scientific reports

Check for updates

OPEN

Opinion manipulation on Farsi Twitter

Amirhossein Farzam [1,2,3✉], **Parham Moradi** [4,5], **Saeedeh Mohammadi** [4,5,6], **Zahra Padar** [7] & **Alexandra A. Siegel** [8✉]

For Iranians and the Iranian diaspora, the Farsi Twittersphere provides an important alternative to state media and an outlet for political discourse. But this understudied online space has become an opinion manipulation battleground, with diverse actors using inauthentic accounts to advance their goals and shape online narratives. Examining trending discussions crossing social cleavages in Iran, we explore how the dynamics of opinion manipulation differ across diverse issue areas. Our analysis suggests that opinion manipulation by inauthentic accounts is more prevalent in divisive political discussions than non-divisive or apolitical discussions. We show how Twitter's network structures help to reinforce the content propagated by clusters of inauthentic accounts in divisive political discussions. Analyzing both the content and structure of online discussions in the Iranian Twittersphere, this work contributes to a growing body of literature exploring the dynamics of online opinion manipulation, while improving our understanding of how information is controlled in the digital age.

The Farsi Twittersphere was initially hailed as an early example of "liberation technology" for its role in the 2009 anti-regime protests in Iran[1]. In the absence of independent media, Farsi Twitter emerged as a public space where uncensored opinions proliferated[2]. But over the past decade, the Farsi Twittersphere has become an opinion manipulation battleground. Diverse actors regularly employ bots, sock-puppets, and other inauthentic accounts to advance their goals and shape online narratives[3]. Examining over one million tweets collected between June 2019 and January 2020, we explore the behavior of inauthentic accounts across trending online discussions spanning diverse social and political issues in the Farsi Twittersphere.

The influence of coordinated inauthentic activity in online political debates is a topic of growing interest among academics and policymakers alike. An emerging body of research has highlighted the considerable presence of inauthentic accounts working to shape debates around elections, political protests, military operations, and other political events in diverse contexts[4–14]. Such accounts include bots, cyborgs, sock-puppets, and trolls (e.g.[4,7,9,13]). The majority of this work explores the use of inauthentic activity by foreign actors to influence politics abroad, with Russian interference in US elections receiving the most scholarly attention. The use of inauthentic accounts by domestic actors to influence domestic politics is particularly understudied (though see[15] for an exception).

We assess the relative prevalence of inauthentic activity in eleven trending Twitter discussion topics crossing the social cleavages of the Iranian society. Specifically, we introduce a typology of authentic and inauthentic Twitter users and examine how different types of users participate in trending discussions over time. To classify users according to our typology, we use Botometer[16–18], a supervised machine-learning algorithm, to measure the Complete Automation Probability (CAP) for the users in our Twitter data. While Botometer does not accurately classify individual accounts, extensive human coding and validation using Farsi Twitter accounts suggests that average CAP scores can be used to assess the average relative prevalence of inauthentic activity across communities or groups of Twitter users.

After developing these relative measures of inauthenticity, we analyze friendship and retweet networks to understand how clusters of inauthentic accounts interact with diverse communities of Iranian Twitter users across the political spectrum. Exploring the network structures of Twitter users across these trending discussions, we find that inauthentic activity is concentrated in divisive political discussions, where retweet and friendship

[1]Present address: Department of Political Science, Duke University, Durham, NC 27708, USA. [2]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208, USA. [3]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany. [4]Center for Complex Networks and Social Data Science, Tehran, Iran. [5]Department of Physics, Shahid Beheshti University, Tehran, Iran. [6]Present address: School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. [7]Department of Computer Science, University of Freiburg, Freiburg, Germany. [8]Department of Political Science, University of Colorado, Boulder, CO 80309, USA. ✉email: a.farzam@duke.edu; alexandra.siegel@colorado.edu

networks are ideologically polarized. By contrast, apolitical discussions have very little inauthentic activity, and these networks are characterized by cross cutting interactions across the political spectrum.

Finally, we compare the content produced by relatively inauthentic groups of accounts to more genuine clusters of accounts across different trending topics. We find that clusters of inauthentic users in divisive political discussions use highly polarizing keywords, targeting particular ideological audiences, relative to clusters of authentic users. Characterizing both the content and structure of online discussions in the Iranian Twittersphere across diverse issue areas, this work contributes to a growing body of literature exploring the dynamics of online opinion manipulation.

Results

Prevalence of inauthentic accounts.

To assess the role of inauthentic accounts in trending topics in the Farsi Twittersphere, we collected data in real time, surrounding 11 trending topics—four apolitical discussions, four non-divisive political discussions, and three divisive political discussions between June 2019 and January 2020. Developing a typology of Twitter users with respect to the level of automation and inauthenticity of the account, we examine the average relative prevalence of different types of accounts in each trending discussion. We use the CAP scores from Botometer as a relative measure of the level of inauthenticity of the accounts. The details of how CAP scores are utilized for this purpose are explained in the "User typology" subsection. As Fig. 1 displays, we see that most influential accounts—the top decile of most retweeted accounts—in divisive political discussions, on average tend to have higher levels of inauthentic activity (higher CAP scores) than non divisive political discussions and apolitical discussions. Performing a two-sided two-sample Kolmogorov-Smirnov test[19] on the average CAP scores for apolitical and divisive political discussions in Fig. 1, we see higher average CAP scores in divisive political discussions ($p < 0.001$). Comparisons using other measures of influence in the retweet network—retweet PageRank and retweet h-index—yields a similar result. This is demonstrated by the violin plots in Fig. 1, which show the CAP distributions in divisive political and apolitical discussions for the top decile of users by each influence measure. Note that while the Kolmogorov-Smirnov test indicates difference in the distributions at the significance level $p < 0.001$ for retweet count and PageRank, this difference is not statistically significant for the retweet h-index. Nevertheless, this suggests that the participants in divisive political discussions come from a population that contains more inauthentic activity than participants in apolitical discussions, particularly among users that tend to be more influential in the retweet network with respect to PageRank centrality and retweet count.

Beyond our relative measure of CAP scores, the deactivation or suspension of accounts shortly after their participation in a trending discussion provides further evidence of inauthentic behavior. Inauthentic accounts are much more likely to be suspended by Twitter or deactivated for suspicious activities than genuine Twitter users. We see higher levels of deactivation or suspension of accounts shortly after participation in divisive political discussions than apolitical discussions. Specifically, we find that the odds ratio of suspension or deactivation of accounts in divisive political discussions to apolitical discussions is approximately 1.06 (with χ^2 statistic > 4 and $p < 0.05$), which indicates that participants in divisive political discussions are more likely to be deactivated or suspended after participating in the discussion. Performing the same analysis on only the top 10% most retweeted accounts, we find that the odds ratio of suspension or deactivation in divisive political discussions to apolitical discussions is even larger, at 1.94 (with χ^2 statistic > 19 and $p < 0.001$). Similar patterns are observed
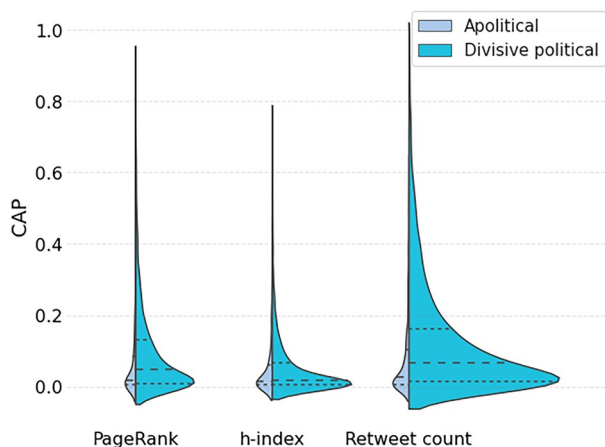

Figure 1. The distributions of CAP scores among the most influential accounts, by PageRank (left), h-index (middle), and retweet count (right), in the retweet network of the Apolitical and Divisive Political discussions. The violin plots, and the quartile lines within each violin plot, show the distribution of CAP scores for users among the top decile of accounts with respect to the corresponding measure, in each topic within each category of discussion. The two-sided Kolmogorov–Smirnov test reveals statistically significant difference between each pair of distributions at significance level $p < 0.001$ for PageRank and retweet count, but this difference is not statistically significant for h-index. The left half of the violin plot (Apolitical) contains 790 CAP scores and the right half (Divisive Political) contains 7385 CAP scores.

for suspension or deactivation among users in the top decile by retweet PageRank and retweet h-index, with odds ratios of deactivation in divisive political to apolitical discussions 1.71 ($p < 0.05$) and 1.47 ($p < 0.05$), respectively. Figure 2 shows these odds ratios, as well as enrichment-depletion patterns for users grouped by activation status and user types with respect to our inauthenticity typology. This indicates that there is even a stronger evidence of inauthentic activities in divisive political discussions if we focus on the most influential accounts in the discussion.

Participation in trending discussions. In addition to prevalence, Twitter data enables us to measure how inauthentic accounts behave during diverse types of trending topics. Discussions on Twitter are bursty, gaining traction, peaking, and re-equilibrating as users move on to the next topic or event. Typically, first movers begin an online conversation, a trending topic peaks within a few days, after which engagement slows and the topic stops trending. Drawing on our typology of authentic and inauthentic accounts, we can examine when different types of accounts entered the discussion of each topic, as well as how they behaved throughout the trending period. We construct the distribution of these patterns using time-series data of the tweets partitioned by user types. Using the CAP scores from Botometer, we group users into three types—automated and semi-automated bots (group A), bot-assisted humans and trolls (group B), and genuine users (group C)— falling within mutually exclusive CAP intervals, in decreasing order from group A to group C. Our manual exploration of the data guides this large-scale systematic analysis of the temporal patterns and the user type configuration of the networks corresponding to each discussion. This includes the manual annotation of over 650 accounts with respect to features that characterize each type in our typology, as well as inspection of most influential (i.e. retweeted)
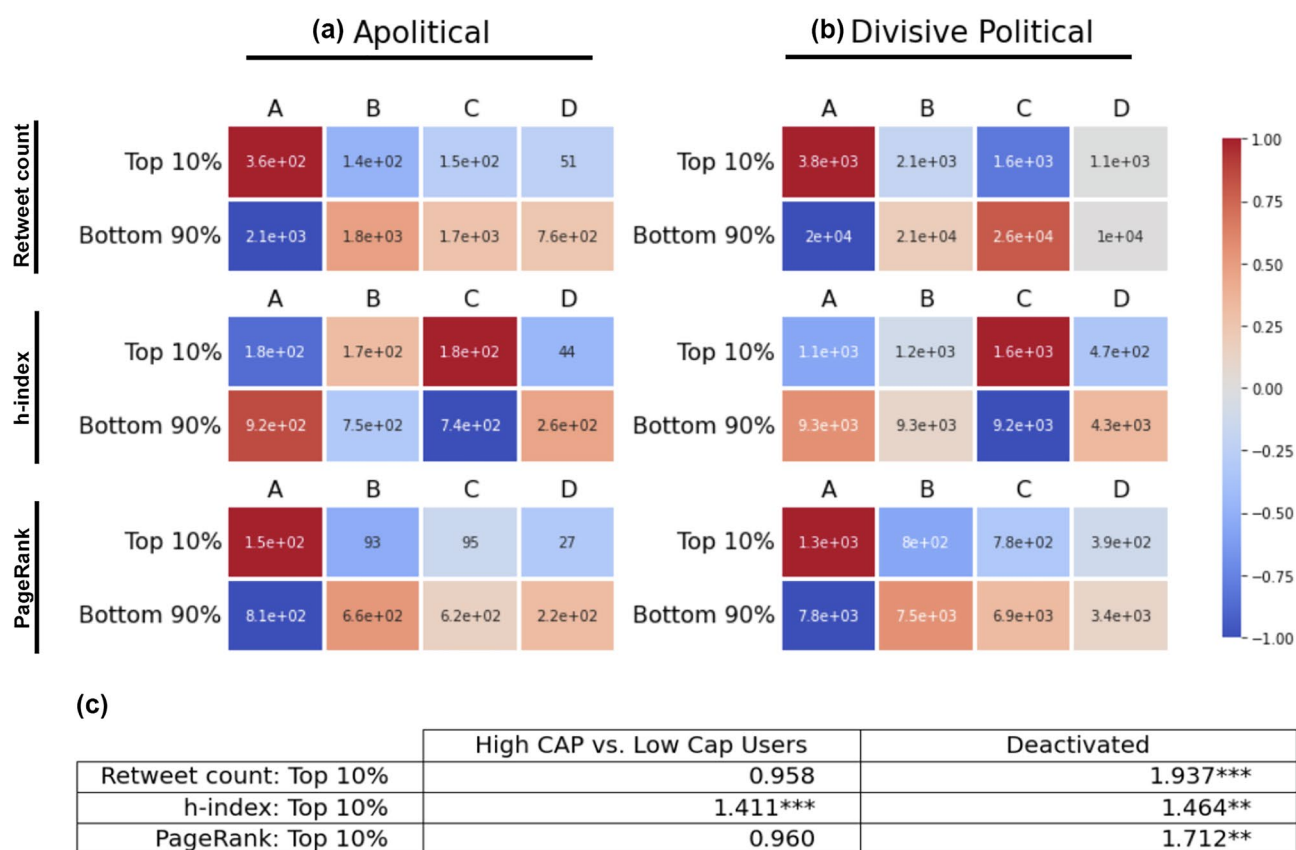

