance of the tweet (supportive/critical);
(iv) the type of language used (e.g. associations among hashtag words/usage of emoticons/idiom used);
(v) the time evolution of topics and stances over time.

The analysis revealed the following patterns:

1) Human hubs predominantly used a personal style of exposition, mainly using the second person plural for reporting simple sentences (e.g. “For us there is no alternative to voting”), a tone creating a sense of involvement and community. The observed tweets are in Catalan and in English. Topics evolve over time: Initially most

of the tweets are critical against terrorism and revolve around the topic of repression, mostly inciting towards the freedom of vote for defending civil rights. On the eve of the referendum, most of the observed tweets conveyed a supportive stance towards the right of voting, including also requests for protecting the urns and highlighting the importance of pacifist protests. The morning of the referendum starts with very positive tweets (e.g. “Good morning, Catalonia! Ready to vote?” or “”) but the trending topic quickly transitions to the violence of the Police (“Shocking scenes in the morning”, “Brutal represion contra personas indefensas”, “761 heridos”). Starting from the referendum day, the type of language used includes also swearing and “police” becomes quickly associated with “brutality” in most of the observed tweets. A small fraction of the content endorsed by human hubs is relative to newspaper titles reported in an impersonal form and supportive of the referendum.

2) Common users follow patterns similar to those of hub users, with concepts such as “manipulation”, “information” and “concern” becoming more frequent during the referendum day. Also common users tend to share messages in second person plural but their content tends to be more personal than the one of human hubs (e.g. “This is probably one of the most emotional moments of my life.”). Also common human users tended to complain about the violent happenings of the referendum day and tended to oppose the vision of “pacific citizens” to the one of the oppressing and “violent Police”. Both hubs and common human users tended to frequently use emoticons in their tweets, with emoticons expressing concern, negative surprise and disbelief as being the most frequently occurring ones.
3) Bot users tend to expose their content mainly in impersonal ways, using mostly third person and passive forms. Their constructs are less elaborate than humans, with the subject of the statement usually in the first position of the sentence. To draw a parallel, the simplicity and impersonality of these tweets make them very similar to newspaper titles. The topics promoted by automated accounts are mainly related to “repression”, “coup”, “tension” and “war”. Tweets coming from bots are multi-language, with English, Italian, Catalan, Spanish and German being the most represented languages. Emoticons are absent and a constant sense of tension and conflict pervades the whole storyline of these tweets, differently from the initially optimistic content of human tweets.

The human analysis of tweets confirms the trends of sentiment quantified in the main paper for human-to-human interactions: Humans start particularly optimistic and then share increasingly more pessimistic and negative content after the onset of the referendum day, as a reaction to the violent acts registered in those hours. Bots tweets reflect the average sentiment of neutrality of bots-to-bots interactions, with some exceptions. The above analysis indicates that bots mainly promoted newsmedia titles, mimicking the trend of human emotions and hence boosting the sentiments of fear, disgust and reprobation rising after the voting day.

12. Influence of social bots over human users

The influence played by bots during the considered voting event can be analysed in terms of (i) the targets of bot interactions and (ii) the volume of such interactions. From the in-degree correlations reported in the main text, we know that bots mainly targeted human hubs. If we define as hubs those nodes in the social bulk network having at least more than 100 in-going endorsements, then we have 28 hubs that might potentially retweet automatic content from bots. Further analysis indicates that bots tend to be largely ignored by human hubs: only one hub out of 28 retweets content generated by bots.

However, a closer look indicates that bots tend to be retweeted mainly by human common users, especially in Group 1. In Figure 3 (B) of the manuscript's main text we report the volumes of endorsement exchanges between humans and bots in the Twitter core network when all considered accounts are partitioned in groups 1 and 2:

- 44% of the endorsements are shared among humans in Group 2;
- 34% of the endorsements are shared among humans in Group 1;
- 10.5% of the endorsements go from bots to humans in Group 1;
- 7.4% of the endorsements go from humans to bots in Group 1;
- 5.8% of the endorsements go from humans to bots in Group2;
- 0.1% of the endorsements go from bots to humans in Group 2.

Compared to Group 2, Group 1 has a large volume of endorsements flowing from human users to bots, indicating that bots are retweeted by common users and hence taken into account.

The picture emerging from the above statistics is that:

1) Bots tend to promote the same content from human hubs and in Group 1 this content has a general negative sentiment polarity;
2) In this way, even though bots are not retweeted by hubs, they can still promote and boost specific semantic content which is read by common users;
3) Common users tend to re-share and retweet the human content shared by bots.

