CEUR-WS.org/Vol-2890/paper2.pdf

Can Celebrities Burst Your Bubble?[⋆]

Tuğrulcan Elmas[0000-0002-4305-1479], Kristina Hardi[0000-0001-6305-600X],
Rebekah Overdorf[0000-0003-3462-9539], and Karl Aberer[0000-0003-3005-7342]

EPFL, Switzerland
{firstname.lastname}@epfl.ch

Abstract. Polarization is a growing, global problem. As such, many social media based solutions have been proposed to try to reduce it. In this study, we propose a new solution that recommends topics to celebrities to encourage them to join a polarized debate and increase exposure to contrarian content — bursting the filter bubble. Using a state-of-the art model that quantifies the degree of polarization, this paper makes a first attempt to empirically answer the question: *Can celebrities burst filter bubbles?* We use a case study to analyze how people react when celebrities are involved in a controversial topic and conclude with a list possible research directions.

Keywords: filter bubble · polarization · twitter

1 Introduction

Polarization is a state in which the public is divided into groups with opposing opinions on an issue [4]. Polarization is regarded as a threat to democracy and is detrimental to healthy dialogue in a community. Echo chambers — the phenomena in which individuals only hear the side of a debate they already agree with — are a primary driver of polarization, as they are where extremist ideas foster [18]. Social media platforms themselves hold some responsibility for the formation of such echo chambers; the algorithms that determine the information diet of users are believed to rank belief-reinforcing information higher as a result of maximizing engagement, which in turn minimizes cognitive dissonance. The term *filter bubble*, coined by Eli Pariser [17], describes this phenomenon by which echo chambers are caused by the design of the system.

Due to their potential detriment to democracy and society, others have proposed methods to burst filter bubbles in order to reduce polarization. Many of these solutions rely on action by the social media platforms, ignoring the fact that social networks have created this problem and may not be incentivised to act, including recommending users [8] (aided by intermediary topics [10]) or content [13] with opposing opinions, and presenting the information in a different

[⋆] Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). Presented at the MISINFO 2021 Workshop, held in conjunction with the 30th ACM The Web Conference, 2021, in Ljubljana, Slovenia.

way (i.e. showing the credibility of a source) [21]. Other work focuses on raising awareness to users and therefore requires action on the part of the users who are in a filter bubble. These include exposing users to contrarian news [7], raising awareness of one's connections' and own biases [6], convincing some users in the social network to reduce the overall polarization via education [14].

We present a new recommendation scheme that requires neither buy-in from the social network nor action on the part of those in the filter bubble. This scheme bursts filter bubbles by recommending polarizing topics to influential users, i.e. celebrities. Prior work has shown that celebrities increase exposure and influence opinion on controversial subjects like vaccination [2], suggesting that their involvement in a debate can reduce polarization by means of exposing users to counter opinions. Other work has shown that users value connections while selecting content [15], so messages conveyed by celebrities whom users are connected to, are likely to be valued over content from non-connections. Finally, this method leverages the fact the social media is a small-world network [22] so people with counter opinions are connected to the same users who do not explicitly posit their opinions. For example, LeBron James is both followed by both liberals and conservatives and also is the only liberal source in most of his conservative followers' profiles and therefore can burst those users' bubbles [12].

We identified the following research questions related to this scheme and answer them in the remainder of this paper:

RQ1 If a celebrity joins a debate on a controversial topic will exposure to the contrarian content be increased and will polarization be reduced?
RQ2 How to select celebrities that would lead to such effects?
RQ3 Will such exposure mitigate the extreme opinions and hence decrease polarization?
RQ4 How would people react when a celebrity joins the debate? Will we observe mitigation of thoughts or backfire effect?

We define celebrity as "anyone popular and although not strictly impartial, not politically polarized." To address **RQ1** and **RQ2**, we empirically show that the inclusion of popular and neutral nodes into a polarized graph decreases the polarization of the graph, hence celebrity inclusion reduces polarization. To address **RQ3** and **RQ4**, we perform a qualitative analysis of users' reactions to celebrities participating in a controversial topic.

2 Empirical Results on Effect of Celebrity Involvement

2.1 Theoretical Background

First, we address **RQ1** to determine the effect that a celebrity joining a debate will have on polarization. We use the quantifying controversy model [9] which quantifies the controversy of a topic by computing how likely a user on one side of a polarized debate is to be exposed to content disseminated by a popular user on the opposing side. As such, this model serves as a proxy for polarization score on a topic. We recap the model briefly before detailing our application of it.

