Data Min Knowl Disc (2017) 31:1480–1505
DOI 10.1007/s10618-017-0527-9

Measuring and moderating opinion polarization in social networks

Antonis Matakos[1] · **Evimaria Terzi**[2] · **Panayiotis Tsaparas**[1]

Received: 6 March 2017 / Accepted: 28 June 2017 / Published online: 15 July 2017

Abstract The polarization of society over controversial social issues has been the subject of study in social sciences for decades (Isenberg in J Personal Soc Psychol 50(6):1141–1151, 1986, Sunstein in J Polit Philos 10(2):175–195, 2002). The widespread usage of online social networks and social media, and the tendency of people to connect and interact with like-minded individuals has only intensified the phenomenon of polarization (Bakshy et al. in Science 348(6239):1130–1132, 2015). In this paper, we consider the problem of measuring and reducing polarization of opinions in a social network. Using a standard opinion formation model (Friedkin and Johnsen in J Math Soc 15(3–4):193–206, 1990), we define the *polarization index*, which, given a network and the opinions of the individuals in the network, it quantifies the polarization observed in the network. Our measure captures the tendency of opinions to concentrate in network communities, creating echo-chambers. Given this numeric measure of polarization, we then consider the problem of reducing polarization in the network by convincing individuals (e.g., through education, exposure to diverse viewpoints, or incentives) to adopt a more neutral stand towards controversial issues. We formally define the ModerateInternal and ModerateExpressed

Responsible editor: Kurt Driessens, Dragi Kocev, Marko Robnik-Šikonja Myra Spiliopoulou.

✉ Antonis Matakos
amatakos@cs.uoi.gr

Evimaria Terzi
evimaria@cs.bu.edu

Panayiotis Tsaparas
tsap@cs.uoi.gr

[1] Department of Computer Science and Engineering, University of Ioannina, Ioannina, Greece

[2] Department of Computer Science, Boston University, Boston, MA, USA

problems, and we prove that both our problems are NP-hard. By exploiting the linear-algebraic characteristics of the opinion formation model we design polynomial-time algorithms for both problems. Our experiments with real-world datasets demonstrate the validity of our metric, and the efficiency and the effectiveness of our algorithms in practice.

Keywords Polarization · Social networks · Opinion formation · Moderation

1 Introduction

In the past decades, online social networks and social media have emerged as the primary vehicle for the public discourse. Today, discussions take place primarily on Facebook and Twitter, where information and viewpoints are exchanged, and opinions are shaped. In this new world, users have easy access to information, but also to a public podium and a broad audience for their opinions.

Empowering ordinary users to express and share their opinions online seems like a step towards making individuals more open to different ideas, cultures, and viewpoints, and thus making societies overall more democratic and diverse. Nevertheless, it has been observed that the easy and uninhibited access to information and expression often leads to the opposite effect. Users tend to create connections with like-minded individuals, and create echo-chambers and filter bubbles that reinforce their existing opinions (Bakshy et al. 2015; Bessi et al. 2016). In such cases, instead of smoothing the differences, online social networks reinforce them, thus leading to increased *polarization*.

Online polarization has been observed over a variety of issues and topics, ranging from frivolous (the dress controversy[1]) to decisive and consequential (the increasing divide in US politics[2]). Polarization separates individuals into sides that have little or no communication with and understanding of each other, and has a corrosive and detrimental effect to the functioning of communities, societies, and democracies. It is thus of critical importance to devise mechanisms for reducing polarization. This is typically achieved by raising awareness and educating individuals about the different sides of an issue, with the goal of moderating extreme opinions and reaching a common ground. This is an arduous and costly process that may span a generation to yield results.

In this paper, we take an algorithmic approach to the problem of measuring and reducing polarization. In order to measure polarization, we consider a popular opinion formation model (Friedkin and Johnsen 1990). In this model, opinions are modeled as real numbers ranging from -1 to 1, depending on the viewpoint of the user. Each user u has an internal opinion s_u that is given as input and it is fixed, and an expressed opinion z_u that depends on their own internal opinion and the expressed opinions in their social network. Using a random walk interpretation of the opinion formation model, we can interpret z_u as the expected opinion that node u will reach when taking a random walk in the social network. High value of z_u implies that the user is surrounded mostly by single-minded individuals with extreme opinions, while low value implies that

[1] https://en.wikipedia.org/wiki/The_dress.

[2] http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/.

the social network of u adopts moderate and diverse opinions. We view the absolute value $|z_u|$ as a measure of the degree of the polarization of user u. Given the vector of expressed opinions $\mathbf{z}$ for the whole network, the length of the opinion vector $\|\mathbf{z}\|^2$ captures the degree of polarization in the network. We refer to $\|\mathbf{z}\|^2$ as the *polarization index* $\pi(\mathbf{z})$ of the network.

Given this numeric measure of polarization, we are interested in algorithms for reducing polarization in the network. We assume that we can reduce polarization by convincing people (through education or other means) to adopt a more moderate opinion. Given a budget value k, we want to find the best set of k individuals in the network, such that convincing them to moderate their opinions (in our model, set their opinion value to zero) will minimize the polarization index of the network. We consider two variants of this problem: the ModerateInternal problem, and the ModerateExpressed problem. In the ModerateInternal problem we moderate the internal opinion of the users, that is, for each user u in the selected set we set $s_u = 0$. This is the case where through education we expose users to the viewpoint of the other side, and lead them to adopt a moderate viewpoint. In the ModerateExpressed problem we moderate the expressed opinion of the users, that is, for each user u in the selected set we set $z_u = 0$. This is a case where we give incentives to users to adopt a moderate public opinion, and propagate a balanced viewpoint.

From the computational point of view, we prove that both problems are NP-hard. We propose algorithms that exploit the properties of the opinion formation model so as to efficiently construct the solution set, as well as efficient heuristics. We experiment on real datasets and we demonstrate the effectiveness of our algorithms in decreasing polarization.

In summary, in this paper we make the following contributions:

- We define a novel polarization index for quantifying polarization in a network, based on the opinions of users under a popular opinion formation model (Friedkin and Johnsen 1990). Our measure takes into account both the existing opinions of the users, and the network structure. To the best of our knowledge we are the first to use this model to measure polarization.
- We define two novel problems, ModerateInternal and ModerateExpressed for reducing polarization in a network. We show that both problems are NP-hard, and propose efficient algorithms for solving them. Our algorithms exploit a linear-algebraic view of the opinion-formation model we adopt.
- We experiment on real data, including a Twitter network from 2016 US Elections. We demonstrate that our polarization index is successful in capturing polarization, and that our algorithms are effective in reducing polarization.