Figure 2. The rows in both the heat maps and the table correspond to retweet count, retweet h-index, and retweet PageRank, from the top row to the bottom row. (**a**, **b**) The heat maps show the enrichment-depletion patterns of users in the top-decile with respect to each measure, with the same vertical order as in the table, in (**a**) apolitical and (**b**) divisive political discussions. User groups A (high-CAP), B (mid-CAP), and C (low-CAP), correspond to the three types with respect to our account automation typology. User group D indicates deactivated or suspended accounts. For each cell, the quantities are centered around the values with a randomly assigned distribution, conditional on the observed population of each row and column, and normalized to a −1 to +1 range. The annotations within each cell show the raw quantities. (**c**) The 'Deactivated' column in the table shows the odds ratios of deactivation or suspension in divisive political discussions compared with apolitical discussions, for users belonging to the top decile by each of the three influence measures—retweet count, retweet h-index, and retweet PageRank. Similarly, the column labeled 'High CAP vs. Low Cap Users' in the table shows the odds ratios of belonging to the high-CAP group of users (A), as opposed to the low-CAP group (C), for the same topics and with respect to the same influence metrics. Chi-square tests are performed to verify statistical significance. '*', '**', and '***' indicate significance levels 0.1, 0.05, and 0.01, respectively.

users in each network cluster for each discussion. We use this manual annotation to confirm the expected order of CAP scores across user types according to our typology. The manual annotation is subsequently matched against the CAP scores to find the interval boundaries for each type. In particular, we find the boundaries on CAP intervals, such that the overall mismatch between manual annotations of the of accounts falling in the same group is minimized. The details of the criteria we use and the validation process are described in the "User typology" subsection.

We find substantially different temporal behavior across different discussion types. To illustrate this dynamic, we provide evidence from one divisive political discussion about a female passenger who was refused service by a driver from Snapp, a popular ride-sharing service, who accused her of not complying with Iran's mandatory hijab law. Given existing controversies around the issue, crossing personal freedom and women's rights, the discussion became a trending divisive political discussion. We contrast this with one apolitical topic about Valentine's day. Although observance of Valentine's day is potentially controversial and could be associated with political or ideological orientations, this Twitter discussion remained apolitical throughout the period when it was trending. Results from all other apolitical and divisive political topics can be found in the Supplementary Information accompanying this paper. In divisive political discussions, we see that automated inauthentic accounts (group A) tweet earlier on and more often than bot assisted humans and trolls (group B) and genuine users (group C), helping to drive the topic to trend. This pattern is displayed in Fig. 3, which shows the temporal dynamics of participation about the Snapp discussion. This stands in contrast to apolitical topics, where all groups follow the same temporal tweeting pattern, as it is demonstrated for the Valentine discussion in Fig. 3. In order to quantitatively validate our observation, we compute the Spearman correlation coefficients for number of tweets per group for each discussion. The results, visualized in Fig. 3, reveal highly correlated temporal dynamics between different groups in apolitical discussions, but negligible correlations between participation pattern of inauthentic users and genuine users in divisive political discussions. This is in agreement with the qualitative difference observed in the tweet counts in Fig. 3.
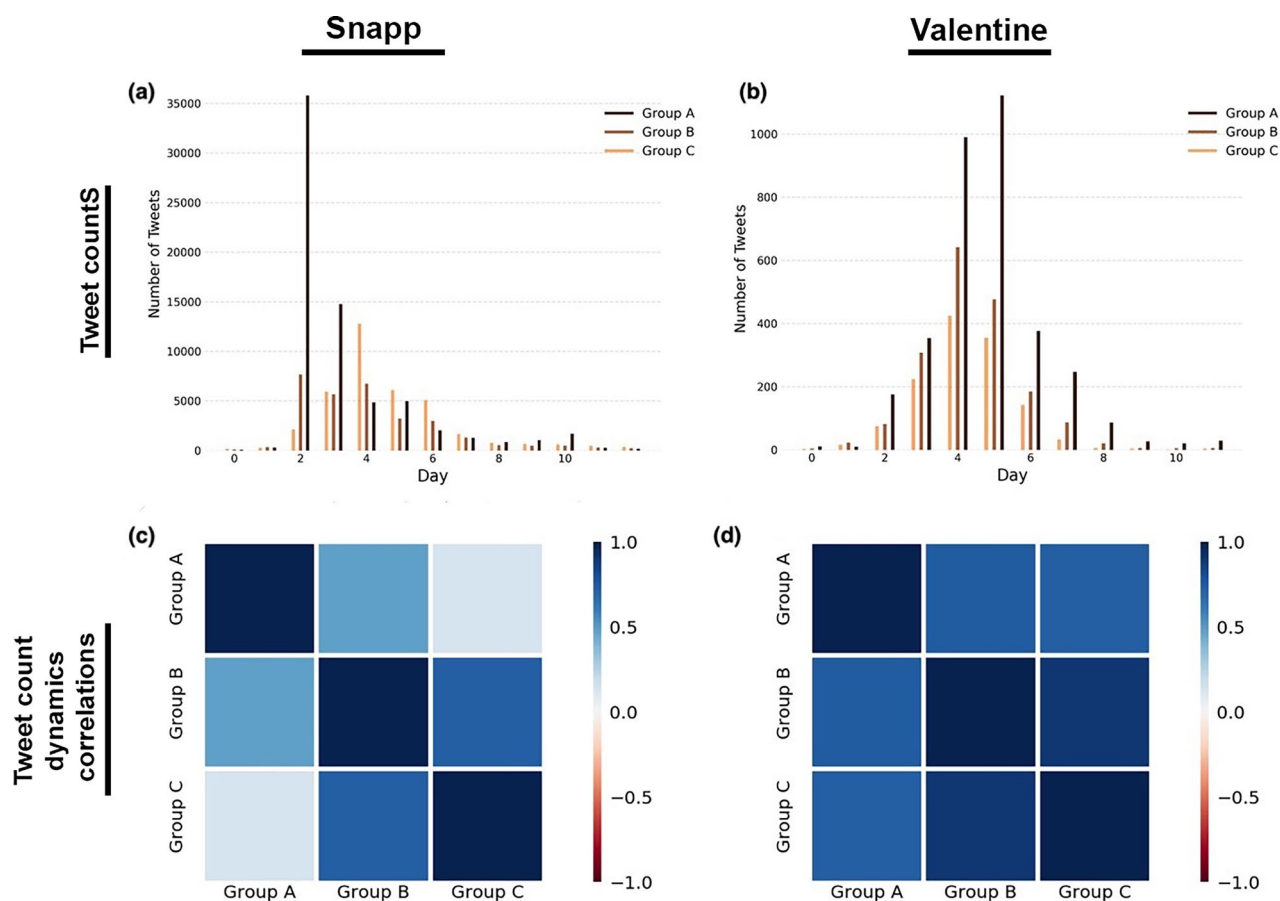

Figure 3. Top row: The number of tweets by each group of users in each 24-hour time interval since the beginning of the discussions. (**a**) Snapp, a divisive political discussion. (**b**) Valentine, an apolitical discussion. Bottom row: The Spearman correlation between the changes in the number of tweets from one day to the next for different groups of users. (**c**) Snapp, a divisive political discussion. (**d**) Valentine, an apolitical discussion. The users are grouped, according to our typology described in the "User typology" subsection, into automated inauthentic accounts (group A), bot assisted humans and trolls (group B), and genuine users (group C). The data for the plots in the left column (Snapp) contains 144347 tweets, and the data for the plots in the right column (Valentine) contains 7962 tweets.

Meso-scale structure and communities of inauthentic accounts.   Examining the structure of communities participating in trending discussions provides insight into the dynamics of inauthentic behavior. In each of the divisive political discussions we observe at least one major cluster with a higher aggregate level of inauthentic activity than that of the entire network. In apolitical discussions on the other hand, we see less variation in levels of inauthentic activity across all clusters. This is illustrated in Fig. 4, which shows the enrichment patterns of each type of account with respect to our typology, and deactivated or suspended accounts, in each of the 4 largest network communities in the friendship and retweet networks. The odds ratios of a user from the high-CAP group belonging to the largest network communities in divisive political discussions to that in apolitical discussions are large and statistically significant with $p < 0.001$ in the Chi-square test (see the table in Fig. 4).

Our analysis of the distribution of different user types across the retweet and friendship networks further reveals differences between divisive political and apolitical discussions in terms of the core-periphery network structure, which is related to amplification of exposure to content across the network on Twitter[20]. The results, visualized in Fig. 4, reveal that inauthentic groups of users tend to be significantly more concentrated in the core, in both friendship and retweet networks and across both types of discussions. On the other hand, while in divisive political discussions low-CAP users are highly concentrated in the periphery and largely absent from the core of the friendship networks, the difference in apolitical networks is relatively less noticeable. There is no considerable difference in the enrichment of genuine users in core and periphery of the retweet networks.

