Supplementary Material

1 MORE INFORMATION ON BOTS FILTERING

The presence of Bots in our dataset may alter the results of our analyses. Usually bots tend to produce a large number of tweets, mainly focused on one or few topics, so not filtering them could lead to large errors on the average similarity between users.

To assure that our dataset only contains a negligible fraction of Bots we implemented three different filtering strategies. First of all, as bots content tends to be dominated by retweets (i.e. see (1)) we decided to exclude all the retweets from our dataset and all the users with a very high percentage of retweets in their timeline. This decision has also been motivated by the fact that, in our analysis, we are more interested in contents produced by the users than in shared information.

To further reduce the probability of finding a bot in our dataset we also removed users with an unfeasible daily rate of tweets. S2 shows, for each user in our dataset, the average number of produced tweets per day in relation to the time they have been active; counted as the time difference between the first and the last tweet in our dataset. As it is clear from the figure, there is a high peak in the activity of users with a short active time suggesting that they produced a high number of tweets and then disappeared. This is another typical signature of bots activity so we decided to filter all the users that have been active for one day or less and those who produced, on average, more than 400 tweets per day. This left 9490 central users and a total population of 608899 over the initial 774596 that used at least one hashtag included in the topics we extracted. The distribution of tweets per user is very heterogeneous, as it it can be seen on Fig. of this population can been in Fig. S1.

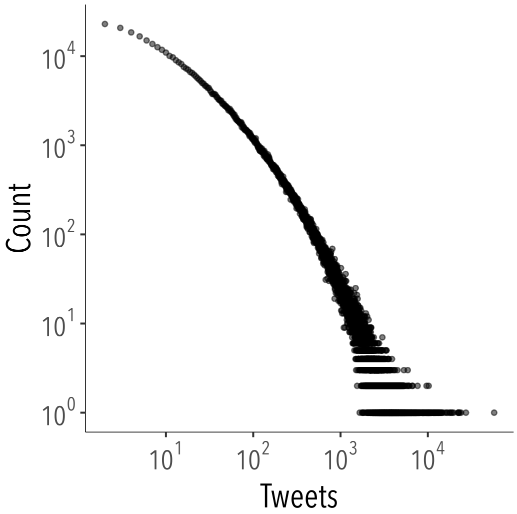

Figure S1. Distribution of tweets per user from the *Population* set.

Finally, we checked the quality of our filtering using the tool developed by Menzcer *et al.* – i.e. (1)–, Botomoter[1]. The tool, starting from a user ID, retrieves its timeline and returns a score between 0 and 1 that the user is a bot. A score of 1 means that almost certainly the user is a bot while 0 stands for a human

[1] https://botometer.iuni.iu.edu/

user. Due to rate limits in the number of users that can be tested per day we decided to check only the central users in our dataset. S3 shows the distribution of scores of the central users with still an active account at the time of the analysis (6539 over the original 9490) while S4 presents the number of users with a score greater or smaller than 0.6. From these results, it is clear that the influence of bots on our analysis is minimal. The vast majority of the users in our dataset have a score lower than 0.25 and only 222 over 6539 have a score higher than 0.6.

To further demonstrate the robustness of our analyses, we also computed the average similarities such as that on Fig 3 of the main text only considering central users whose score is smaller than 0.5 (S5). Comparing the original figure with the new one, it is clear that, even using this conservative threshold, the possible influence of bots in our results is insignificant.

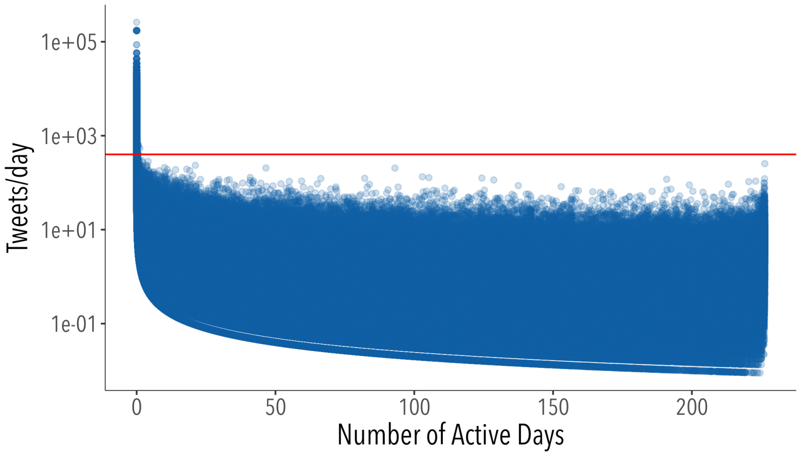

Figure S2. Average number of tweets per day generated by each user as function of the active time (in days) calculated as the time difference between the first and last tweet in our dataset.