The remaining of the paper is structured as follows. In Sect. 2 we review related work on measuring and reducing polarization, opinion formation models, and opinion mining. In Sect. 3 we define the polarization index, and provide intuition for our definition. In Sect. 4 we define and study the ModerateInternal problem, and in Sect. 5 the ModerateExpressed problem. Section 6 presents the experimental evaluation of our metric and our algorithms, and Sect. 7 concludes the paper. “Appendix A” contains the full proof for the NP-hardness of the ModerateInternal problem.

2 Related work

Although, to the best of our knowledge, we are the first to introduce and study the ModerateExpressed and the ModerateInternal problems, our work is related to recent work on polarization and the study and application of opinion formation models.

Filter bubbles and echo chambers. While social media have the potential to expose individuals to more diverse viewpoints, they can also limit exposure to attitude-challenging information, which leads to a radicalization of attitudes and false perceptions about events. This has led to theories about the effects of "echo-chambers" (Bakshy et al. 2015; Bessi et al. 2016; Garrett 2009), where users are only exposed to information by like-minded individuals, and "filter bubbles" (Bakshy et al. 2015; Pariser 2011), where algorithms only present personalized content that agrees with the user's attitude. Recent lines of work (Garrett 2009) have investigated the strength of echo chambers and filter bubbles, and found that opinion-challenging information reduces the likelihood of a news story's exposure.

Quantifying and reducing polarization. The phenomenon of polarization has been the subject of study in social sciences for decades (Isenberg 1986; Sunstein 2002). There has been a lot of work on measures for quantifying the polarization observed in online social networks and social media (Akoglu 2014; Conover et al. 2011; Garimella et al. 2016; Guerra et al. 2013; Amelkin et al. 2015; Dandekar et al. 2013) and model its emergence (Dandekar et al. 2013; Vicario et al. 2016). The main characteristic of those works is that the measures proposed are based on the structural characteristics of the underlying graph and they do not consider the existing opinions, or an opinion formation model, when quantifying polarization. Vicario et al. (2017) study polarization while incorporating opinion dynamics, assuming a variation of the Bounded Confidence Model (BCM). This variation of the model, has the limitation that it can only converge in states where opinions form clusters of a single value. The closest to our definition is the notion of tension for measuring polarization (Bindel et al. 2015; Dandekar et al. 2013), which focuses on pairwise disagreements of opinions over the edges of the network. The metric does not consider the overall distribution of opinions, and it does not work as well in the presence of echo-chambers in the network, where like-minded individuals only interact with each other.

Given the negative effects of polarization and fragmentation on the well-being and the well-functioning of societies there has been work that focuses on methods for decreasing polarization. Such studies focus on proposing mechanisms that will expose online social-media users to content that is not necessarily aligned with their prior beliefs. The work in this direction can be split into work that focuses on (*a*) *how* to present information to users and (*b*) *who* to approach with the new information. In terms of (*a*) there has been work focusing on user studies as well as the design of the appropriate interfaces that predispose users positively towards diverse ideas presented to them (Munson and Resnick 2010; Munson et al. 2013; Vydiswaran et al. 2015). Clearly, our work is complementary to the above as we focus on the algorithmic aspects of decreasing polarity.

In terms of who to approach, the recent work by Garimella et al. (2017) considers the introduction of edges that will reduce the observed polarization in a social network. Although this work is related to ours, it focuses on graph-theoretic measures of polarization and does not take into consideration the opinions of individuals. Furthermore, it considers the addition of links, rather than the moderation of opinions. Therefore, both our model and our problem are different from that in Garimella et al. (2017).

Opinion mining. In our work we assume that we are given user opinions as input, and we focus on using these opinions to measure and moderate polarization. In our experiments we assume that opinions can be inferred by the actions of the users (membership in known communities, or following specific accounts). In networks where we have information about the attributes of the users, or the content they contribute, it may be possible to obtain more fine-grained and nuanced opinion values by applying opinion mining and sentiment analysis techniques (Liu 2012). Opinion mining deals with the inference of the semantics of a given text. Concept-based techniques have come to prominence recently (Cambria et al. 2015, 2016), along with new neural network approaches, using deep neural networks (Poria et al. 2016; Chen et al. 2016). Our framework for measuring and moderating polarization can be extended to include an opinion mining algorithm as the first step of the pipeline.

Influence and opinion maximization. At a high level, our work is also related to the line of work on influence and opinion maximization (Kempe et al. 2003; Gionis et al. 2013). In these works the goal is to select a set of individuals that will adopt a product or an opinion so as to maximize the overall adoption in the network. The closest to our work is the work of Gionis et al. (2013), where the goal is to find a set of individuals who will change their opinion (internal or expressed), such that the sum of expressed opinions is maximized. Both works assume the same opinion-formation model. However, our goal is different: rather than maximizing the positive expressed opinion we aim at minimizing the polarization index. The difference in the objectives results in differences in the problem properties and the algorithmic techniques that need to be developed.

3 The polarization index

In this section, we define the polarization index we will use in the paper, and we provide the necessary background for understanding and analyzing our metric.

Throughout the paper, we consider a social graph $G = (V, E)$ with n nodes and m edges. Each edge (i, j) is associated with a weight $w_{ij} \geq 0$, which expresses the strength of the connection between i and j, and the influence they exert to each other.

We adopt the opinion-formation model of Friedkin and Johnsen (1990), which assumes that every person i in the network has a persistent *internal opinion* s_i, and an *expressed opinion* z_i which depends both on their internal opinion s_i and the expressed opinions of their neighbours. More precisely, the expressed opinion of node i is computed as the weighted average of their internal opinion and the expressed opinions of the neighbours of i, $N(i)$, in G:

$$z_i = \frac{w_{ii} s_i + \sum_{j \in N(i)} w_{ij} z_j}{w_{ii} + \sum_{j \in N(i)} w_{ij}}, \qquad (1)$$

where w_{ii} denotes the importance that node i places on their own opinion. It has been shown that if every person i iteratively updates their expressed opinion, then the expressed opinions converge to a unique opinion vector $\mathbf{z}$.

In our setting, opinions can be both positive and negative. Thus, we assume that they take values in the interval $[-1, 1]$, where -1 reflects a negative opinion, and 1 a positive one, while 0 corresponds to a neutral position.