The distribution of user types over the friendship networks also reveals significant differences between apolitical and divisive political discussions with respect to another meso-scale structural partitioning of the network—bow-tie structure[21]. The bow-tie structure of directed networks identify where the nodes stand with respect to the direction of information flow, and are known to correspond to discursive communities in online social networks[22]. We consider the strongly connected core of the bow-tie ('S'), the immediate incoming gate of the S component ('IN'), and the immediate outgoing gate of the S component ('OUT') in the friendship network. More details on how we detect and utilize the bow-tie structure can be found in the "Network analysis" subsection. In the context of friendship networks of Twitter discussions, the components of the bow-tie mark the direction of exposure, or equivalently the potential propagation of content, among users engaged in the discussion. Since the direction of the edges are from followers to friends, the OUT component could be viewed as the primary source of content propagation. Our findings show relatively high concentration of high-CAP users in the OUT component of the bow-tie within the divisive political discussions, while this is not the case for apolitical discussions (see Fig. 4). This signals relatively higher potential of high-CAP users for impacting the content of the discourse in divisive political discussions. On the other hand, high-CAP users are largely absent from the S component of the bow-tie in divisive political discussions, which indicates that genuine users have higher potentials for facilitating the circulation of content in these discussions.

Differences in content.   Comparing the content of tweets produced by communities with high and low levels of inauthentic activity, in divisive political discussions, we notice a more polarizing language in communities with higher levels of inauthentic activity. However, when we look at apolitical discussions, we see little difference in the language used by communities with higher and lower levels of inauthentic activity. This can be seen from Figs. 5 and 6, which, respectively, display the most frequent words and salient topics, among users grouped by our inauthenticity typology. The word clouds in Fig. 5 compare the highest frequency words (translated from Farsi to English) in communities of genuine users (group C) to those produced by automated and semi-automated bots (group A), in the divisive political discussion about Snapp and the apolitical discussion about Valentine. More specifically, they show the words that are among the top 5% most frequent words in the tweets posted by automated inauthentic users, but not among the top 10% most frequent words used by genuine users. While here we discuss the Snapp and Valentine discussions in detail, the main observations in these word clouds generally hold across other discussions of the same type, i.e. divisive political and apolitical discussions. We further use an unsupervised transformer-based topic detection algorithm to find salient topics along with most representative words within each topic for each group of users within each discussion. Additional details about the content analysis are explained in the "Content analysis" subsection. Our observation from comparing top words within the first 4 topics (Fig. 6) is aligned with what is observed in the word clouds, highlighting the polarizing language within each group in divisive political discussions.

In order to interpret the content analysis, we provide further context for the Snapp discussion. The Snapp discussion occurred after a driver employed by Snapp, the country's biggest ride-sharing enterprise, refused to take a passenger to her destination because she had not been wearing a hijab, which is mandatory according to a controversial clothing law for women in Iran. The topic started to trend on Twitter after the passenger tweeted about the incident, criticizing the driver for making her leave the car, and demanding accountability from the company, Snapp. Subsequently, a dispute erupted between Twitter users expressing a range of opinions from supporting the passenger to reproaching her. In Fig. 5 we observe that inauthentic users (group A) frequently used polarizing words ideologically aligned with the Iranian state. For example, the words 'Law-breaker', 'Law-lessness', and 'Unprincipled' carry a critical tone towards the clothing of the passenger involved in the incident. Moreover, the words 'Value', 'Norm', and 'Dignity' are also representative of the vocabulary corresponding to the same ideology. However, genuine users (group C) more frequently used words such as 'Bitter', 'Punishment', and 'the saddest', signifying their sympathetic stance with the passenger in the story. Similarly, the words with highest word scores among the top-4 detected topics show qualitative differences between the content of tweets from inauthentic and genuine users, shown in Fig. 6. For instance, 'Observe', 'Respect', 'Muslim', representative words of topics discussed by high-CAP group, imply advocacy in favor of the driver involved in the event, which is in line with the position implied by the most common words in the corresponding word cloud. Meanwhile, words
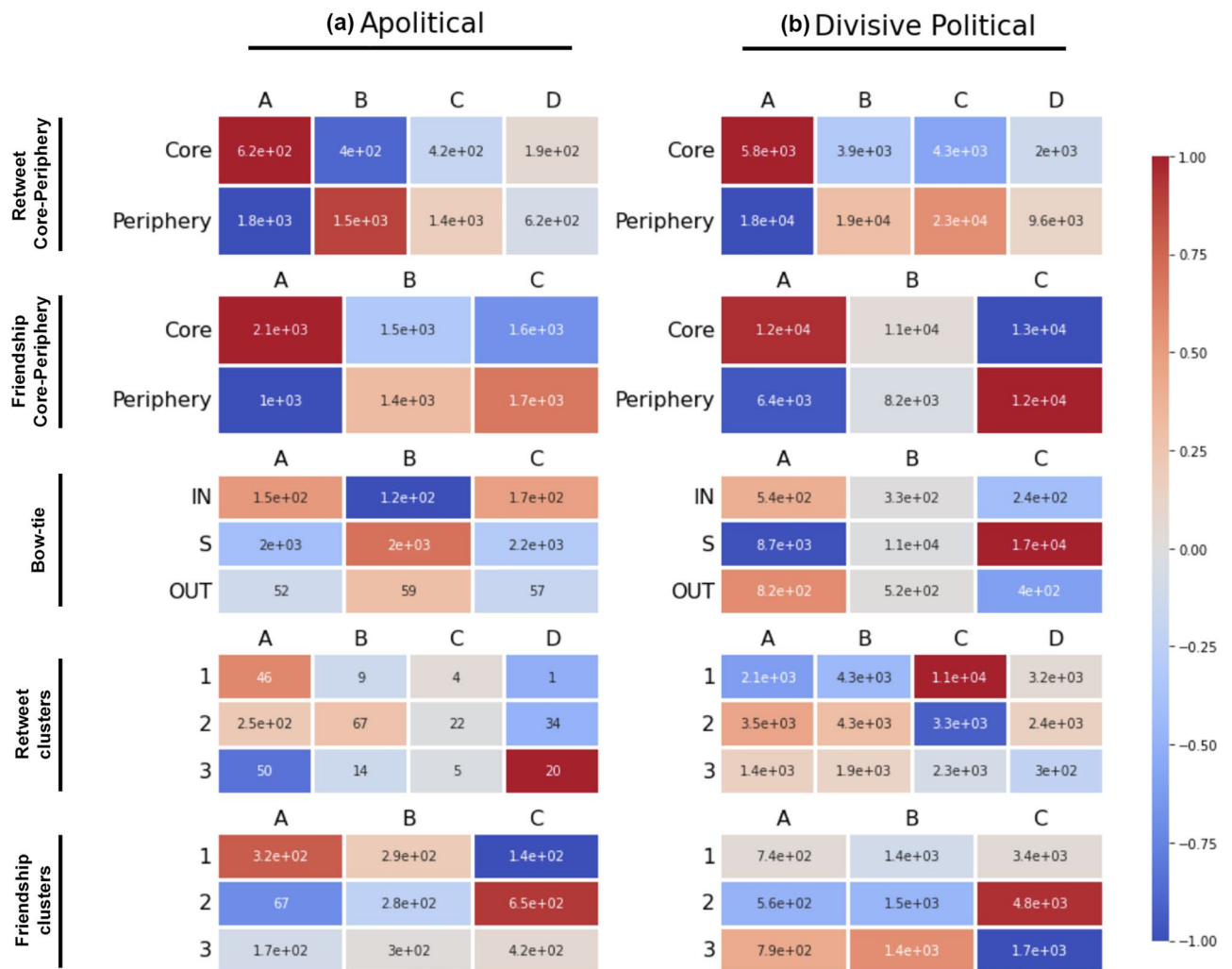
Figure 4. The rows in both the heat maps and the table correspond to retweet core-periphery, friendship core-periphery, friendship bow-tie, retweet clusters, and friendship clusters, from the top row to the bottom row. (**a**, **b**) The heat maps show the enrichment-depletion patterns of users in the top-decile with respect to each measure, with the same vertical order as in the table, in (**a**) apolitical and (**b**) divisive political discussions. Users groups A (high-CAP), B (mid-CAP), and C (low-CAP), correspond to the three types with respect to our account automation typology. User group D indicates deactivated or suspended accounts. For each cell, the quantities are centered around the values with a randomly assigned distribution, conditional on the observed population of each row and column, and normalized to a −1 to +1 range. The annotations within each cell show the raw quantities. Note that, since deactivation of an account disrupts the collection of neighbors of the account in the friendship network, nodes from group D are not considered in the heat maps corresponding to the friendship networks. (**c**) The 'Deactivated' column in the table shows the odds ratios of deactivation or suspension in divisive political discussions compared with apolitical discussions, for users belonging to each group of nodes in the retweet or friendship networks. That is, from the top row to the bottom row, core or periphery in retweet and friendship networks, bow-tie components in friendship networks, and 4 largest communities in retweet and friendship networks. Similarly, the column labeled 'High CAP vs. Low Cap Users' in the table shows the odds ratios of belonging to the high-CAP group of users (A), as opposed to the low-CAP group (C), for the same topics and with respect to the same partitioning of the network. Chi-square test is performed to verify statistical significance. '*', '**', and '***' indicate significance levels 0.1, 0.05, and 0.01, respectively.
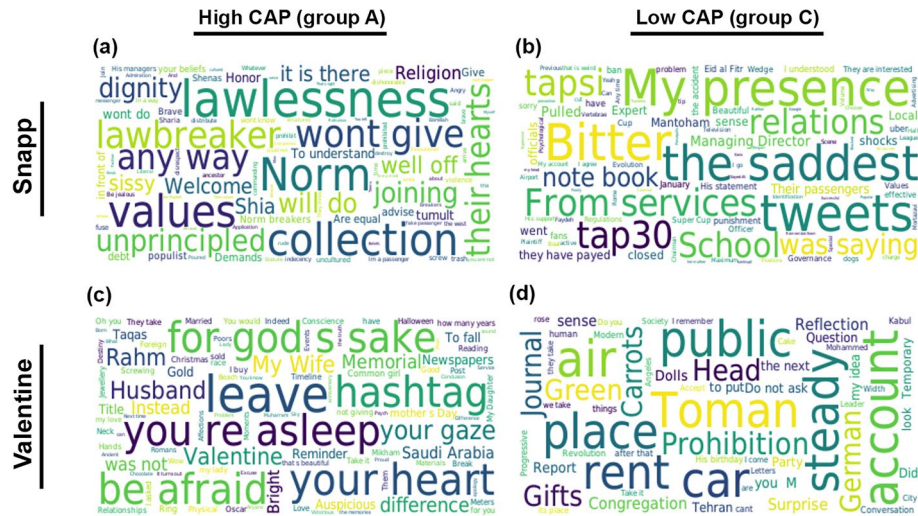
Figure 5. Difference word clouds by CAP for two discussions. (**a**, **b**) Snapp, a divisive political discussion. (**c**, **d**) Valentine, an apolitical discussion. (**a**) and (**c**) include the top 5% most frequent words across the high-CAP group which are absent from the top 10% most frequent words used by the low-CAP group. (**b**) and (**d**) include the top 5% most frequent words across the low-CAP group which are absent from the top 10% most frequent words used by the high-CAP group. All words are translated from Farsi to English for ease of visualization.

such as 'IRGC', 'Tap30', and 'Mandatory', which are representative of salient topics in the content produced by the low-CAP accounts, support the opposite stance.