Quantifying Controversy Model First, consider a social graph $G(V, E)$ in which vertices V are users who hold an opinion on a topic and edges E are the social connections between them. G is partitioned into two disjoint sets of users, X and Y, which possibly correspond to the two different sides of the discussion. For each node in a set of randomly selected nodes, a random walk is started and concludes when it reaches any high-degree user. Let P_{AB} be the probability that a random walk begins in partition A and ends in partition B. The "Random Walk Controversy Score" (RWC) is the difference of the probabilities that a random walk begins and ends in the same partition (P_{XX} and P_{YY}) and that a random walk begins and ends in the other partition (P_{XY} and P_{YX}).

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX}$$

The resulting RWC score is inversely correlated with the likeliness of exposure to popular content from the opposite side and implies polarization of the debate on the topic.

Since the users in the same partition are well connected due to the homophily principle, we assume that in a polarized network content produced in the same partition has the same stance. Conversely, content produced by users in different partitions have different stances. Thus, in a polarized network, content that is quickly reached by a user (via a random walk) is from the same stance as the user. Hence, the user is trapped in an echo chamber. We leave a model for echo chambers that works without this assumption about stances to future work.

In the context of Twitter, the topic is modeled as tweets containing relevant hashtags to a seed hashtag that defines the topic. The social network, G, is built by including users who authored these tweets. The links between the nodes of these networks could be following, retweeting, or both. We use following relationships as they are a better proxy to measure exposure to content from the connected user. As such, G is directed and users receive incoming links from their followers. The users who are recommended the topic will be added to the G if they accept the recommendation. Links between users already in G and a newly added user will be added to G if these users already follow the newly added user. Our problem is then to identify such nodes to recommend the topic so that their inclusion in the network will decrease RWC of G.

One issue with this topic modeling approach is that it draws mostly from politically interested users and hence exaggerates the polarization of a popular topic that involves many hashtags and keywords [19]. However, this approach is plausible when you consider the scenario in which a user clicks on a hashtag about a controversial topic that is trending. We assume that the tweets from users' connections are more likely to be ranked higher and hence the user will be in a filter bubble when presented with tweets on that topic.

2.2 Node Addition Problem

In order to determine which celebrities to select (**RQ2**), we define the Node Addition Problem as determining which nodes to add to the network in order to maximize the reduction of the controversy of the topic. Consider a topic T and a social graph $G = (V, E)$ made up of users who participated in a debate about T. Let the controversy score of G be $RWC(G)$. Let the *Potential Social Graph* $G^{**}(V^{**}, E^{**})$ be the union of $G = (V, E)$ and all the vertices that are connected to V but did not discuss T and the edges connecting them to V. The node addition problem is to find a set of k nodes V' not in G but in G^{**} to add to G and obtain the *Augmented Graph* $G^* = (V^*, E^*)$ which maximizes $RWC(G)$ - $RWC(G*)$.

We hypothesize that k nodes that maximize the decrease of RWC will be those who have 1) high in-degree 2) edges distributed evenly between two partitions. We leave mathematical proof for future work and only present empirical results. To find those k nodes, we use the Fagin algorithm [5] to rank nodes by 1) their in-degree and 2) their minimum ratio of connections' to one partition over all connections. The first is a proxy for popularity and the second a proxy for neutrality. We compute RWC for each candidate and choose the k nodes which yield the largest RWC decrease. We assume these k nodes will consist of candidates which minimize RWC individually to avoid computing RWC for every k combination of nodes, which is very costly.

2.3 Experimental Results

For the empirical results, we used the follower data of users who participated in the debate about the topic #Russia_March. This topic is studied in [9] and is already found to be polarized. We first created the follower graph G which involves only the users who tweeted with #Russia_March and relevant hashtags. Then we collected all the followees of users in the G to create the augmented graph. See Figure 1 for the two graphs.