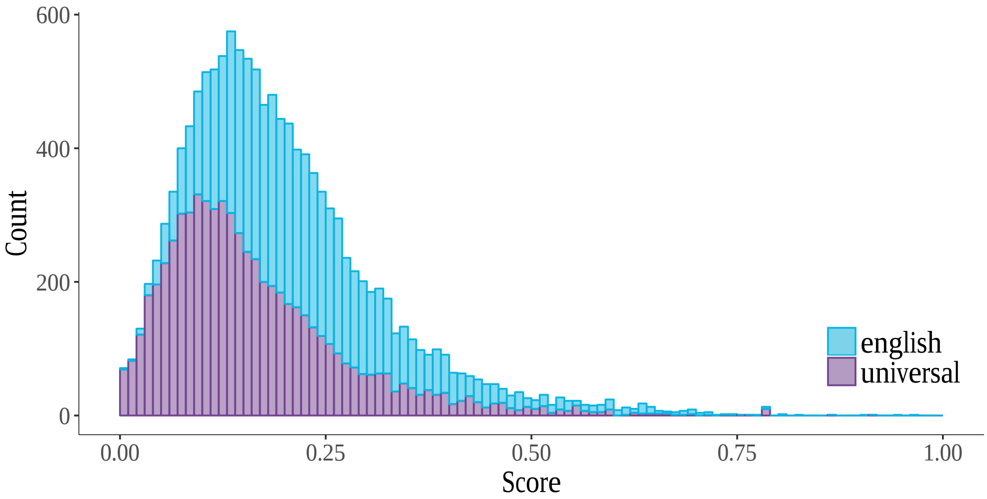

Figure S3. Bots score distribution for the central users of our dataset as given by the Botometer tool (https://botometer.iuni.iu.edu/).

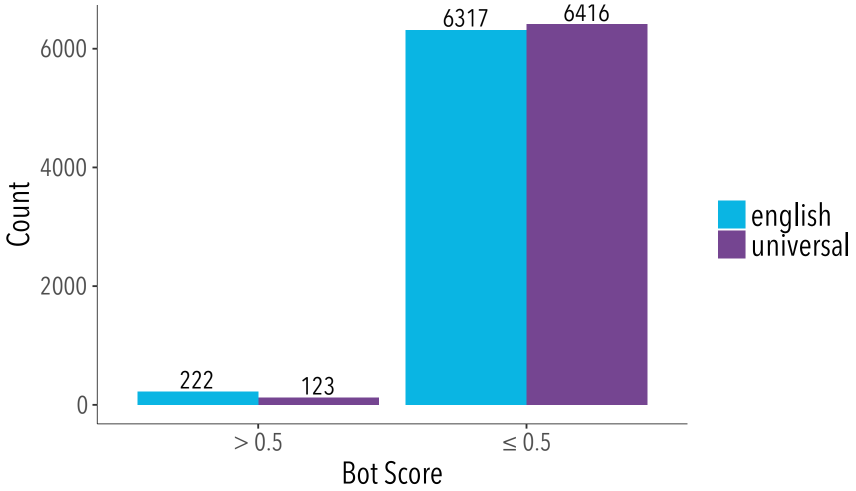

Figure S4. Number of central users classified as bots (score $>$ 0.6) by the Botometer tool(https://botometer.iuni.iu.edu/).