By contrast, the other pair of word clouds in Fig. 5, which display the characteristic words across authentic and inauthentic users in the apolitical discussion about Valentine's day, do not display meaningful differences between the two groups. Neither the words appearing as highly frequent words only among the inauthentic users (group A) nor those frequently used only by the genuine users (group C), convey any particular meaning or otherwise signal a divide in the content of discussion between the low-CAP and high-CAP users. This also holds for the representative words of the topics detected through topic modeling on the tweets in the Valentine discussion, where we do not observe a substantive difference between the topics corresponding to the two groups of users (see Fig. 6).

In addition to comparing the content of tweets from authentic and inauthentic users (as measured by their CAP scores), we make the same comparison for users belonging to different clusters in friendship networks of each trending discussion. Figure 7 displays the difference between the high-frequency words used by users belonging to two major communities in the friendship network of the Snapp discussion, which is a divisive political discussion. This is shown in word clouds constructed in the same fashion as the word clouds in Fig. 5, i.e. each word cloud is showing the words that are among the top 5% most frequent words in one community but not among the top 10% in the other. The words in the word cloud on the left clearly signal a conservative rhetoric in the corresponding tweets, while the word cloud on the right includes words from the common vocabulary of the dissidents and the groups that are critical of the status quo. Similar to the comparison between the content from low-CAP and high-CAP users, we also compare the detected topics and their representative words in each of the two friendship communities. This comparison, shown in Fig. 8, provides further evidence for polarization of content between two major communities in the friendship network of the Snapp discussion.

Notably, we observe a difference between the communities in the friendship network. Friendship networks are structural constructs formed by follower-friend connections between the participants in the discussion. These connections are often made prior to the discussion, and have no direct link to the content of the tweets. Meanwhile, the network clustering algorithm is also completely blind to the text of the tweets or otherwise-content-related features. Despite this, we can observe a clear difference between the content of tweets tweeted by users from different network clusters, which is indicative of the political orientation of these users. Due to how user activities appear in their followers' feeds, a community in the friendship network approximately corresponds to a neighborhood of users with significantly overlapping exposure to content. Hence, the clear difference in the content of the tweets tweeted by users from different network communities speaks to the differences in the exposure of these users. In other words, users belonging to different communities are exposed to qualitatively different content, with opposing political rhetoric, which suggests a presence of echo chambers in divisive political discussions. By contrast, there is no meaningful difference between the detected topics and highly frequent words used by different friendship communities in apolitical discussions, as it is shown in Figs. 7 and 8.

Inauthentic activity in echo chambers. Our comparison of the community structures in the retweet and friendship networks provides more insight into the communities in which inauthentic accounts insert themselves. This analysis reveals that echo chambers are much more prevalent in divisive political conversations relative to apolitical conversations. Figure 9 compares the communities in the retweet and friendship networks of the Snapp (divisive political) and the Valentine (apolitical) discussions. The upper triangle in each plot on
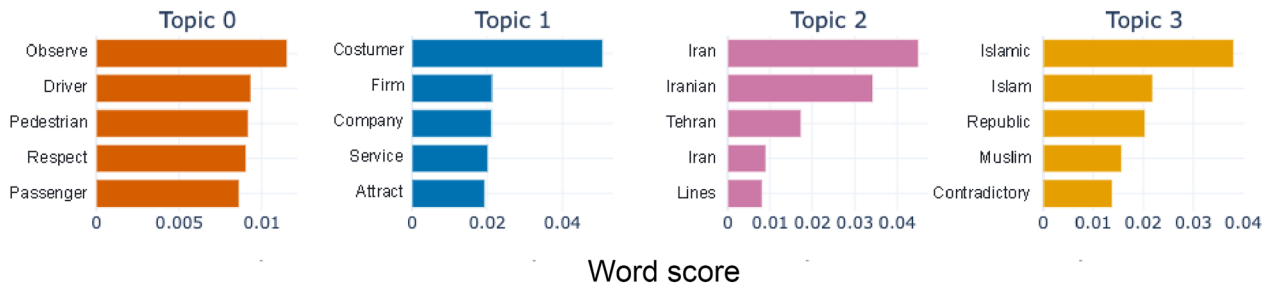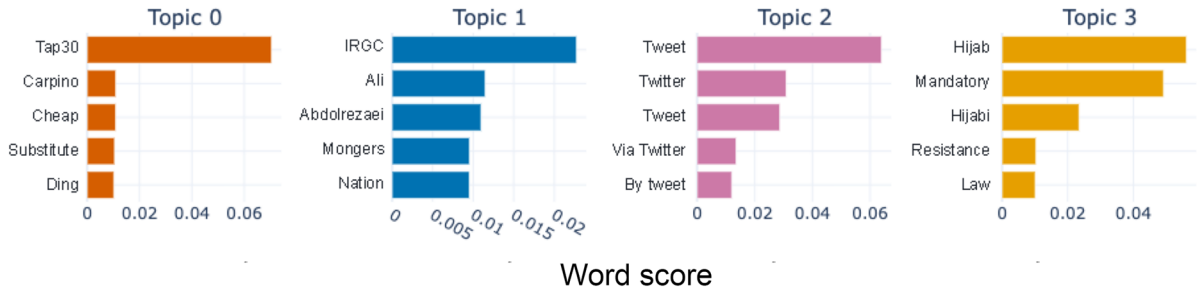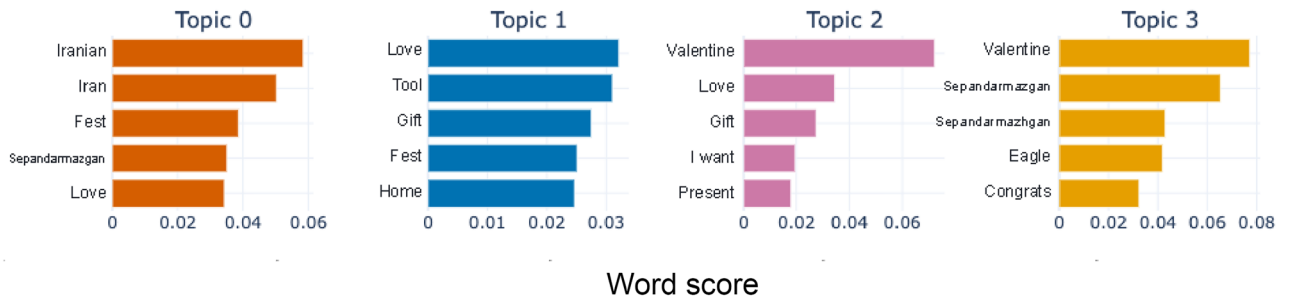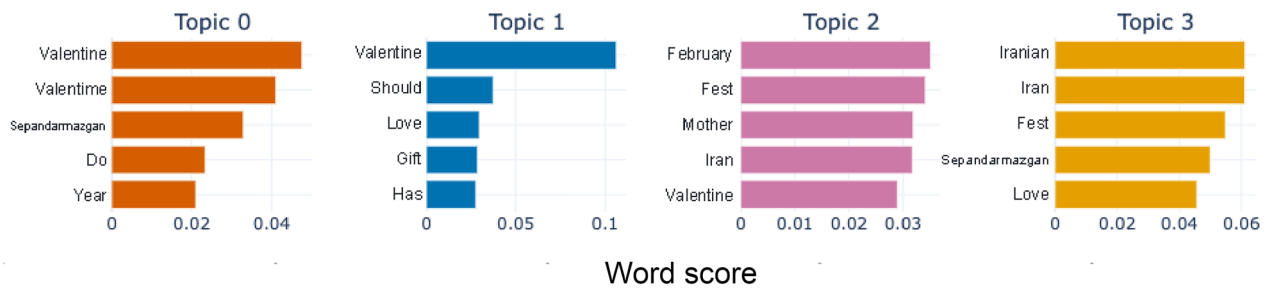
Figure 6. Salient topics and highest-scored words within each topic, obtained from BERTopic, for tweets from (**b**, **d**) low-CAP (C) and (**a**, **c**) high-CAP (A) user types in two discussions. (**a**, **b**) Snapp, a divisive political discussion. (**c**, **d**) Valentine, an apolitical discussion. All words are translated from Farsi to English for ease of visualization.

Fig. 9 shows the symmetrized adjacency matrix of the retweet network with a permutation of rows and columns that groups the nodes by their retweet community. The lower triangle on the other hand, shows the same symmetrized adjacency matrix but with an alternative permutation which groups nodes by the friendship communities they belong to. As the lower triangle of the right plot in Fig. 9 suggests, the friendship communities of the Snapp discussion induce a clustered structure on the retweet network. Furthermore, comparing the lower and upper triangles, we can see that this induced clustering has a rather considerable overlap with the retweet communities. This, however, is not the case for the Valentine discussion, as we can see from the plot on the left of
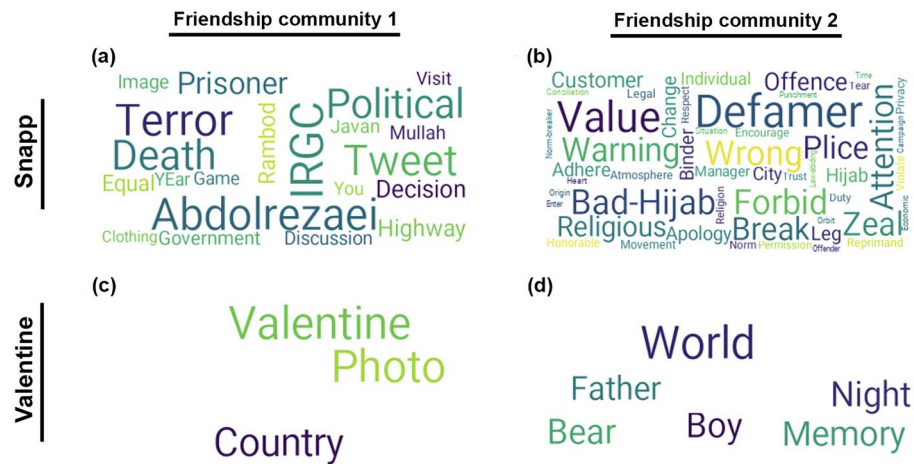
Figure 7. Difference word clouds for friendship network communities corresponding to two discussions. (**a**, **b**) Snapp, a divisive political discussion. (**c**, **d**) Valentine, an apolitical discussion. (**a**) and (**c**) include the top 5% most frequent words used by users in community 1 which are not among the top 10% most frequent words used by users in community 2. (**b**) and (**d**) include the top 5% most frequent words used by users in community 2 which are not among the top 10% most frequent words used by users in community 1. All words are translated from Farsi to English for ease of visualization.

Fig. 9, which does not show a qualitatively meaningful clustering of the retweet network induced by the friendship communities.

Note that a retweet community is the group of users among which circulation of content takes place, while a friendship community signifies the neighborhood in the network where users are exposed to a mutual content. The overlap between the two communities, which we see in the Snapp discussion in Fig. 9, suggests the presence of echo chambers. We see a similar pattern across divisive political discussions, indicating the presence of echo chambers, which we do not observe in apolitical discussions, where the friendship communities do not induce a pronounced clustering on the retweet network. This can be seen for the Valentine discussion, as an example of an apolitical discussion, in Fig. 9. As the sparsity pattern of the adjacency matrix of the retweet network ordered by the friendship communities shows (lower triangle), the firendship network communities are scattered across the retweet network of the Valentine discussion. This is consistent across other apolitical discussions as well, which indicates lack of evidence for formation of echo chambers in this category of discussions.