Hence, bots act as influential nodes in the sense that they promote human-generated content (rather than automated tweets) "borrowed" from hubs and taken into consideration by non-trivial fractions of human users (as suggested by the percentages of endorsements in Group 1).

References

1. Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. "The rise of social bots." Communications of the ACM 59, no. 7 (2016): 96-104.
2. Varol, Onur, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. "Online human-bot interactions: Detection, estimation, and characterization." In Eleventh International Conference on Weblogs and Social Media. 2014 AAAI (2017).
3. Messias, Johnnatan, Lucas Schmidt, Ricardo Augusto Rabelo de Oliveira, and Fabrício Rodrigues Benevenuto. "You followed my bot! Transforming robots into influential users in Twitter." (2013).
4. Yang, Zhi, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. "Uncovering social network sybils in the wild." ACM Transactions on Knowledge Discovery from Data (TKDD) 8, no. 1 (2014): 2.
5. Gilani, Zafar, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. "Stweeler: A framework for twitter bot analysis." In Proceedings of the 25th International Conference Companion on World Wide Web, pp. 37-38. International World Wide Web Conferences Steering Committee, 2016.
6. Subrahmanian, V. S., Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. "The DARPA Twitter bot challenge." Computer 49, no. 6 (2016): 38-46.
7. Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. "Botornot: A system to evaluate social bots." In Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273-274. International World Wide Web Conferences Steering Committee, 2016.
8. Bessi, Alessandro, and Emilio Ferrara. "Social bots distort the 2016 US Presidential election online discussion." (2016).
9. Aral, Sinan, and Dylan Walker. "Identifying influential and susceptible members of social networks." Science (2012): 1215842.
10. Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." Proceedings of the National Academy of Sciences 111, no. 24 (2014): 8788-8790.
11. Ferrara, Emilio, and Zeyao Yang. "Measuring emotional contagion in social media." PloS one 10, no. 11 (2015): e0142390.
12. Ferrara, Emilio, and Zeyao Yang. "Quantifying the effect of sentiment on information diffusion in social media." PeerJ Computer Science 1 (2015): e26.
13. Mønsted, Bjarke, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. "Evidence of complex contagion of information in social media: An experiment using Twitter bots." PloS one 12, no. 9 (2017): e0184148.
14. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12, no. Oct (2011): 2825-2830.

15. Christopher, M. Bishop. PATTERN RECOGNITION AND MACHINE LEARNING. Springer-Verlag New York, 2016.
16. Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 963-972. International World Wide Web Conferences Steering Committee, 2017.
17. Hutto, CJ and Gilbert, Eric. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In Eighth International Conference on Weblogs and Social Media. (2014)
18. Tamersoy, Acar, Munmun De Choudhury, and Duen Horng Chau. "Characterizing smoking and drinking abstinence from social media." In Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 139-148. ACM, 2015.
19. Won, Donghyeon, Zachary C. Steinert-Threlkeld, and Jungseock Joo. "Protest Activity Detection and Perceived Violence Estimation from Social Media Images." In Proceedings of the 2017 ACM on Multimedia Conference, pp. 786-794. ACM, 2017.
20. Kumar, Srijan, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. "An army of me: Sockpuppets in online discussion communities." In Proceedings of the 26th International Conference on World Wide Web, pp. 857-866. International World Wide Web Conferences Steering Committee, 2017.
21. Ribeiro, Filipe N., Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. "SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods." EPJ Data Science 5, 1 (2016): 23.
22. Cruz, Fermín L., José A. Troyano, Beatriz Pontes, and F. Javier Ortega. "Building layered, multilingual sentiment lexicons at synset and lemma levels." Expert Systems with Applications 41, no. 13 (2014): 5984-5994.
23. Newman, Mark. Networks: an introduction. Oxford university press, 2010.
24. Amancio, Diego R., Osvaldo N. Oliveira Jr, and Luciano da F. Costa. "Unveiling the relationship between complex networks metrics and word senses." EPL (Europhysics Letters) 98, no. 1 (2012): 18002.
25. Kenett, Yoed N., David Anaki, and Miriam Faust. "Investigating the structure of semantic networks in low and high creative persons." Frontiers in Human Neuroscience 8 (2014): 407.
26. Stella, Massimo, Nicole M. Beckage, and Markus Brede. "Multiplex lexical networks reveal patterns in early word acquisition in children." Scientific Reports 7 (2017): 46730.
27. Borge-Holthoefer, Javier, and Alex Arenas. "Semantic networks: Structure and dynamics." Entropy 12, no. 5 (2010): 1264-1302.
28. Batrinca, Bogdan, and Philip C. Treleaven. "Social media analytics: a survey of techniques, tools and platforms." Ai & Society 30, no. 1 (2015): 89-116.