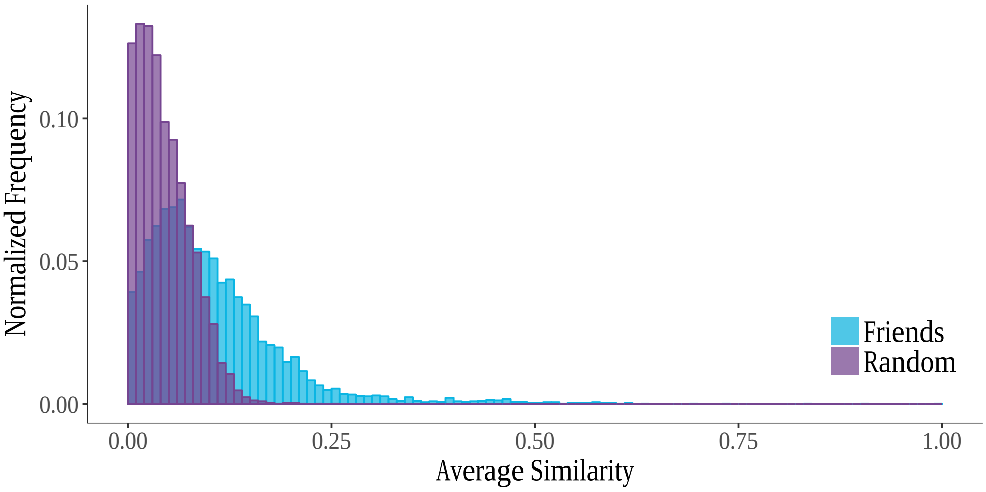

Figure S5. Distribution of average similarity between central users and their followees(purple) and between central users and randomly selected groups of the same size (yellow). Distributions have been calculated considering the whole set of 9,490 central users.

2 TOPICAL SIMILARITY VERSUS HASHTAGS SIMILARITY

One of the main innovations of our work is that we use topics instead of hashtags to calculate similarity between users. At this point one can argue what are the advantages of using topics instead of hashtags, as extracting topics from hashtags co-occurence network is a costly process. A similar approach could be considering directly vectors that describe hashtags usage instead of topics. This method, however, disregard dyads wherein users do not use the same hashtags but are interested in the same issues. To test if our method gives better results than only using hashtags we repeated the analysis in Fig 3 of the main text calculating the average similarity also using hashtags vectors. As demonstrated by S6, the distribution of average similarity calculated using hashtags is more peaked and centered at lower values. This is due, as shown in the inset of S6, to the presence of a significant amount of dyads with a low similarity. This means that most connections use the same topic but not the same hashtags. Thus, we believe that using topics to detect similarity is more robust and allows to uncover relationships that would go unnoticed using hashtags.

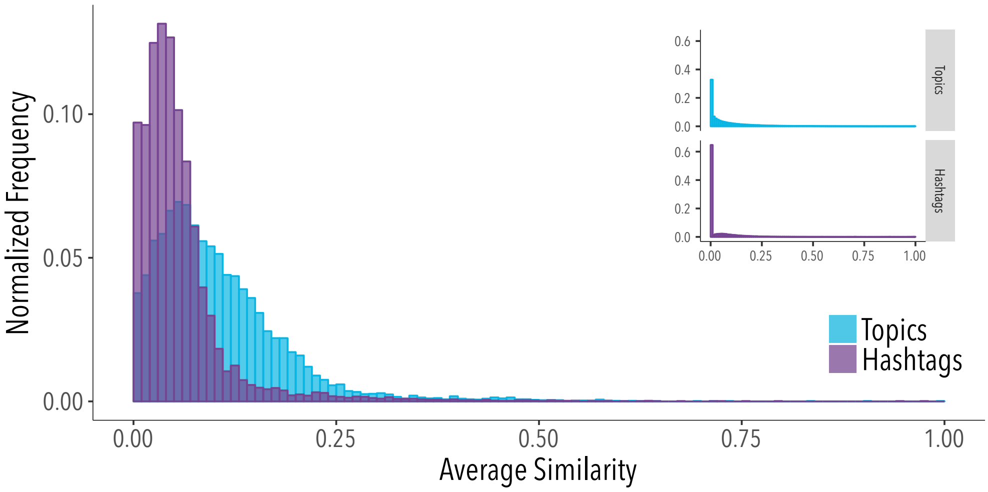

Figure S6. Distribution of average similarities of central users with their followees computed considering topics and hashtags. The inset shows the distribution of similarities of all the dyads.