In the absence of any polarization all users would express a neutral opinion, i.e., $z_i = 0$ for all $i \in V$. The absolute value of the opinion of a user $|z_i|$ captures how extreme the user opinion is. We quantify polarization in the network by measuring how far we are from the state of complete neutrality. We measure this by looking at the length of the vector $\mathbf{z}$ under the L_2^2 norm. To make the value of our metric independent of the size of the network, we normalize by the number of nodes in the graph. More formally, we have the following definition.

Definition 1 *(Polarization index)* Given a network $G = (V, E)$, a vector of internal opinions $\mathbf{s}$ defined over the nodes of the network, and the resulting vector of expressed opinions $\mathbf{z}$, we define the **polarization index** $\pi(\mathbf{z})$ as: $\pi(\mathbf{z}) = \frac{\|\mathbf{z}\|^2}{n}$.

We now give some additional background and intuition behind our metric. First, an equivalent way of obtaining the *expressed opinion* vector $\mathbf{z}$ from the *internal opinion* vector $\mathbf{s}$ is the following (Bindel et al. 2015): if $\mathbf{L}$ is the Laplacian matrix of graph $G = (V, E)$, and $\mathbf{I}$ is the identity matrix, then $\mathbf{z} = (\mathbf{L} + \mathbf{I})^{-1}\, \mathbf{s}$. We will refer to the matrix $\mathbf{Q} = (\mathbf{L} + \mathbf{I})^{-1}$ as the *fundamental matrix*.

Second, there is a direct connection between the opinion formation model, and random walks with absorbing nodes, as it is shown in Gionis et al. (2013). More specifically, given the graph $G = (V, E)$, with n vertices, and weights w_{ij} for the edges $(i, j) \in E$, we construct the *augmented graph* $H = (V \cup X, E \cup R)$ as follows. For each vertex $v_i \in V$, we add a new vertex x_i in X. We also add a *directed* edge (v_i, x_i) in R, with weight w_{ii}. The node x_i corresponds to the internal opinion of node v_i.

Now consider a random walk on graph H that starts from a vertex $v \in V$. The nodes in X are *absorbing*. That is, when reaching these nodes, the random walk terminates. For each absorbing node x_i and non-absorbing node v_j, we can compute the probability $P(x_i \mid v_j)$, that the random walk that started from v_j terminates at node x_i. It was shown in Gionis et al. (2013) that $\mathbf{Q}(j, i) = P(x_i \mid v_j)$, that is, the j-th row of the matrix $\mathbf{Q}$ is a probability distribution over all nodes in X. Therefore, we have that

$$z_j = \sum_{i=1}^{n} P(x_i \mid v_j) s_i.$$

We can think of the probability $P(x_i \mid v_j)$ as the probability that node v_j adopts the opinion of node v_i. This probability depends on the structure of the graph: the more the paths that connect v_j with node v_i, the higher the probability $P(x_i \mid v_j)$.

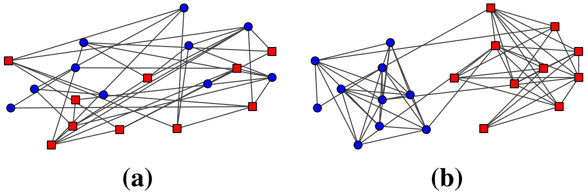


Fig. 1 Three examples of graphs for the polarization index. **a** G_1: Random graph. **b** G_2: Echo chamber graph. **c** G_3: Community structure with random opinion assignment

The probability is also affected by the weights w_{ij} and w_{ii} since they determine the probability that a specific edge is followed. For example, high w_{ii} weight means that the user is more likely to be absorbed in her own opinion node than follow a path in the graph to some other node. The expressed opinion z_j of node v_j is the expected value of the internal opinion of the node at the point of absorption.

The implications of this connection are the following. For a specific node v_j, the value $|z_j|$ is minimized if node v_j has equal probability to reach positive and negative opinions, that is, it has a balanced view of the opinions in the network. On the other hand, if the user is trapped in a filter-bubble of like-minded friends, all with extreme opinions, the value of $|z_j|$ will be high. The polarization index becomes high if we have echo chambers in the network, that is, we have communities in the graph, that are homogeneous with respect to their internal opinions.

To illustrate this point, consider the three graphs shown in Fig. 1. The graph G_1 in Fig. 1a consists of 20 nodes, 10 with opinion -1, and 10 with opinion $+1$, that are randomly interconnected. The graphs in Fig. 1b, c are the same and they consist of two densely connected subgraphs of size 10 that are sparsely interconnected. In Fig. 1b, the opinions are aligned with the communities in the graph: the nodes in the left community have opinion -1 (blue round nodes), while the nodes in the right community have opinion $+1$ (red square nodes). In Fig. 1c, the opinions are randomly assigned.

We compute the polarization index for all three graphs. In the first graph G_1, edges are created at random, and thus the graph has no community structure, and opinions mix randomly in the network. Therefore, each node has more or less equal probability to adopt a positive or negative opinion, resulting in a low polarization index of 0.03. In the second graph G_2, there is a clear echo chamber effect: positive nodes speak mostly with positive nodes, and negative nodes speak mostly with negative nodes. As a result the polarization index is high, 0.30. In the third graph G_3, although there is a clear community structure in the graph, the opinions are equally distributed in the two communities. Therefore, although the nodes tend to communicate mostly with the nodes within their community (the probability of adopting the opinion of a node in a different community is small), both opinions are equally represented in each community, resulting in a small polarization index of 0.03.

Given this measure of polarization in the network, our next goal is to minimize it by convincing a small set of users users to adopt more neutral positions. We consider two possible ways to achieve this. The first is to convince users, via education and exposure to different viewpoints, to change their internal opinions. The second is by giving incentives to users to express and propagate a neutral opinion. In both cases we say that we *moderate* the opinions of the users. Depending on whether we moderate the internal or the expressed opinions of users we define the ModerateInternal and the ModerateExpressed problems respectively. We define and study these two problems in the following sections.

4 Moderating internal opinions

In this section we define the ModerateInternal problem, we analyze its complexity, and we design efficient and effective algorithms for solving it.

4.1 Problem definition

When moderating internal opinions, we seek a small set of nodes, T_s, whose *internal* opinions would be set to zero, such that the polarization index is minimized. We use $\pi(\mathbf{z} \mid T_s)$ to denote the polarization index after setting the internal opinions of the nodes in T_s to zero. The formal problem definition is the following.