Figure 9 visualizes the difference between divisive political and apolitical discussions in terms of the overlap between retweet and friendship communities in the case of two example discussions. In order to further verify this result, we quantify the overlap between retweet and friendship communities, as well as polarization of the network structure, for each type of discussion. Given the randomness involved in the community detection method and the sensitivity of measures of similarities between two clusterings (see[23]), instead of comparing only one pair of clustering outcomes, we consider the distribution of pairwise clustering similarities between retweet and friendship networks in an ensemble of network clusterings. Using two measures of clustering similarities, Fig. 10 compares the empirical cumulative distribution functions (CDF) of the similarities between the retweet and friendship networks in apolitical and divisive political discussions. The details about the similarity measures and how the empirical CDF's are obtained are explained in the "Network analysis" subsection. As we can see in Fig. 10, the empirical CDF corresponding to divisive political discussions falls below that of the apolitical discussions across almost the entire range of similarity values, according to both measures. Furthermore, the empirical CDF's corresponding to apolitical discussions reach 1 very rapidly, at very small similarity values. This means the retweet and friendship networks of apolitical discussions tend to be consistently dissimilar across several runs of the clustering algorithm, while the structure of friendship and retweet networks of divisive political discussions can lead the clustering algorithm to output considerably overlapping friendship and retweet communities. In other words, in the light of the discussion above about the qualitative connection between echo-chambers and overlapping friendship and retweet communities, Fig. 10 suggests that the network structure of divisive political discussions is conducive to formation of echo-chambers, while this is not the case in apolitical discussions.

As an additional assessment of structural polarization of the network, we compute the modularities of both types of networks for each discussion, the distributions of which, over ensembles of multiple runs of the clustering algorithm, are shown in Fig. 11. A large modularity means users are densely connected within each community, with relatively few out-community connections[24]. Hence, modularity is a measure of segregation of users into clusters, and is used to evaluate the potential for structural polarization in online social networks (see e.g.[25,26]). We compute the modularity for 100 runs of the clustering algorithm on each network, which yields a distribution of modularity values for each network and discussion type. The details are described in the "Network analysis" subsection. These distributions are shown by the histograms in Fig. 11, where we can see that the friendship networks in divisive political discussions tend to have relatively large modularity values concentrated over a narrow range, while this is not the case for apolitical discussions, where there seems to be a lack of inherently
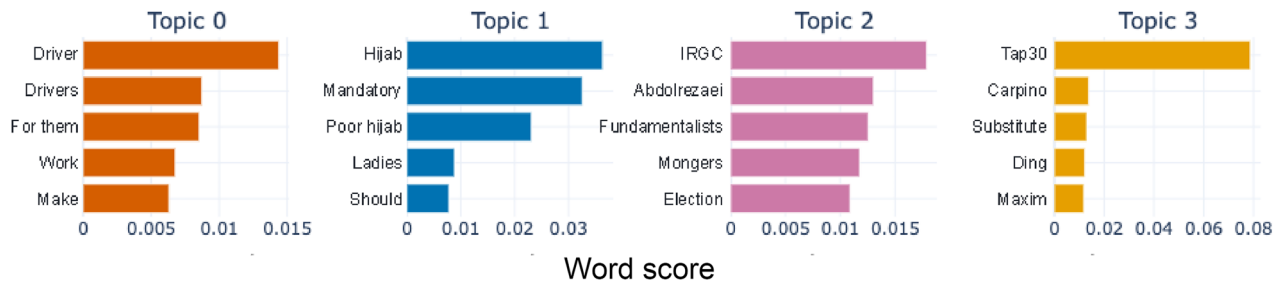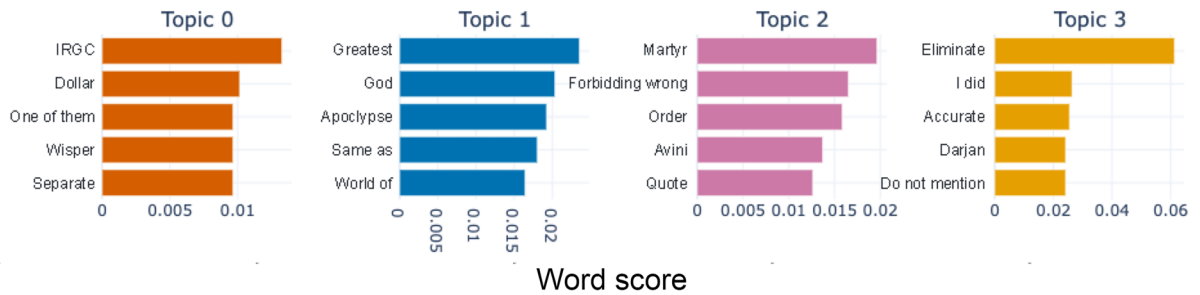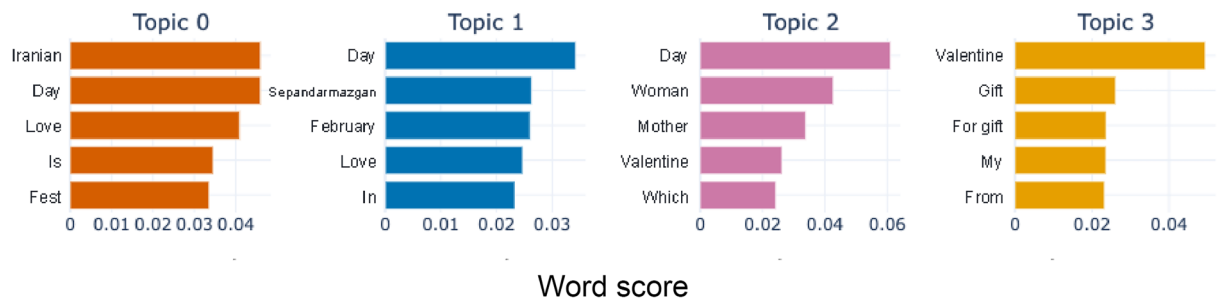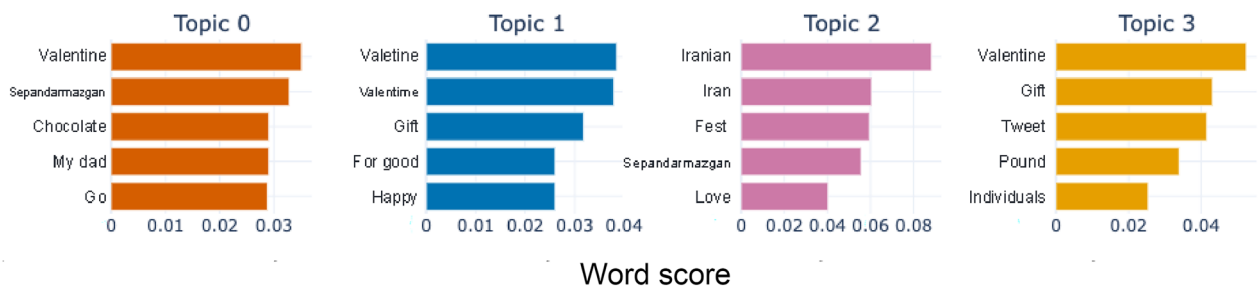
Figure 8. Salient topics and highest-scored words within each topic, obtained from BERTopic, for tweets from users in two friendship communities. (**a**, **b**) Snapp, a divisive political discussion. (**c**, **d**) Valentine, an apolitical discussion. All words are translated from Farsi to English for ease of visualization.

segregated network structure. While the modularity values for the retweet networks in divisive political discussions have a narrower range, and those for apolitical discussions tend to have larger values, given what the retweet network represents, this does not convey information on echo chambers on its own. However, in light of our observation on the difference in friendship modularities, the relative similarity in retweet modularities suggests that, while users retweet each other in a rather modular and segregated fashion in both types of discussion, the friendship structure of divisive political discussions makes them more prone to segregated exposure to the retweets. This provides further evidence for the relationship between network structure and polarization as well
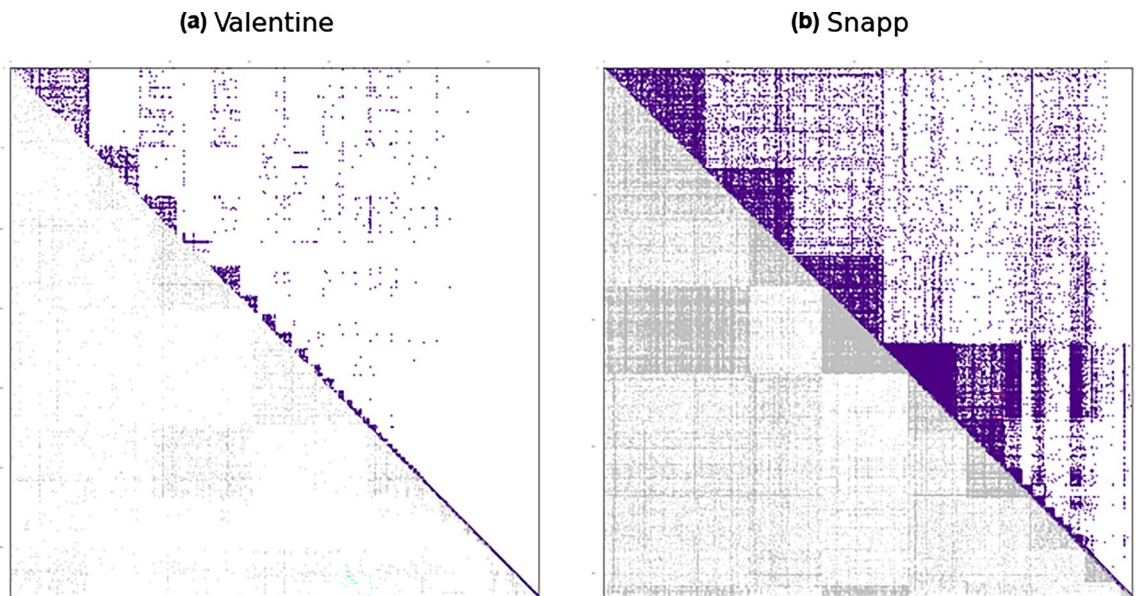
Figure 9. The symmetrized retweet networks for the Twitter discussions (**a**) Valentine and (**b**) Snapp. In the upper triangle the rows and columns are reordered such to group nodes by their retweet communities. In the lower triangle, the reordering groups nodes by the community they belong to in the corresponding friendship network. Excluding the isolated nodes, the retweet network of the Snapp discussion has 21029 nodes and 69565 edges, and the retweet network of the Valentine discussion has 3705 nodes and 4491 edges.
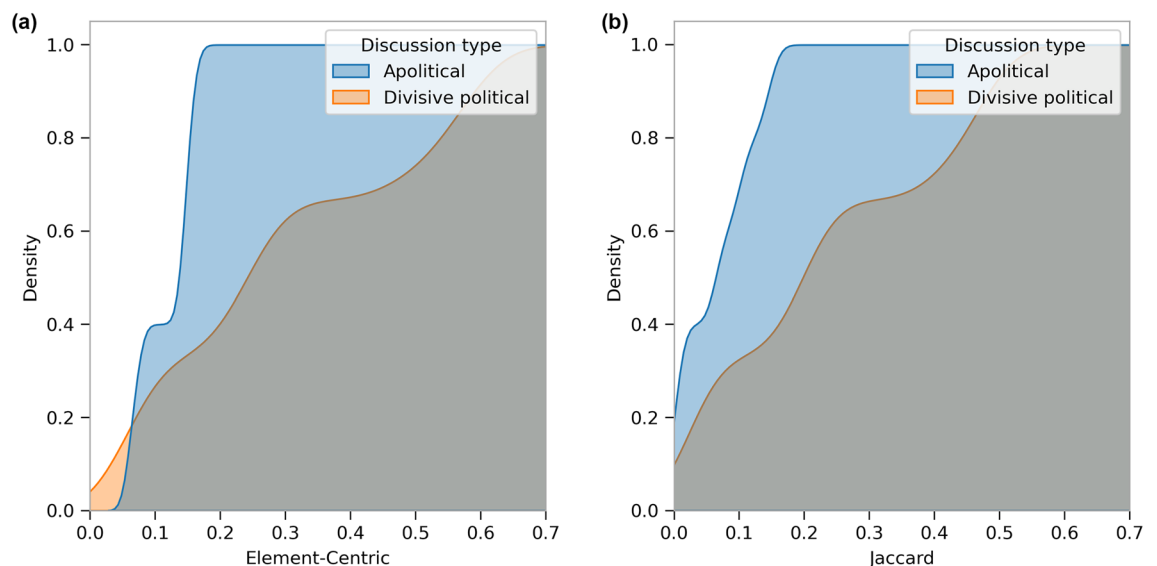