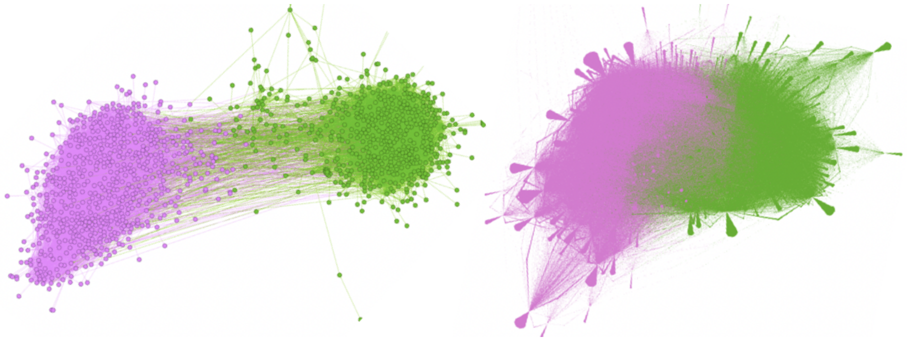

Fig. 1. Social Graph (left) and Potential Social Graph (right). The nodes are users and the edges indicate follows. The colors indicate partitions. Force Directed Layout is used to visualize graphs. Notice that Potential Social Graph is not polarized like Social Graph.

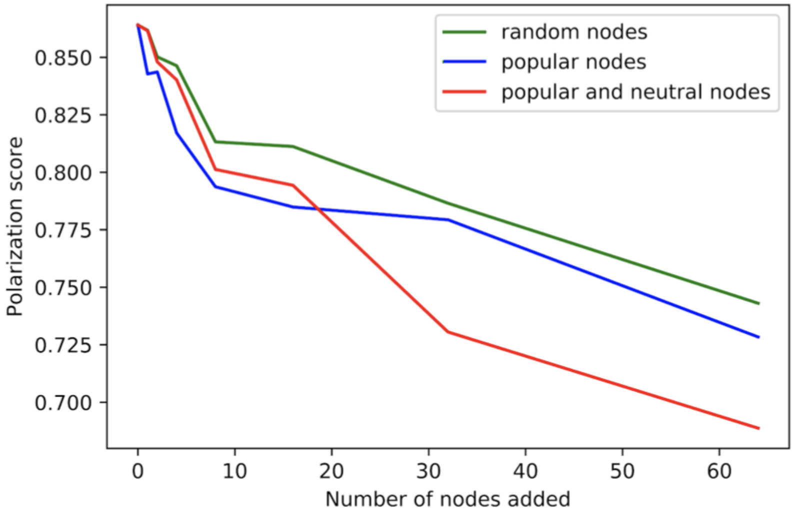

Fig. 2. Decrease in polarization score according to the number of nodes added. Colors denote the method to choose the nodes to be added. Notice that choosing popular and neutral nodes to recommend a topic is much more effective than merely choosing popular nodes after 20 additions.

We used two baselines to evaluate our node selection process. First, the most popular nodes to study the effect neutrality and second "random nodes", which are artificially created nodes that have fixed degree (50) and are connected to 25 randomly chosen nodes in each partition to study the effect of popularity. Figure 2 shows the final polarization score with respect to the number of nodes added. Colors denote the method used to choose nodes. Although adding popular nodes seems beneficial initially, they become ineffective after 20 nodes. As the results indicate, most popular and neutral nodes reduce polarization, who happen to be celebrities by our definition.

A possible side effect of this method is that users will unfollow celebrities who discuss controversial topics and join the debate. To simulate this, we randomly break incoming links of the k nodes. As seen in Figure 3, the polarization is still reduced unless the celebrities lost most of their followers, which is unrealistic.

The empirical results show that polarization as measured by RWC reduces when celebrities join a controversial debate. It is not clear, however, how much reduction in RWC equates to real-life implications. In the next section, we explore what happens in a real-life study of celebrities weighing in on a controversial topic.