Problem 1 *(ModerateInternal)* Given a graph $G = (V, E)$, a vector of internal opinions $\mathbf{s}$, and an integer k, identify a set T_s of k nodes such that changing the internal opinions of the nodes in T_s to 0, minimizes the polarization index $\pi(\mathbf{z} \mid T_s)$.

4.2 Problem complexity

We prove the following Theorem for the hardness of the the ModerateInternal problem.

Theorem 1 *The* ModerateInternal *problem is NP-hard.*

Proof We only give the intuition of the proof here. The full proof appears in the "Appendix A".

Our proof uses a reduction from the m-SubsetSum problem, where given a set of N positive integer numbers $v_1, \ldots, v_N$, a value m, and a target value b, we ask if there is a set of numbers B of size m, such that $\sum_{v_i \in B} v_i = b$.

Given an instance of the m-SubsetSum problem, we construct an instance of ModerateInternal as follows. The graph is a star with $N + 1$ nodes: we have a central node u_0, and a spoke node u_i for each integer v_i. For the center of the star (node u_0) we have that $w_{00} = t$, for an appropriately selected value of t (we will discuss this below), and $s_0 = -1$. The weight of the edge (u_0, u_i) from the center to node u_i is $w_{0i} = v_i$, and the weight of node u_i to its internal opinion is also $w_{ii} = v_i$. The opinion of all spoke nodes is $s_i = 1$. We set $k = N - m$, and we ask for a set of nodes T_s, $|T_s| = k$, such that, when setting $s_i = 0$ for $u_i \in T_s$, $\pi(\mathbf{z} \mid T_z) = \|\mathbf{z}\|^2$ is minimized.

Assume that we have selected the set T_s, $|T_s| = k$. We can prove that

$$\pi(\mathbf{z} \mid T_s) = \frac{N+4}{4} z_0^2 + \frac{N-k}{2} z_0 + \frac{N-k}{4}.$$

Therefore, $\pi(\mathbf{z} \mid T_s)$ is determined by the expressed opinion of the center node z_0. Let $R = V \backslash T_s \cup \{u_0\}$ denote the set of spoke nodes whose opinion was *not* set to 0. Using the equations for the expressed opinions of the opinion formation model we can show the following for the value of z_0 (details in the "Appendix A"):

$$z_0 = \frac{\sum_{u_i \in R} v_i - 2t}{W + 2t}.$$

For the sake of the argument, assume for a moment that we achieve the minimum $\pi(\mathbf{z} \mid T_s)$ for $z_0 = 0$. Then clearly, we need to select a set of nodes in T_s, such that for the nodes in R we have $\sum_{u_i \in R} v_i = 2t$. Setting $t = b/2$ we can prove that we minimize $\pi(\mathbf{z} \mid T_s)$ if and only if there is a set of nodes R such that $\sum_{u_i \in R} v_i = b$, which proves the reduction. However, the value $z_0 = 0$ does not minimize $\pi(\mathbf{z} \mid T_s)$. In the full proof, we determine the optimal value of z_0, and the value of t that achieves this optimal when there is a set of nodes R such that $\sum_{u_i \in R} v_i = b$, and thus complete the reduction. □

Furthermore, we observe that $\pi(\mathbf{z} \mid T_s)$ is not monotone with respect to T_s. That is, it is not necessarily true that the more nodes we make neutral, the lower the polarization. This can be seen by considering a simple graph consisting of two nodes, u and v, with internal opinions -1 and 1 and $w_{uu} = w_{vu} = w_{vv} = 1$. In this case $\pi(\mathbf{z}) = 2/9$. If we change the internal opinion of the negative node to neutral, then $\pi(\mathbf{z}) = 5/9$. Thus, making a node neutral, causes the polarization index to increase. This observation implies that designing an algorithm for ModerateInternal is challenging.

4.3 Algorithms

In this section, we present our algorithms for ModerateInternal. For the following, we assume that the matrix $\mathbf{Q}$ has been pre-computed, and it is given as input to the algorithm.

The BOMP algorithm: The *Binary Orthogonal Matching Pursuit* (BOMP) algorithm is inspired by the connection of the ModerateInternal problem to the problem of sparse approximation (Natarajan 1995). First, we establish this connection and then we describe the BOMP algorithm.

As we have already discussed in Sect. 3, *expressed opinion* vector $\mathbf{z}$ can be computed as $\mathbf{z} = \mathbf{Qs}$, where $\mathbf{Q} = (\mathbf{L} + \mathbf{I})^{-1}$. Note that we have that $\mathbf{Qs} = \mathbf{QS1}$, where $\mathbf{S}$ is the diagonal matrix with $\mathbf{S}_{ii} = s_i$, and $\mathbf{1}$ is the vector of all ones. For the rest of the discussion we will use $\mathbf{R} = \mathbf{QS}$.

Now, let $\mathbf{s}'$ denote the vector $\mathbf{s}$ after we set k of its entries to zero – these entries will correspond to users whose internal opinions become neutral. Our goal is to find the vector $\mathbf{s}'$ that minimizes $\|\mathbf{Qs}'\|^2$. Note that $\mathbf{Qs}' = \mathbf{R1} - \mathbf{Rx}$, where $\mathbf{x}$ is a vector with 1's at the positions of the selected nodes, and zeros everywhere else. Since $\mathbf{R1} = \mathbf{z}$, the original expressed opinion vector, our problem can be stated as follows: Find the best *binary* vector $\mathbf{x}$ with k non-zero entries (i.e., $\|\mathbf{x}\|_0 = k$) such that $\|\mathbf{z} - \mathbf{Rx}\|^2$ is minimized. This is the definition of the sparse approximation problem (Natarajan 1995), where we restrict the solution to binary vectors.

Inspired by Lappas et al. (2012) we will approximate the solution to this problem using a variation of a known algorithm from signal processing (Mallat 2008; Davis et al. 1994) called *nonnegative orthogonal matching pursuit* (`NNOMP`). The `NNOMP` algorithm is designed to find a sparse vector $\mathbf{x}$ (with no more than k non-zero entries) with *non-negative* yet *real* coefficients which when multiplied with a matrix $\mathbf{R}$ is minimizes $\|\mathbf{z} - \mathbf{R}\mathbf{x}\|^2$ for a target vector $\mathbf{z}$.