Figure 10. The empirical cumulative distribution functions (CDF) of the similarities between the retweet and friendship networks of the apolitical (blue) and divisive political (orange) discussions, over 20 different runs of Louvain clustering for each network. That is, for each discussion, the ensemble contains 400 pairs of similarity values. Note that the diagrams use a Kernel Density Estimate to obtain smoothed distributions. (**a**) The similarities are computed according to element-centric similarity[23]. (**b**) The similarities are computed according to Jaccard index.

as the emergence of echo chambers in divisive political discussions, which is consistent with our interpretation of the overlap between friendship and retweet communities.

Discussion

Taken together, our analysis of inauthentic activity across trending topics in the Farsi Twittersphere demonstrates that inauthentic accounts are very active in divisive political discussions, where they often initiate trending topics and use polarizing language to advance partisan agendas. This stands in contrast to their activity in apolitical discussions, where inauthentic accounts are less prevalent and engage in similar discourse to genuine users. The
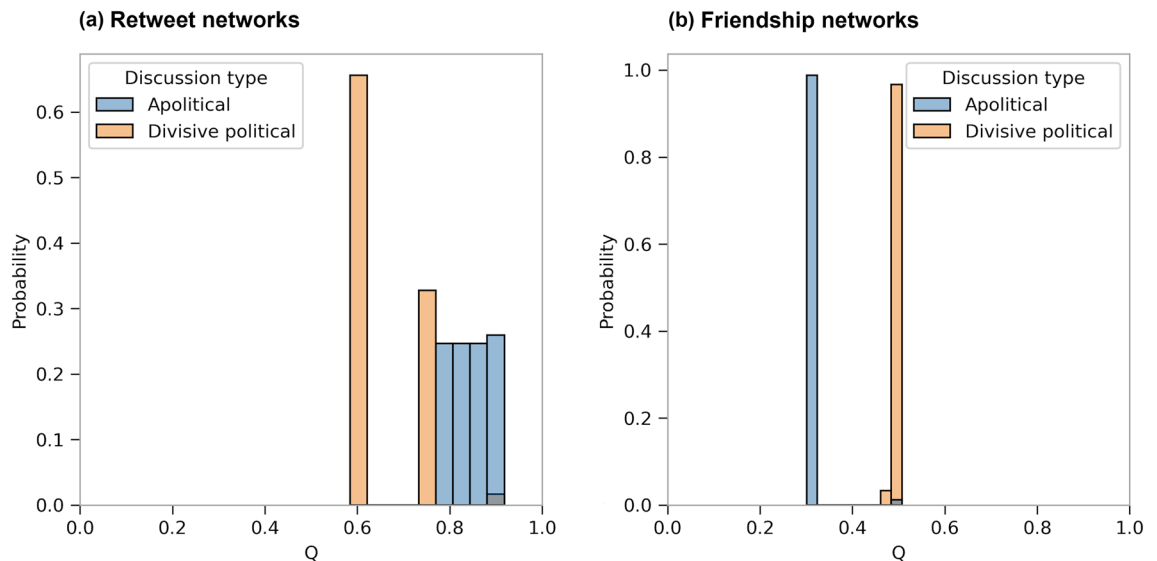
Figure 11. The distribution of network modularities (denoted *Q* on the horizontal axis) in the (**a**) retweet and (**b**) friendship networks of the apolitical (blue) and divisive political (orange) discussions, over 100 different runs of Louvain clustering for each network.

network structure of these conversations provides insight into the communities in which inauthentic accounts operate. Divisive political conversations occur within echo chamber environments, where retweet networks and friendship networks overlap to a high degree. By contrast, apolitical conversations bridge the partisan divide, reaching users of diverse ideological persuasions. Inauthentic accounts are therefore able to advance partisan narratives in echo chamber environments where they can target specific partisan audiences.

By analyzing the dynamics of inauthentic activity in the Iranian Twittersphere, our study contributes to a growing body of literature exploring online influence operations in several ways. First, we provide a typology of accounts enabling us to distinguish between automated inauthentic accounts, human-controlled inauthentic accounts, and genuine users. Second, we provide evidence examining the content, structure, and temporal dynamics of inauthentic activity, providing a rich multi-method characterization of inauthentic online behavior. Third, we provide evidence from the Farsi Twittersphere, an understudied context where inauthentic activity is quite prevalent. Fourth, unlike most existing studies, we cover diverse issue areas, enabling us to compare inauthentic behavior across divisive political, non-divisive political, and apolitical topics in the same analysis. The multifaceted analysis of several Twitter discussion spaces, allowed us to overcome common challenges posed by errors in bot-detection methods, missing data points, and method-induced ambiguities in isolated analysis techniques. As a result, we are able to provide systematic descriptive evidence on how inauthentic accounts engage in online discourse on Farsi Twitter.

Despite these contributions, our study has several limitations that suggest avenues for future research. Our study is limited to just one platform—Twitter—and therefore does not allow us to characterize online inauthentic activity more generally. Having said that, existing research on inauthentic accounts and opinion manipulators on Facebook, reveals partial similarities between features of these accounts and those identified in our typology for inauthentic Twitter accounts. For instance, inauthentic accounts carrying out information operations on Facebook are deployed for content creation and false amplification of content[27]. Additionally, inauthentic accounts on both Facebook and Twitter have been used for engagement with authentic accounts to manipulate engagement statistics[28]. For detection of inauthentic activities on Facebook, combining manual and automatic labeling to improve reliability, previous works have found similar validation techniques to what is used in our study to be effective[29]. The authors have also pointed to the potential utility of network properties in improving detection methods on Facebook[29], and our study can point to directions for guided investigations of network features that could be most helpful to consider, given our findings regarding the patterns of inauthentic activities across network neighborhoods. We hope that future research will include cross-platform analyses that enable us to better understand the broader ecosystem of online coordinated inauthentic activity. Additionally, measuring inauthentic activity is very challenging. While human validation of our measurement approach suggests it performs well at an aggregate level, existing automated approaches prevent us from accurately classifying accounts at the individual level, an important step for future research.

Recent research demonstrates that state and non-state actors are increasingly leveraging social media platforms to run influence operations. Understanding how these actors operate across diverse issue areas and global contexts is therefore crucial from a policy perspective. We hope that online platforms will continue to make data on information operations available to researchers, to help us to improve our characterization and measurement of online inauthentic activity. In the absence of better data and measurement tools, we encourage researchers to draw on similar methodological approaches to those we present here to continue to characterize inauthentic activity across different issue areas, time periods, and contexts.

Methods

In this section, we describe our topics of discussion, clarify our dataset and details of our data collection, and explain our data analysis methods. Moreover, in order to understand the configuration of our discussion spaces in terms of user types with respect to automation and authenticity, we introduce a typology which helps reduce ambiguities in our observations.

Selected topics. Different topics of discussions in Twitter can shape discussion spaces with different qualities. For instance, Smith et al.[30] introduce six different types of political discussions in Twitter with respect to their social structure. Our exploratory analysis of trending discussions in Farsi Twitter revealed three main types of topic of discussion with qualitative differences in the development of their trend, network structure, and participation of users. In order to verify consistency in our observations across various discussions of each type, for the targeted phase of our data collection, we selected an overall of eleven topics including four apolitical discussions, four non-divisive political discussions, and three divisive political discussions. The analysis in the "Results" section focuses on the characteristics of divisive political discussions compared against apolitical discussions. Please note that the border between non-divisive and divisive political discussions is rather blur, and there is a transition zone between these two types. While the contrast between divisive political and apolitical discussions is more evident in our results, allowing us to more clearly characterize these two categories of discussions, non-divisive political discussions mostly show mixed features. The discussion topics are described in detail in the Supplementary Information.

In order to make sure every discussion space has the potential to show certain characterizing features such as polarization or formation of echo chambers, all chosen subjects cross socio-political cleavages in the Iranian society. Therefore, although not directly political, all events can turn into a political handle for supporters of major Iranian political groups in Twitter. This, in turn, could create incentive for political organizations to engage in opinion manipulation, giving rise to inauthentic activities.

Data. *Twitter data.* We collected tweets by keywords at a time in the vicinity of the peak trendiness of our target topics in Farsi Twitter. Our data collection machine uses Twitter's Standard Search API, filtered by keyword and language (Farsi). All tweets made available through the API were collected over a period spanning the rise and fall of the trend. We used PostgreSQL[31] as a relational database to store references for tweet and user objects and SQLite[32] was used to re-index the objects obtained through the API for each individual topic. Using API endpoints of Twitter for followers and friends, we also collected followers and friends of users in our data, indexed them in a designated SQLite database, and constructed the edge lists. If an account, X, is private or deactivated before its followers/friends are requested through the API, it appears in friends/followers list of its follower/friend, Y, if Y is public and active. However, queries to the API for obtaining the friends/followers of X fail. We combined the data from friends and followers lists to recover some of the missing links due to an account being private or its deactivation in the gap between our collection of the tweets and that of the friends and followers.

Systematic bot detection. To perform big data analysis on account types with respect to the typology described in the "User typology" subsection, we use Botometer[16–18], a supervised machine-learning algorithm, to collect the botscores and Complete Automation Probability (CAP) for the users in our Twitter data. We use universal CAP (the language-agnostic complete automation probability) from Botometer results for each account as their botscore. We include only accounts that were still active (not deactivated or suspended) by the time we collected the botscores.

Manual bot detection. In an attempt to find an effective and practical approach to using the data obtained from Botometer, we used 1163 manually annotated randomly-sampled accounts: 495 accounts annotated by members of our group and 668 more labeled by the participants in a workshop, who were trained on our typology. The annotations were used to verify the validity of results from Botometer for the purpose of our study and to find a correct threshold for distinguishing between genuine users and likely opinion manipulators. Realizing the subjectivity in bot detection, we distributed the accounts among the participants in our workshop in a way that each account is annotated by up to three different participants. The Kendall's τ for inter-annotator agreement[33] was approximately 0.67. We then chose the CAP intervals such that the variation of labels within each interval is minimized.