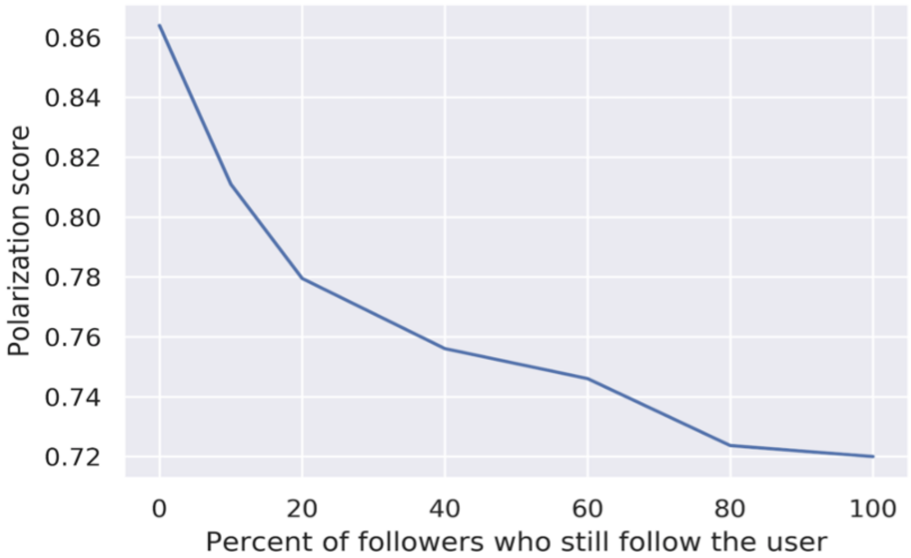

Fig. 3. Simulation of a scenario where users unfollow the celebrity who joined the debate. The polarization is still reduced effectively even when 20% of the users still follow the celebrity.

3 The Case of 2019 İstanbul Election Rerun Decision

The 2019 İstanbul Election Rerun was a controversial decision by the Supreme Electoral Council when the opposition's candidate won, overruling the AKP's candidate by a slim margin despite high voter turnout. The decision was deemed unfair by supporters of the opposition. Many celebrities started to tweet after the opposition candidate, Ekrem İmamoğlu, gave a speech and said "Everyone should speak, the celebrities should speak!" [11]. This serves as a suitable case study as those celebrities' messages reached a very wide audience; many have more than 100,000 followers.

We address **RQ4** and determine how users react to celebrities joining a debate by analyzing immediate reactions to the celebrity tweets regarding this decision. We also address **RQ3** on this real-world data set and determine via a longitudinal study if this celebrity intervention actually made a difference in this case.

We selected 81 celebrities from a list of Turkish celebrities [1] who tweeted in favor of the opposition's candidate on the night of 6 May 2019. We collected their tweets, retweets, and replies to their tweets for one week using Twitter's Streaming API. By manual inspection, we found that 47 of them are in cinema-tv business, 24 in music, and 10 in other fields. Judging from the tweets since September 2019, 43 celebrities were already found to be posting frequently about controversial but apolitical topics and 7 were found to be occasionally posting about such topics. The remaining 31 only post personal and professional updates.

No celebrity showed explicit political affiliation or criticism towards a party or government, but 25 criticized recent government policies, 15 infrequently posted content that could be interpreted as anti-government, and the remaining 41 appear politically neutral. This suggests that for most of the celebrities in our dataset, the İstanbul Election Rerun was the first time they spoke out on a political topic on Twitter.

For those celebrities with more than one relevant tweet, we selected the tweet that received the most replies for each celebrity in 6-7 May 2019, then annotated each celebrity according to their stance according to that tweet. Note that 60 of the celebrities showed explicit support to the opposition's candidate (30 used the opposition slogan #HerŞeyÇokGüzelOlacak), while 8 celebrities only commented on the unfairness of the decision of rerun. In addition, 11 celebrities called for citizens to vote in the rerun, and 2 called for other celebrities to tweet.

We randomly sampled 10 direct replies per celebrity tweet. We annotated these replies according to 1) stance on the celebrity (positive, negative, neutral) and 2) narratives they contain. Note that not all replies had a narrative. We removed the celebrity tweets that were irrelevant or had less than 10 replies. We annotated 679 tweets in total. We found that 434 tweets had a positive stance towards the celebrity and their idea, 178 tweets had a negative stance (with 31 containing insults), and 60 tweets had a neutral stance. Our analysis indicates the following narratives are prevalent in tweets, which have a non-positive stance unless otherwise specified.