In our problem, the vector $\mathbf{x}$ is a binary vector and thus $\mathbf{x}$ essentially selects a subset of columns from $\mathbf{R}$ and uses their sum to approximate the target vector $\mathbf{z}$. Our algorithm, *Binary Orthogonal Matching Pursuit* (`BOMP`), is a variant of `NNOMP` and it proceeds in iterations. At iteration t, `BOMP` starts with a vector $\mathbf{x}^{t-1}$ with $(t-1)$ entries of value 1. These entries correspond to the columns of the matrix $\mathbf{R}$ that have been selected up to this iteration. Let $\hat{\mathbf{z}}^{t-1} = \mathbf{R}\mathbf{x}^{t-1}$ denote the approximation of the target vector $\mathbf{z}$ constructed so far. The algorithm selects the column from $\mathbf{R}$ (not selected so far) that has the largest dot-product with the residual $\mathbf{z} - \hat{\mathbf{z}}^{t-1}$ of the target vector. The set of selected indexes is augmented with this new index to produce vector $\mathbf{x}^t$. The algorithm terminates when we have selected k columns. The set of columns define the set T_s of nodes whose internal opinions will be set to zero.

The computational complexity of the `BOMP` algorithm is $O(kn^2)$. In each of the k iterations, the algorithm computes the dot-product of every candidate index to be added to set of selected indices with the residual vector. This step is the most computationally expensive, requiring time $O(n^2)$. All the other steps require at most $O(n)$ time, resulting in $O(kn^2)$ complexity in total.

The `GreedyInt` algorithm: We also consider a greedy algorithm for the problem. The algorithm builds the selected set of nodes T_s iteratively. It starts with an empty set T_s^0, and at each step t it adds to the existing solution T_s^{t-1} the node v which, when added to the solution, $T_s^t = T_s^{t-1} \cup \{v\}$, it causes the largest decrease $\pi(\mathbf{z} \mid T_s^{t-1}) - \pi(\mathbf{z} \mid T_s^t)$ in the objective function. We will denote this algorithm as `GreedyInt`.

The `GreedyInt` algorithm can be implemented efficiently by exploiting the observation that the effect of neutralizing a node in the graph in the expressed opinion $\mathbf{z}$ can be computed by the subtraction of the corresponding column of the matrix $\mathbf{R}$ from $\mathbf{z}$. Therefore, for each candidate node v we need time $O(n)$ to compute $\pi(\mathbf{z} \mid T_s^{t-1} \cup \{v\})$, resulting in complexity $O(n^2)$ for each iteration, and $O(kn^2)$ for the algorithm in total.

5 Moderating expressed opinions

In this section, we define the ModerateExpressed problem, we analyze its complexity, and we design an efficient algorithm for solving it.

5.1 Problem definition

When moderating expressed opinions, we seek a small set of nodes T_z to set their *expressed* opinions to zero, such that the polarization index is minimized. We use $\pi(\mathbf{z} \mid T_z)$ to denote the polarization index after setting the expressed opinions of the nodes in T_z to zero. The formal problem definition is the following.

Problem 2 *(ModerateExpressed)* Given a graph $G = (V, E)$, a vector of internal opinions $\mathbf{s}$, the resulting vector of expressed opinions $\mathbf{z}$, and an integer k, identify a set T_z of k nodes such that fixing the expressed opinions of the nodes in T to 0, minimizes the polarization index $\pi(\mathbf{z} \mid T_z)$.

5.2 Problem complexity

We prove the following Theorem for the hardness of the ModerateExpressed problem.

Theorem 2 *The* ModerateExpressed *problem is NP-hard.*

Proof The proof of the theorem follows closely the proof of hardness in Gionis et al. (2013), so we only provide the correspondence between the two proofs. The proof exploits the equivalence between the opinion formation model and absorbing random walks, shown in Gionis et al. (2013).

Similar to the proof in Gionis et al. (2013) our proof uses a reduction from the Vertex Cover on Regular Graphs problem (VCRG) (Feige 2003). We show that there exists a set of nodes Y for a regular graph G_{VC} in the VCRG problem such that $|Y| \leq K$ and Y is a vertex cover if and only if there exists a set T_s for a graph G in the ModerateExpressed problem, such that $|T_z| \leq k$ and $\pi(\mathbf{z} \mid T_z) \leq \theta$, for $\theta = \frac{n}{2(d+1)^2}$. In our construction we set $G = G_{VC}$, and we initialize the vector $\mathbf{s}$, such that $s_i = 1$ for all $i \in V$. The proof then proceeds in the same way as in Gionis et al. (2013). We can show that we can achieve a value $\pi(\mathbf{z} \mid T_z)$ less than θ if and only if the nodes T_z that we select in G (to make absorbing) define a vertex cover in G_{VC}. □

Using a similar example as the one we used in Sect. 4.2 we can show that $\pi(\mathbf{z} \mid T_z)$ is also not monotone with respect to T_z, implying again that it is not straightforward to design an algorithm for solving ModerateExpressed.

5.3 Algorithms

Our algorithm for ModerateExpressed is a greedy algorithm, which we call `GreedyExt`. `GreedyExt` is an iterative algorithm which starts with an empty set T_z^0. At each step t the algorithm adds to the existing solution T_z^{t-1} the node v_i, which, when setting $z_i = 0$, it causes the largest decrease $\pi(\mathbf{z} \mid T_z^{t-1}) - \pi(\mathbf{z} \mid T_z^t)$ in the objective function.

A naive implementation of the `GreedyExt` algorithm is computationally expensive. At each step of the algorithm we need to check n nodes, and for each node compute the new opinion vector after setting the expressed opinion of the node to zero. The most straightforward way to do this is by multiplying $\mathbf{s}$ directly with $\mathbf{Q}$, in $O(n^2)$ time; recall that $\mathbf{Q} = (\mathbf{L} + \mathbf{I})^{-1}$, and thus it is a dense matrix. Alternatively, one can iteratively apply Eq. (1) and achieve the same computation in time $O(mI)$, where I the number of iterations it will take until convergence and m the number of edges of G. In our experiments, this computation converges in about a hundred iterations. Thus if we implement `GreedyExt` using the first method its running time becomes $O(kn^3)$, while with the second method the running time is $O(knmI)$. Our experiments with large graphs show that both these computations are impractical when dealing with medium-size datasets.