User typology. Opinion manipulation in Twitter through different types of bogus accounts is a well-studied subject. Gorwa et al. comprehensively review prior investigations and typologies for major categories of online bots[9]. They group bogus online accounts into six main types: Web robots, chatbots, spambots, social bots, sockpuppets and trolls, and cyborgs and hybrid accounts. Our study concerns accounts that fall in the latter half of these categories - social bots, sockpuppets, and cyborgs. Comparing the findings of Chu et al.[34] with that of Subrahmanian et al.[35] shows the significant improvement of Twitter bots in imitating human behavior, a fact confirmed in previous studies[35–38], as well as our manual observations. Therefore, inauthentic accounts could be extremely difficult to detect by a simple set of measures such as temporal activity patterns, and a more complex combination of measures are necessary for distinguishing them from genuine human users. In this subsection, we introduce a taxonomy of users, with respect to their level of automation, authenticity, and opinion manipulation behavior, which we use to analyze each discussion space. Based on our observations through manual exploration of the Farsi Twittersphere, we group the users into three main groups:

A *Automated and semi-automated bots* Twitter bots, often part of a bot squad[39,40], are used to perform structur-ally repetitive tasks at a noticeably higher rate compared to humans. These are Twitter accounts that could be fully automated without direct involvement of a human, or they can be semi-automated.

B *Bot-assisted humans or human controlled campaigning accounts* These accounts, although may seem like an account belonging to an ordinary person, show a strong similarity to a campaigning account. The content generated by these accounts is strongly oriented towards supporting the standpoint of a group, as one expects to see from an account which belongs to a campaign. These accounts can be divided into the following sub-categories.

B.1 *Deployed users* These are agents that are deployed to make and control accounts and increase the repre-sentation of the stance of an organization. Such accounts are often referred to as trolls[13]. Their behavior is typically hardly distinguishable from that of an official campaign or a propaganda news agency, apart from the user information. They could be bot-assisted, i.e. equipped with a machine to enhance their performance. The main characteristics of these accounts that separates them from Automated and semi-automated bots (group A) and Unusually dedicated users (group B.2) are that, unlike Automated and semi-automated bots, most often they show a cognitive ability deemed to be exclusively possessed by humans, yet unlike Unusually dedicated users, personal content is noticeably absent in their activities.

B.2 *Unusually dedicated users* This group consists of users whose accounts seem dedicated to supporting a cause or an organization. They differ from deployed users in that they show personal activities in their tweets significantly more often, e.g. occasionally tweet personal content, engage in personal interactions with others, or express beliefs that do not fit within the main-stream agenda of the organization they side with. Their general behavior is nevertheless similar to that of Deployed users (group B.1).

C *Genuine users* The users in this group are humans using Twitter to engage in social activities and express their stance. Apart from showing human-like cognitive behavior, the users belonging to this group often have less homogeneous activities. Multidimensional characteristics, personal content, and heterogeneous support of others' stance in a discussion space are among the most outstanding characterizing features of the accounts in this category.

Notice that users that fall in the Automated and semi-automated bots category, as well as those we refer to as Deployed users, could be hybrid accounts controlled partially by bots and partially by humans. However, the hybrid accounts that are Automated and semi-automated bots are human-assisted bots, while Deployed users are bot-assisted humans, i.e. the former are accounts whose performance in cognitive and content-related aspects is improved by humans and the latter are accounts utilizing machines to enhance the quantitative aspects of their activities.

Given this typology, we expect the reflection of authentic users on Twitter space to be mostly contained in group C –Genuine users– and be limited to Genuine users and Unusually dedicated users. Unusually dedicated users are where we expect most of the error in bot-detection to lie. Many accounts belonging to this group may be indistinguishable from those in deployed or genuine users. Neither automated nor manual examination of the accounts could detect structural or content-based features that are exclusive to this group. Therefore, it is impractical to label any account as one belonging to Unusually dedicated users with a high degree of confidence.

In our analysis, we target automated and semi-automated bots (group A) and compare their activities across the discussions we analyze against that of genuine users (group C). In particular, in our analysis of inauthentic activities in Farsi Twitter, group A serves as the target group, group C as the observation group, and group B is primarily a buffer zone which provides a safety margin separating the inauthentic automated agents from authentic accounts. Although in principle one could study group B as a target group itself and such a study could have useful implications, that does not fit within the objectives of this paper, and we do not conduct a substantial analysis of activities of users in this group.

Using bot-detection methods systematically on big data for mapping users to one of these categories is a challenging task, given the error in bot detection. As the Botometer team warn users of their bot detection tool, Botometer is rather a complement for human judgement and cannot be relied on for classifying a Twitter account as bot or human[41]. We, however, verify that, in average, Botometer provides a working estimate that can be used to group users by the types defined in this section. Figure 12 shows the distribution of CAP in each of the three groups, with group B divided into its two subgroups, for manually annotated accounts. A wide range of CAP is seen for users in all groups, and significant overlap confirms the anticipated misjudgements and errors. Nevertheless, both mean and median decrease from automated and semi-automated bots to genuine users, which indicates that, at an aggregate level, we can obtain meaningful estimates from Botometer on the behavior of users with respect to the criteria related to opinion manipulation. Therefore, although using Botometer for micro-scale analysis of bot-like behavior or labeling of individual accounts according to our typology should be strictly avoided, Botometer can be used to make meaningful implications for macro-scale analysis of opinion manipulation in Twitter when dealing with big data.

Temporal analysis. In order to study the emergence of a trend in Farsi Twitter, we extract time-series data from the tweets that would help us analyze the dynamics of the discussion space. This is done through dividing the time interval where the participation in the discussion is significantly higher than before or after that interval into a number of intervals of equal length (24 hours) and binning the data into these smaller intervals. The data in each time interval is further binned into three user types described in the "User typology" subsection, using their CAP scores. We then study the dynamics of the share of each user type in the discussion over time by com-puting number of tweets tweeted by users from each type within each time interval. The results, discussed in the
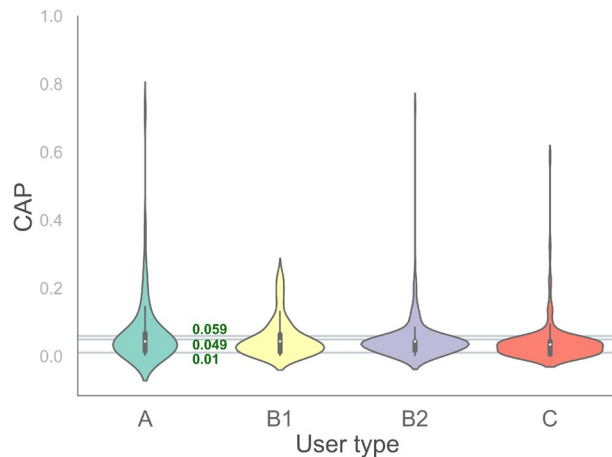
Figure 12. The distribution of CAP in each group of 668 manually labeled users. The separation lines are at CAP values 0.010, 0.049, and 0.059.

Participation in Trending "Discussions" subsection, reveal qualitative differences between apolitical and divisive political discussions with respect to the temporal dynamics of participation in the discussion.

Network analysis. We form two types of networks for each discussion: friendship network, and retweet network. In both networks the nodes are the users participating in the discussion, i.e. tweeted or retweeted a post containing a corresponding keyword. In a friendship network there is an edge from user j to user i if j follows i, while in a retweet network, a directed edge (j, i) indicates that j has (at least once) retweeted i. Both networks are directed, however, depending on our analysis, we may ignore the direction of the edges. For each topic, both types of networks are restricted to the corresponding discussion. For analyzing communities in our networks, we use the Louvain clustering algorithm[42] to detect communities and bin the relevant data by the induced subgraphs. Note that different runs of the Louvain algorithm could yield different results in a fixed number of iterations. Considering the use of community structures in our data analysis, we perform a sanity check, computing similarities between several random runs of the Louvain clustering for each network in our dataset, which is included in the Supplementary Information accompanying this paper. Our validation analysis reveals that different runs of the clustering algorithm consistently output highly similar community structures, which confirms the validity of our analysis and interpretations, considering that we only draw aggregate-level conclusions about our observations.

Although retweet networks are more commonly studied for analyzing Twitter discussions, the structure of the friendship network of participants in a discussion could be indicative of various qualitative properties of the discussion space (see e.g. the study by Gonçalves et al.[43]). Importantly, the friendship network is the primary medium for circulation of content, and as such, its structural properties could reveal potential patterns of exposure to content, such as conduciveness to polarization and echo chambers, as we discuss in the "Differences in content" subsection. Therefore, analyzing the structure of friendship networks can help us connect the content of discussion to the structure of the corresponding network. We discuss an example of how this connection can be studied from a joint analysis of friendship networks and the text of tweets in the "Differences in content" subsection, where we analyze word clouds and topics corresponding to different friendship communities. Additionally, since the retweet network is formed through resharing tweets, which is an action users take considering the content of the tweets, the retweet network is inherently content related. Thus, a comparative analysis of the retweet and friendship networks gives a clearer picture of the relationship between the content and structure of discussion spaces in Twitter, as it is discussed in the Inauthentic Activity in Echo Chambers subsection.

In order to perform the comparison between retweet and friendship networks, we compare the community membership of users with respect to the clustering of each network type. In a network with segregated communities, the adjacency matrix could be represented as a block-diagonal matrix through a permutation of rows and columns such that nodes belonging to the same community are grouped together. In any given network, the more segregated the communities are, the closer will such a permutation of the adjacency matrix look to a block-diagonal matrix, with denser diagonal blocks and fewer non-zero off-diagonal entries. We use this observation in Fig. 9 to visually compare the community structure of the retweet networks with the partitioning induced by friendship communities. Moreover, we use two clustering similarity measures—Jaccard index and element-center clustering similarity[23,44]— to quantify this observation. To assess the similarities of the retweet and friendship communities, we first form an ensemble of 20 clusterings for each network, obtained from 20 runs of the Louvain clustering algorithm using different random seeds. This yields 400 pairwise similarity values for each discussion space, with respect to each of the two similarity measures, which we then pool by discussion type to compare apolitical and divisive political discussions. The distributions of the similarity values allow us to draw conclusions on the structural potential of the retweet and friendship networks for yielding similar or dissimilar community structures, as we discuss in the Inauthentic Activity in Echo Chambers subsection. Performing this analysis over multiple runs of the clustering algorithm allows us to comment on the overlap between retweet

and friendship communities despite the randomness in the algorithm and sensitivity of the similarity measures to changes in community memberships.

In order to further quantify the polarized structure of the networks, we compute network modularities[24] with respect to the communities obtained from 100 runs of the Louvain clustering algorithm, to obtain a distribution of modularity values for the networks in each type of discussion. The modularity of a network with adjacency matrix A is defined as follows

$$Q = \frac{1}{2m} \sum_{(v,u) \in V \times V} \left[A_{v,u} - \frac{k_v \, k_u}{2m} \right] \delta_{c_v, c_u}$$

where m is the number of edges in the network, V is the set of nodes, k_v is the degree of node v, c_v is the cluster node v belongs to, and δ_{c_v, c_u} is the Kronecker delta. Modularity is a commonly used measure for evaluating the degree of community seggregation and potential for polarization in social networks[25,26]. As we discuss in the Inauthentic Activity in Echo Chambers subsection, our assessment of the distribution of modularity values confirms our findings from analyzing the overlap between retweet and friendship communities.