3 THE ROLE OF HOMELESS NODES IN TOPICS DETECTION

The extraction of topics based on community detection highlighted a high number (14118) of "homeless nodes" – hashtags that do not belong to any community, so the algorithm creates a new community with only one node –. Given their high number with respect to larger communities, it could be important to asses how those topics affect our results. Our intuition is that homeless hashtags are usually typos or ambiguous words not so common as hashtags. Thus, they should appear in few tweets and be employed by few users. To verify this hypothesis, Fig. S7 presents the number of homeless hashtags used by distinct users. As expected, more than half of the homeless hashtags have been used by only one user and almost the totality by no more than five. This supports our intuition that those hashtags play a minimal role in our results as topics used by only one user are not considered in the similarity.

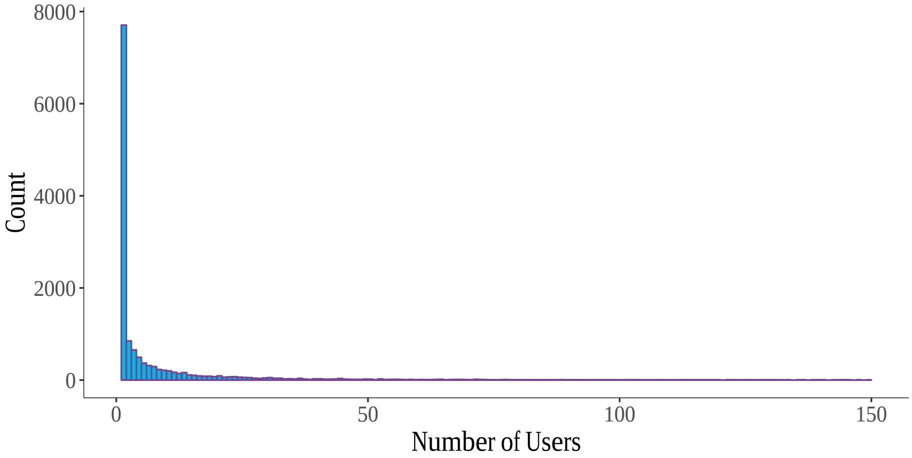

Figure S7. Number of the homeless hashtags used by distinct users.

4 A DESCRIPTION OF TOPICS

The majority of topics considered in this project contains few hashtags, as it can be seen in Fig. S8. The average number of hashtags without the homeless nodes is of 45.6 and of 6.7 taking the homeless nodes into account. Communities overlap and, on average, each hashtag belong to 1.04 communities. Communities are also quite heterogeneous with respect to their number of users, with an average of 622 users per topic.

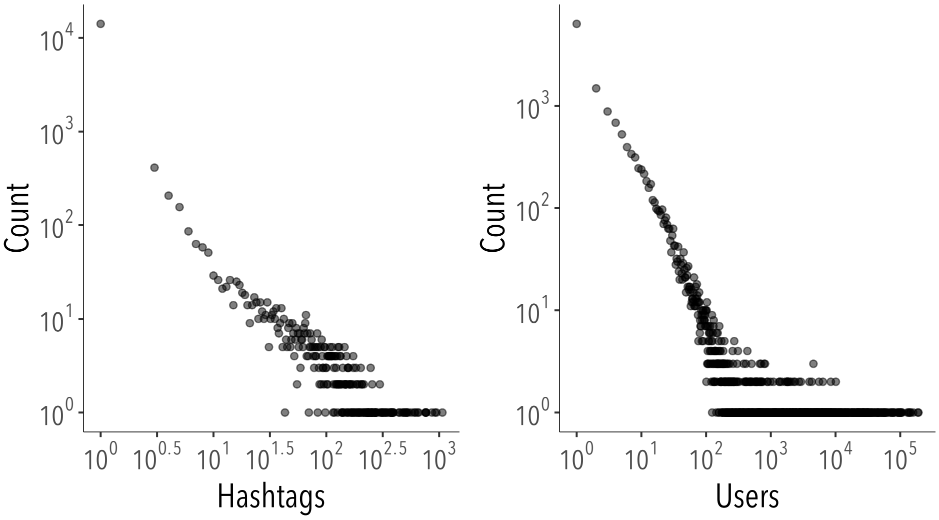

Figure S8. Distribution of hashtags (left) and users (right) of the topics.

REFERENCES

[1]Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The Rise of Social Bots. Commun ACM. 2016;59(7):96–104. doi:10.1145/2818717.