In order to improve the overall running time of `GreedyExt`, we exploit the *Sherman–Morrison* formula (Hager 1989), a special case of the *Woodbury matrix identity*, to speed up the computation of the updated polarization index after adding a new node to the solution set. The identity states that the inverse of a matrix after adding a *rank-1* correction matrix to it can be computed by doing a *rank-1* correction to the inverse of the original matrix. Formally, given an invertible matrix **A** and vectors **u** and $\mathbf{v}^T$, the Sherman–Morrison formula states that:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}} \tag{2}$$

Consider now the case where we want to add node v_i to the solution set T_z. Let $\pi(\mathbf{z}')$ denote the new polarization index after setting $z_i = 0$. We can express the polarization index as $\pi(\mathbf{z}') = \|\mathbf{Q}'\mathbf{s}'\|^2$. In this equation, $\mathbf{Q}' = \left(\mathbf{L}' + \mathbf{I}\right)^{-1}$, where $\mathbf{L}'$ is the updated Laplacian with a row of zeros at the i-th index, and $\mathbf{s}'$ is the updated internal opinion vector, with $s_i = 0$ at the i-th entry. To understand the update process of **Q** and **s**, note that in the random walk interpretation, setting $z_i = 0$ is equivalent to removing all outgoing edges from node v_i and keeping only the edge to the node x_i, while setting $s_i = 0$.

We can express the Laplacian matrix $\mathbf{L}'$ as a *rank-1* correction of the Laplacian **L**, that is, $\mathbf{L}' = \mathbf{L} + \mathbf{u}\mathbf{v}^T$, where **u** is the unit vector with 1 at the i-th entry, and $\mathbf{v}^T$ is the negative i-th row of the Laplacian. Following the Sherman–Morrison formula (Eq. 2) we have that $\mathbf{Q}' = \mathbf{Q} - \mathbf{B}$, where

$$\mathbf{B} = \frac{\mathbf{Q}\mathbf{u}\mathbf{v}^T\mathbf{Q}}{1 + \mathbf{v}^T\mathbf{Q}\mathbf{u}},$$

We can also write $\mathbf{s}' = \mathbf{s} - \bar{\mathbf{s}}$, where $\bar{\mathbf{s}}$ is a vector with $\bar{s}_i = s_i$, and zero in all other entries. Thus, we have:

$$\begin{aligned}\left\|\mathbf{Q}'\mathbf{s}'\right\|^2 &= \|(\mathbf{Q} - \mathbf{B})(\mathbf{s} - \bar{\mathbf{s}})\|^2 \\ &= \|\mathbf{Q}\mathbf{s} - \mathbf{B}\mathbf{s} - \mathbf{Q}\bar{\mathbf{s}} + \mathbf{B}\bar{\mathbf{s}}\|^2 \\ &= \left\|\mathbf{z} - \frac{\mathbf{Q}\mathbf{u}\mathbf{v}^T\mathbf{z}}{1 + \mathbf{v}^T\mathbf{Q}\mathbf{u}} - \mathbf{Q}\bar{\mathbf{s}} + \frac{\mathbf{Q}\mathbf{u}\mathbf{v}^T\mathbf{Q}\bar{\mathbf{s}}}{1 + \mathbf{v}^T\mathbf{Q}\mathbf{u}}\right\|^2.\end{aligned} \tag{3}$$

In order to efficiently compute the quantity in Eq. (3), we perform the operations in such an order, so that we never need to compute any $n \times n$ matrix. As a result we can compute Eq. (3) in time $O(n)$, which is better than the $O(mI)$ complexity of the power-iteration, given that $m = nd$, where d is the average degree of the graph.

First, we compute the vector $\mathbf{w} = \frac{\mathbf{Q}\mathbf{u}}{1+\mathbf{v}^T\mathbf{Q}\mathbf{u}}$. This can be computed in linear time. Since **u** is the unit vector with 1 in one entry, and zero everywhere else, **Qu** can be obtained in $O(1)$ via column selection. Given the vector **Qu** we can compute $\mathbf{v}^T\mathbf{Q}\mathbf{u}$

in $O(n)$ time, and then obtain $\mathbf{w}$ by scaling $\mathbf{Qu}$, bringing the total computational cost of $\mathbf{w}$ to $O(n)$.

Given the vector $\mathbf{w}$ we can now compute $\mathbf{w}\mathbf{v}^T\mathbf{z}$ (the second term in Eq. (3)) in linear time, by first computing the dot-product $\mathbf{v}^T\mathbf{z}$, and then scaling the vector $\mathbf{w}$ with the result. Also, we can compute the vector $\mathbf{Q}\bar{\mathbf{s}}$ in $O(n)$ time, by first selecting the column of $\mathbf{Q}$ and then scaling it by s_i. The term $\mathbf{w}\mathbf{v}^T\mathbf{Q}\bar{\mathbf{s}}$ (the last term in Eq. (3)) can be computed as before in linear time. All other computations are computations on vectors, resulting in $O(n)$ total cost for the computation of Eq. (3).

We repeat the above procedure n times to find the best candidate node. For the selected node, we compute the updated matrix $\mathbf{Q}' = \mathbf{Q} - \mathbf{B}$ using the Sherman–Morrison formula in $O(n^2)$. This brings the total computational cost of `GreedyExt` to $O(kn^2)$.

6 Experiments

In this section, we present an experimental evaluation of the polarization index, and of our algorithms for both problems. The goals of our experiments are to validate the polarization index, study the properties of the proposed algorithms, and evaluate their performance and scalability.

6.1 Datasets

We consider five datasets representing different types of social networks. We use networks that are partitioned into opposing communities, and there is ground-truth data about the community membership of the nodes. Thus, we can naturally assign internal opinions -1 and 1 to the nodes depending on their community membership. We consider the following datasets:

Karate[3]: This dataset represents a social network of friendships between 34 members of a karate club at a US university in the 1970s. The social network is partitioned into two distinct equal-sized communities that correspond to two fractions built around two rival instructors.

Books[4]: This is a network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing. Books are classified as *Liberal*, *Conservative*, and *Neutral*. There are in total 43 liberal books, 49 conservative, and 13 neutral . We handled the neutral nodes by assigning to them internal opinion zero.

Blogs[5]: A directed network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance (2005). Blogs are classified as either *Liberal* or *Conservative*. We converted the social graph into an undirected one and only kept the largest

[3] https://networkdata.ics.uci.edu/data.php?id=105.

[4] https://networkdata.ics.uci.edu/data.php?id=8.

[5] https://networkdata.ics.uci.edu/data.php?id=102.

connected component. The resulting dataset contains two communities with 636 and 586 nodes each, and 19,089 edges.