In addition to the communities, we perform two additional analyses of the meso-scale structure of the networks using core-periphery and bow-tie structures. Previous studies point to a connection between the core-periphery structure and propogation of content on social networks[20,45]. In the core-periphery analysis, the users are partitioned into two groups of core—a densely connected cohesive block of nodes—and periphery—a sparser and relatively more distant set of nodes[46]. We use the algorithm proposed by Kojaku and Masuda[47], which allows for detecting multiple cores and peripheries within the same network, in order to find the core-periphery structure of each network in our dataset. The bow-tie structure on the other hand, yields an alternative partitioning of the nodes in directed networks into seven components[21,48,49]. Recent studies have found connection between the bow-tie structure and discursive communities in online social networks[22]. We use three components of interest: A strongly connected component ('S'), the nodes at the tail of out-going edges to the S component which act as its 'IN' gate, and the nodes at the head of in-coming edges from the S component which act as its 'OUT' gate. In our friendship networks, the 'S' component is a densely connected group where there exists a directed path in both directions between every pair of nodes, and as such, it is particularly important for robust propagation of content. The IN and OUT components on the other hand could be thought of as primary consumers and sources of content in a friendship network, given that an edge in our friendship networks runs in the direction of exposure, from follower to friend.

Content analysis. In order to get an overall view of the dominant content, we plot word clouds visualizing frequent words in the tweets. To construct the word clouds, we first process the text of the tweets and apply appropriate filters. We partition the words pool obtained through this process by subpopulations of interest in order to compute word counts for each subpopulation. For demonstrating differences between the tweet contents of subpopulations i and j, we plot the word cloud corresponding to words that are among the top 5% most frequent words in i but not among the top 10% of j, and vice versa. This allows us to observe existing differences between the dominant content of tweets posted by users belonging to different subpopulations. As discussed in the "Differences in content" subsection, this analysis is performed once when the grouping of users is done by user types, and once when users are grouped by friendship network communities.

In the interpretation of the results described in the "Differences in content" subsection, we use the word clouds corresponding to tweets by groups of users as a proxy for the common concern and vocabulary of that group (see[50] for an example of a similar approach). To further investigate common concerns, we detect salient topics in the set of tweets from each group of users, and consider the most representative words among the top-4 most salient topics in each discussion space. In order to obtain the topics and their representative words from our tweets, we use BERTopic[51], an unsupervised transformer-based topic modeling tool, which has been shown to be well-suited for topic modeling on Twitter posts[52]. Note, as we explain in the "Differences in content" subsection, when comparing the content of tweets of users in different friendship communities, the network clustering is solely based on the structure of friendship networks, while the text analysis is blind to the network structure. Thus, deviation from randomness in the results is an indication of the presence of a relationship between content and structure. Our results, described in the "Differences in content" subsection, further suggest that such deviations relate to the topic of the discussion and the position of the participants on the relevant socio-political spectrum.

Data availability

Encrypted tweet IDs are available at https://github.com/afarzam/FarsiTwitter_code/tree/main/tweetIDs and the CAPs for encrypted user IDs are available at https://github.com/afarzam/FarsiTwitter_code/tree/main/CAPs along with the encrypted keys and Python programs for decryption. The data collected for this project is subject to Twitter policies regarding data collection for academic research. Moreover, for privacy concerns, we are unable to make our dataset available in its original format. Once these considerations are realized, we may be able to provide the private keys for decryption upon request and further consultation with Twitter. The code for reproducing the main figures and the preceding data preparation is available at https://github.com/afarzam/FarsiTwitter_code.

References

1. Diamond, L. Liberation technology. *J. Democr.* **21**, 69–83 (2010).
2. Ketabchi, K., Asadpour, M. & Tabatabaei, S. A. Mutual influence of twitter and postelection events of Iranian presidential election. *Procedia Soc. Behav. Sci.* **100**, 40–56 (2013).
3. Thieltges, A., Papakyriakopoulos, O., Serrano, J. C. M. & Hegelich, S. Effects of social bots in the iran-debate on twitter. http://arxiv.org/abs/1805.10105 (2018).
4. Ananyev, M. & Sobolev, A. Fantastic beasts and whether they matter: Do internet "trolls" influence political conversations in russia. In *Meeting of the Midwest Political Science Association* (2017).
5. Bessi, A. & Ferrara, E. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday* **21**. https://doi.org/10.5210/fm.v21i11.7090 (2016).
6. Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M. & Saracco, F. The role of bot squads in the political propaganda on twitter. *Commun. Phys.* **3**, 1–15 (2020).
7. Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, 21–30 (2010).
8. Forelle, M., Howard, P. N., Monroy-Hernández, A. & Savage, S. Political bots and the manipulation of public opinion in venezuela. SSRN 2635800 (2015).
9. Gorwa, R. & Guilbeault, D. Unpacking the social media bot: A typology to guide research and policy. *Policy Internet* **12**, 225–248 (2020).
10. Howard, P. N. & Kollanyi, B. Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. SSRN 2798311 (2016).
11. Shao, C. *et al.* The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 1–9 (2018).
12. Schäfer, F., Evert, S. & Heinrich, P. Japan's 2014 general election: Political bots, right-wing internet activism, and prime minister shinzō abe's hidden nationalist agenda. *Big Data* **5**, 294–309 (2017).
13. Tucker, J. A. *et al.* Social media, political polarization, and political disinformation: A review of the scientific literature. in *Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018)* (2018).
14. Woolley, S. C. Automating power: Social bot interference in global politics. *First Monday* (2016).
15. Barrie, C. & Siegel, A. Kingdom of trolls? influence operations in the saudi twittersphere. *J. Quant. Descrip. Dig. Media* **1**. https://doi.org/10.51685/jqd.2021.012 (2021).
16. Davis, C. A., Varol, O., Ferrara, E., Flammini, A. & Menczer, F. Botornot: A system to evaluate social bots. in *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274 (2016).
17. Varol, O., Ferrara, E., Davis, C., Menczer, F. & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11 (2017).
18. Yang, K.-C. *et al.* Arming the public with artificial intelligence to counter social bots. *Hum. Behav. Emerg. Technol.* **1**, 48–61 (2019).
19. Hodges, J. L. The significance probability of the smirnov two-sample test. *Ark. Mat.* **3**, 469–486 (1958).
20. Barberá, P. *et al.* The critical periphery in the growth of social protests. *PLoS ONE* **10**, e0143611 (2015).
21. Broder, A. *et al.* Graph structure in the web. *Comput. Netw.* **33**, 309–320 (2000).
22. Mattei, M., Pratelli, M., Caldarelli, G., Petrocchi, M. & Saracco, F. Bow-tie structures of twitter discursive communities. http://arxiv.org/abs/2202.03316 (2022).
23. Gates, A., Wood, I., Hetrick, W. & Ahn, Y. Element-centric clustering comparison unifies overlaps and hierarchy. *Sci. Rep.* **9**, 8574 (2018).
24. Newman, M. E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**, 8577–8582 (2006).
25. Conover, M. *et al.* Political polarization on twitter. *Proc. Int. AAAI Conf. Web Soc. Media* **5**, 89–96 (2011).
26. Livne, A., Simmons, M., Adar, E. & Adamic, L. The party is over here: Structure and content in the 2010 election. *Proc. Int. AAAI Conf. Web Soc. Media* **5**, 201–208 (2011).
27. Weedon, J., Nuland, W. & Stamos, A. Information operations and facebook. *Retrieved from Facebook:* https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf *(2017).*
28. Schliebs, M., Bailey, H., Bright, J. & Howard, P. China's public diplomacy operations: Understanding engagement and inauthentic amplifications of prc diplomats on facebook and twitter (2021).
29. Gupta, A. & Kaushal, R. Towards detecting fake user accounts in facebook. in *2017 ISEA Asia Security and Privacy (ISEASP)*, 1–6 (IEEE, 2017).
30. Smith, M. A., Rainie, L., Shneiderman, B. & Himelboim, I. From polarized crowds to community clusters. in *Policy Commons, Mapping twitter topic networks* (2014).
31. PostgreSQL, G. Postgresql, global development group. http://www.postgresql.org (2008).
32. Hipp, R. *et al.*. Sqlite (Version 3.8.10.2) [Computer Software]. *Sqlite Development Team*. (2015)
33. Kendall, M. G. *Rank Correlation Methods* (Griffin, 1948).
34. Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. *IEEE Trans. Depend. Secure Comput.* **9**, 811–824 (2012).
35. Subrahmanian, V. S. *et al.* The darpa twitter bot challenge. *Computer* **49**, 38–46 (2016).
36. 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 *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972 (2017).
37. Luceri, L., Deb, A., Giordano, S. & Ferrara, E. Evolution of bot and human behavior during elections. *First Monday* (2019).
38. Mittal, S. & Kumaraguru, P. Broker bots: Analyzing automated activity during high impact events on twitter. http://arxiv.org/abs/1406.4286 (2014).
39. Blum, J., Cyphers, B., Nachum, O., & Sobel, L. Finding the bots in the flock (Tweetnet, 2014).
40. Echeverria, J. & Zhou, S. Discovery, retrieval, and analysis of the'star wars' botnet in twitter. in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 1–8 (2017).
41. Botometer. Faq retrieved from.
42. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008 (2008).
43. Gonçalves, B., Perra, N. & Vespignani, A. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS ONE* **6**, e22656 (2011).
44. Gates, A. J. & Ahn, Y.-Y. Clusim: A python package for calculating clustering similarity. *J. Open Source Softw.* **4**, 1264 (2019).
45. Saxena, A., Iyengar, S. & Gupta, Y. Understanding spreading patterns on social networks based on network topology. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 1616–1617 (2015).
46. Borgatti, S. P. & Everett, M. G. Models of core/periphery structures. *Soc. Netw.* **21**, 375–395 (2000).
47. Kojaku, S. & Masuda, N. Finding multiple core-periphery pairs in networks. *Phys. Rev. E* **96**, 052313 (2017).
48. Yang, R., Zhuhadar, L. & Nasraoui, O. Bow-tie decomposition in directed graphs. in *14th International Conference on Information Fusion*, 1–5 (IEEE, 2011).
49. van Lidth, J. Bow-tie_detection. https://github.com/jeroenvldj/bow-tie_detection (2019).

50. Amor, B. R. C. *et al.* Community detection and role identification in directed networks: understanding the twitter network of the care data debate. *CoRR* **abs/1508.03165** (2015).
51. Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. http://arxiv.org/abs/2203.05794 (2022).
52. Egger, R. & Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **7** (2022).

Acknowledgements

Author contributions

A.F. and P.M. designed research and all authors contributed to conceptualization; P.M. designed and developed the data collection machine; P.M. and A.F. collected data; A.F., P.M., Z.P. and S.M. curated data and implemented the computer codes; A.F. designed the methodology and performed the data analysis; A.F. and S.M. prepared and visualized data; All authors contributed to the interpretation of results; A.F. and A.S. wrote the main body of the manuscript and all authors contributed to its writing and preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-26921-5.

Correspondence and requests for materials should be addressed to A.F. or A.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

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