1. Counter argument: The celebrity is wrong as the opposition has committed voter fraud and the decision was correct. (n = 29)
2. Ad hominem: The celebrity is wrong or does not deserve a voice on the matter due to their past political actions, or their character is not harmonious with their idea. (n = 26)
3. Self-interests: The celebrity is behaving this way not because of patriotism but for self interest because they are not successful in their work. (n = 19)
4. Questioning authority: The celebrity does not have a right to speak because they are not a real celebrity. (n = 16)
5. Whataboutism: The celebrity's patriotism is in question as they did not react to soldiers killed by terrorist attacks or on the night of the July 15, 2016 coup. (n = 11)
6. Too late: Positive with the celebrity's opinion but blames them or celebrities in general for acting too late. (n = 10)
7. Reactionary: The celebrities should not expose their political beliefs or champion one political side or should be remembered with their art only. (n = 5)
8. Hopeless: The situation is hopeless and they will not win the rerun, although the celebrity is championing hope. (negative: n = 3), (positive: n = 2)
9. Mitigate: Indicates a non-polarized affiliation, but agrees with the celebrity on that issue with positive stance. (n = 2)
10. Backfire: Threatens the celebrity (n = 9), or indicates they will no longer follow them (n = 3).

Based on this analysis, we make the following observations:

The celebrities' messages reached users with the opposite stance: The presence of negative replies from the opposite side shows that the goal of bursting the filter bubbles has been achieved.

The source of the message matters: if the celebrity is not having a successful career and is not respected, they have negative reactions implying those elements. Even the supportive replies had sarcastic elements in some cases. Their past political deeds also matter especially if they took a pro-government stance before.

The content of the message also matters: if the content is sarcastic or has some logical flaw, the replies indicate this rather than agreement or disagreement which causes distraction. The reactions which would lead to meaningful discussions (although still rare) come when the celebrity's tweet contains an argument.

There is no evidence of correlation between political activity on Twitter and stances of replies a celebrity gets: We averaged the stances of replies each celebrity gets (1 for positive, 0 for neutral and -1 for negative.) and ran a t-test for those who were political / commented on contemporary issues on Twitter and those who do not. The stance of replies turned out to be independent of both factors as the p-value was insignificant.

The negative replies mainly come from politically motivated users: We inspected 100 users who had a negative stance and replied to celebrities. Three of these users had deleted their accounts and 15 were suspended. Among the 82 remaining users, 55 were very polarized — their account seemed to be opened only to share pro-AKP content, and they constantly spread fake news about the opposition. This suggests that the replies should not be taken as a genuine public reaction during analysis as they also likely part of coordinated attacks. However, the narratives they contain are still important as they may influence genuine Twitter users.

Both mitigation and backfire effect appear to be small: Inferring from the reactions, we had only two cases where an artist's presence made a positive effect. The backfire effect is also small; follower counts increased rather than decreased, which may show that they do not go out of favor dramatically.

4 Open Questions

Celebrity acceptance: Would a celebrity accept the recommendation to join the debate 1) by public request, 2) by platform request, 3) by a fellow celebrity, or 4) not at all?

Factors on user's reactions: Are user's reactions to celebrities joining a political debate dependant on the side they join, on whether they try to mitigate extreme opinions, or whether multiple celebrities observe the same behavior?

Modification of the platform: What would be the effects of a modification to the platform so that 1) it recommends topic to users that would increase exposure to the contrarian content and decrease polarization and 2) it recommends content by such users to users with extreme views?

Categorization of celebrity candidates: Not all popular accounts are suitable to recommend topics to comment on, i.e. corporate and media related accounts may be less likely to take the recommendation for fear of backlash.

When do celebrities weigh-in on a controversial topic? Do they join in early and help the topic spread or do they join later? Is it due to peer-pressure, self-interest, or neither? Such analysis would be useful in determining if recommending topics to celebrities is realistic or helpful.

Non-political users: If many celebrities are tweeting about political topics due to this method, users who use Twitter for entertainment purposes and not for political engagement may be negatively impacted or leave the platform.

Revision of quantifying polarization: Current algorithms do not scale, do not take temporal signals into account, and do not take graph modularity due to factors like language into account.

RT = Endorsement? Quantifying polarization studies assume that social connections like retweets and follows are endorsements without justification. However, this assumption often falls apart in real-world applications. Many users even *explicitly state* that retweets are not endorsements on their profiles. In some cases, endorsement occurs without a like: videos involving #BLM (Black Lives Matter) protests and police intervention on Facebook were not liked but shared [20]. In Twitter terms, this would mean that newsworthy posts with a negative sentiment are not liked but retweeted, which breaks this assumption. A survey among 316 users revealed that only 68% of people endorse what they retweet, and 73% of users agree with what they retweet [16]. Thus, coming up with better connection models is needed.