Elections: This dataset is the network between the Twitter followers of Hillary Clinton and Donald Trump collected in the period 15/12/2016-15/01/2017–around the time of the 2016 presidential elections. Members of this network are assigned an internal opinion of 1 or -1 based on which one of the two candidates they follow. Followers of both candidates are assigned a neutral opinion. Since the dataset is prohibitively large (20M followers), we only considered the network formed by the first 50,000 users, according to their user id. We took the largest connected component and iteratively pruned nodes to guarantee that every node has degree greater than 1. The resulting network had a disproportionately large number of Clinton followers so we subsampled her followers to ensure that the ratio of followers for each side reflected the one in the full dataset. In the resulting network there are 7715 Hillary Clinton followers, 8336 Donald Trump followers, and 2216 Neutral followers, for a total of 18,267 users with 204,040 connections between them. As before, we treat the network as undirected.

Hashtags: Using the followers of Clinton and Trump that we collected, we also created "topical" networks, where we assign the opinions according to the specific hashtag that the users tweeted. We considered two pairs of hashtags: The `#maga` and `#imwithher` hashtags, which we expect to be polarized, and the `#halloween` and `#walkingdead` hashtags for which we do not expect to have polarization. We selected these hashtags since they are among the most popular in the dataset. We sampled users that have tweeted at least one hashtag from both pairs, and we created the follow network between them. Again, we kept the largest connected component and iteratively pruned nodes to guarantee that every node has degree greater than 1. The resulting network has 18,890 nodes and 269,696 edges. Using this graph, we consider two possible settings for the opinions: In the first, we assign opinion -1 if the users have tweeted the hashtag `#maga`, 1 if they have tweeted `#imwithher`, and 0 if they have tweeted both. We will refer to this dataset as *Hashtags P*. In the second, we assign opinion -1 if the users have tweeted the hashtag `#halloween`, 1 if they have tweeted `#walkingdead`, and 0 if they have tweeted both. We will refer to this dataset as *Hashtags NP*. These two different settings allow us to study the behavior of our metric in a polarized (*Hashtags P*), and non-polarized network (*Hashtags NP*).

Table 1 summarizes the statistics of our datasets. For all datasets we treat the graphs as undirected. When applying the opinion formation model, we set all edge weights, and all opinion weights to be 1. In order to handle the cases where there is an imbalance of opinions, we normalize the opinion values by subtracting the mean opinion and dividing by the difference of the maximum and the minimum. In this way, the mean opinion value becomes zero, which we consider to be the moderate stance.

6.2 Evaluation of the polarization index

In this section we evaluate the metric in its ability to identify polarization. We compute the value of the metric for the five different datasets. Table 2 shows the values we obtain. In order to understand if the values are indicative of polarization, we perform

Table 1 Dataset statistics

Dataset	Nodes	Edges	Avg degree	Diameter	Positive	Negative	Neutral
Karate	34	78	4.58	5	17	17	0
Books	105	441	8.4	7	43	49	13
Blogs	1222	16,717	27.36	8	636	586	0
Elections	18,267	204,040	22.33	8	7715	8336	2216
Hashtags P	18,890	269,696	28.55	7	12,281	6612	0
Hashtags NP	18,890	269,696	28.55	7	12,408	4102	2383

Table 2 Dataset polarization index and randomization values

Dataset	π	Mean π for random assignments	Std. dev.
Karate	0.089	0.022	0.00499
Books	0.107	0.007	0.00172
Blogs	0.029	0.012	0.00027
Elections	0.012	0.011	0.00007
Hashtags P	0.028	0.005	0.00004
Hashtags NP	0.0049	0.0044	0.00005

a *randomization* test, where we randomly assign the internal opinions on the graph. A randomized assignment of opinions that is independent of the network structure does not create any opinion clusters, and thus it corresponds to a non-polarized state. We compare the value of the π with that of the random assignment, in order to understand the significance of the polarization index value.

We create 100 random assignments of the opinion values, and we report the average and standard deviation of the polarization index values we obtain for these cases. We observe that the polarization index values are significantly higher than those in the randomized datasets in all networks except for the *Elections* and *Hashtags NP* datasets, where the π values are small, and close to that of the random assignment. In the case of the *Hashtags NP* dataset we obtain essentially the same π value.

It is interesting to contrast the π values for the *Hashtags P* and *Hashtags NP* datasets. In the first case, the polarization is much higher, which agrees with our intuition that these hashtags are adopted by different communities that do not interact with each other. In the second case the polarization index is much lower, and close to that of the random assingment. This suggests that the distribution of the opinions in the second case cuts across the natural communities that appear in the graph. Although users are organized in two well-separated communities, positive and negative opinions appear in both, and as a result there is no polarization and echo-chamber effect. This experiment highlights the importance of taking the opinions into account for measuring polarization; looking only at the network structure it is not possible to differentiate between these two cases.

6.3 Heuristic algorithms for opinion moderation

In addition to the algorithms we described in Sects. 4 and 5, we also consider a few more scalable heuristics. Our reasoning in the design of the heuristics is that in order to moderate the overall expressed opinion we need to convert to neutral the opinions of individuals that express extreme opinions, individuals belonging to extreme neighborhoods, or individuals that are influential in the network. The following algorithms implement this reasoning.