Revision of identification of a topic: Hashtags do not capture all the discussion on a topic and focus attention on already polarized users, thus creating biased results. Therefore, the methodology to model a topic should be revised.

Backfiring effect: Anti-polarization tools assume that views will be moderated when a user is connected to users of opposite views by the implicit assumption that views will be averaged, ignoring the possibly backfire effect.

Universal Interest: There is an underlying assumption that an unbiased user has a medium opinion. Most works do not consider that a user may have no opinion on a topic.

Lurkers Matter: The observations from Twitter analysis are based only on the audience that actively reacts. However, Facebook users were found to underestimate their audience by 27% [3]. We expect a more dramatic result on Twitter since most profiles are public and timelines are created based on more than simple follow relationships. Thus, future work is needed to verify these results.

Not all views should be moderated:. In fact it can be harmful in some cases to encourage users towards some position. For example, encouraging normal users to read anti-vaccination content could be detrimental to public health.

References

1. 6 mayıs 2019 imamoğlu lehine tweet atan ünlüler. https://eksisozluk.com/6-mayis-2019-imamoglu-lehine-tweet-atan-unluler–6030169?a=nice, accessed: 2019-05-06
2. Alatas, V., Chandrasekhar, A.G., Mobius, M., Olken, B.A., Paladines, C.: When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Tech. rep., National Bureau of Economic Research (2019)
3. Bernstein, M.S., Bakshy, E., Burke, M., Karrer, B.: Quantifying the invisible audience in social networks. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 21–30. ACM (2013)
4. DiMaggio, P., Evans, J., Bryson, B.: Have american's social attitudes become more polarized? American journal of Sociology **102**(3), 690–755 (1996)
5. Fagin, R., Lotem, A., Naor, M.: Optimal aggregation algorithms for middleware. Journal of computer and system sciences **66**(4), 614–656 (2003)
6. Gao, M., Do, H.J., Fu, W.T.: Burst your bubble! an intelligent system for improving awareness of diverse social opinions. In: 23rd International Conference on Intelligent User Interfaces. pp. 371–383. ACM (2018)
7. Garimella, K., De Francisc iMorales, G., Gionis, A., Mathioudakis, M.: Mary, mary, quite contrary: Exposing twitter users to contrarian news. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 201–205. International World Wide Web Conferences Steering Committee (2017)
8. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Reducing controversy by connecting opposing views. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 81–90. ACM (2017)
9. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. ACM Transactions on Social Computing **1**(1), 3 (2018)
10. Graells-Garrido, E., Lalmas, M., Quercia, D.: People of opposing views can share common interests. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 281–282 (2014)
11. Gumrukcu, T.: Turkish artists stand by istanbul's ousted mayor. Reuters (2019), https://www.reuters.com/article/us-turkey-election-celebrities/turkish-artists-stand-by-istanbuls-ousted-mayor-idUSKCN1SG1S7
12. Hayat, T., Galily, Y., Samuel-Azran, T.: Can celebrity athletes burst the echo chamber bubble? the case of lebron james and lady gaga. International Review for the Sociology of Sport p. 1012690219855913 (2019)
13. Lex, E., Wagner, M., Kowald, D.: Mitigating confirmation bias on twitter by recommending opposing views. arXiv preprint arXiv:1809.03901 (2018)
14. Matakos, A., Terzi, E., Tsaparas, P.: Measuring and moderating opinion polarization in social networks. Data Mining and Knowledge Discovery **31**(5), 1480–1505 (2017)
15. Messing, S., Westwood, S.J.: Selective exposure in the age of social media. Communication Research **41**, 1042–1063 (2014)
16. Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., Finn, S.: What do retweets indicate? results from user survey and meta-review of research. In: Ninth International AAAI Conference on Web and Social Media (2015)
17. Pariser, E.: The filter bubble: What the Internet is hiding from you. Penguin UK (2011)
18. Sunstein, C.R.: Republic. com. Princeton university press (2001)
19. Tufekci, Z.: Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)

20. Tufekci, Z.: Twitter and tear gas: The power and fragility of networked protest. Yale University Press (2017)
21. Vydiswaran, V.V., Zhai, C., Roth, D., Pirolli, P.: Overcoming bias to learn about controversial topics. Journal of the Association for Information Science and Technology **66**(8), 1655–1672 (2015)
22. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. nature **393**(6684), 440 (1998)