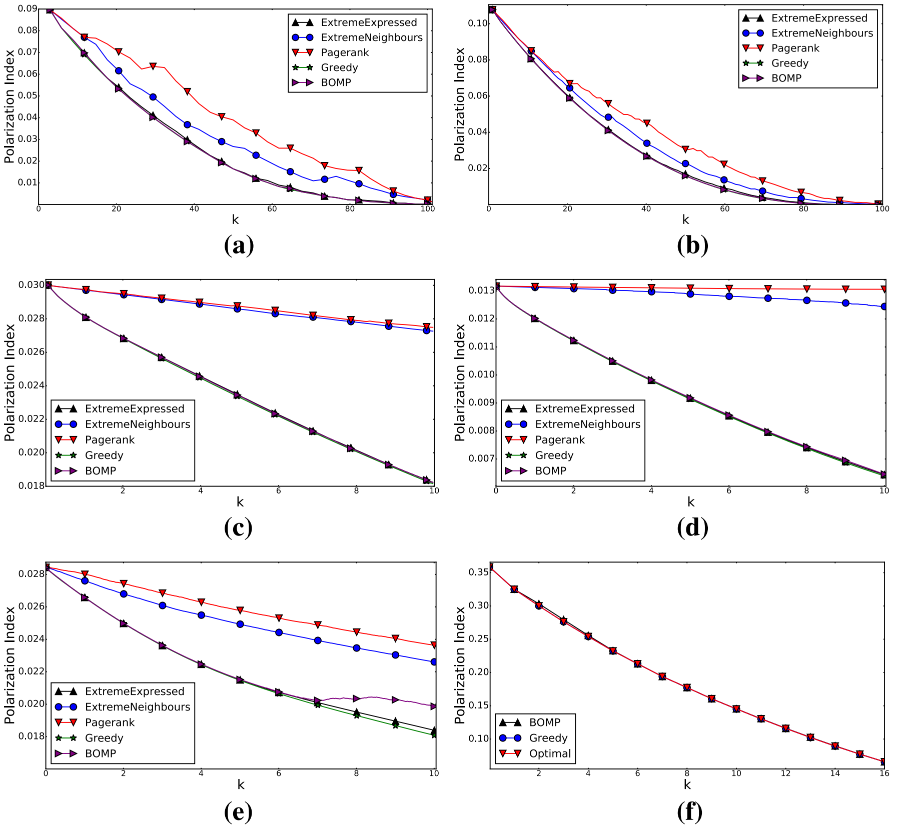
`ExtremeExpressed`: This heuristic works iteratively and at each step it selects to neutralize the node v with the highest expressed opinion $|z_v|$. Since it requires $O(n)$ time to find the most extreme node, the complexity of the algorithm is determined by the time required to compute the updated $\mathbf{z}$ vector after neutralizing a node. In the case of ModerateInternal, as we have shown in Sect. 4.3, we can efficiently calculate the new $\mathbf{z}$ vector by subtracting the column of $\mathbf{Q}$ corresponding to the neutralized node from the current $\mathbf{z}$ vector. Therefore, the algorithm has complexity $O(kn)$. In the case of the ModerateExpressed problem, the fastest way to compute the new $\mathbf{z}$ is by iteratively updating the z_i values as defined in Eq. (1), until convergence. The updates are implemented using efficient matrix-vector multiplication. This takes time $O(mI)$, where I is the number of iterations required for convergence, and m is the number of edges in the graph, leading to complexity $O(kmI)$ for the algorithm. In practice, we have found that the algorithm converges in less than 100 iterations.

`ExtremeNeighbors`: In this heuristic we select the next node to neutralize based on how extreme the neighborhood of the node is. The intuition is that neutralizing this node will have an effect on many extreme nodes. The algorithm at each step changes the opinion of the node v whose neighbors have the highest absolute sum of expressed opinions, that is, $v = \arg\max_{i \in V} |\sum_{j \in N(i)} \mathbf{z}_j|$. For every node we need to check its neighbors, which takes $O(m)$ time, and then update $\mathbf{z}$, accordingly. Therefore, if we use the efficient update of $\mathbf{z}$ as above, the complexity of the algorithm for ModerateInternal is $O(k(n+m))$. Using the iterative method to compute $\mathbf{z}$, the complexity of the algorithm for ModerateExpressed is $O(k(n+m)I)$.

`Pagerank`: The idea behind this heuristic is that in order to moderate the overall opinion, it is a good idea to neutralize the nodes that are central in the network. This will result in maximum spread of a balanced viewpoint. We use PageRank (Lawrence et al. 1998) to measure the centrality of a node. The algorithm selects the nodes in decreasing order of their PageRank value. The complexity of the algorithm is $O((m+n)I + n\log n)$, where $O((n+m)I)$ is the time to compute the PageRank values, and $O(n\log n)$ is the time required to sort the nodes.

6.4 Evaluation of algorithms for ModerateInternal

We first evaluate our algorithms with respect to the value they achieve for the objective function π. We evaluate on all five networks. For the *Hashtags* network we use the hashtags `#maga` and `#imwithher` to set the opinions, since the `#halloween` and `#walkingdead` hashtags are already moderate. Figure 2 shows the value of $\pi(\mathbf{z} \mid T_s)$ for different sizes of the solution set $|T_s| = k$, for all datasets. For the smaller datasets



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Fig. 2 Performance of the algorithms for the ModerateInternal problem on all datasets. **a** *Karate*. **b** *Books*. **c** *Blogs*. **d** *Elections*. **e** *Hashtags P*. **f** Comparison with optimal

Karate and *Books* we let k range over the full size of the dataset. This is impractical for the larger *Blogs*, *Elections* and *Hashtags* datasets, hence we consider T_s up to 10% of the dataset; we plot the value of π in increments of 1%.

As expected, the GreedyInt algorithm achieves the best performance in all datasets. The performance of GreedyInt is consistently matched by BOMP and ExtremeExpressed. The Pagerank and ExtremeNeighbors algorithms are significantly worse, and in the big datasets they achieve only a minimal reduction in π. While we expected the BOMP algorithm to be competitive with GreedyInt the performance of ExtremeExpressed was a surprise. We also compare the BOMP and GreedyInt algorithms against the optimal for k up to 50% of the graph, for the smallest dataset *Karate*, where this computation is possible. We observe that the GreedyInt algorithm behaves optimally, while BOMP achieves performance very close to optimal for $k \leq 6$, and coincides with it for $k > 6$.

Our results indicate that in order to minimize polarization in the ModerateInternal problem, the best strategy is to moderate the nodes with the most extreme opinions. The Pagerank and ExtremeNeighbors algorithms that take into account how well a given node is connected to the network do not perform well.

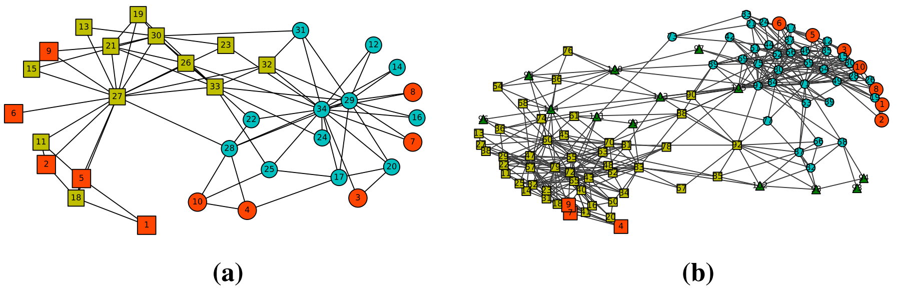


Fig. 3 Selected nodes by `Greedy` on the *Karate* and *Books* datasets for the ModerateInternal problem. **a** *Karate*. **b** *Books*

We further investigate this observation by visualizing the nodes selected by `GreedyInt` in Fig. 3, for the two smaller datasets, *Karate* and *Books*. In the visualization, we assign different color and shape to the nodes of the different communities. The nodes are numbered according to their selection order by `GreedyInt`. The first ten nodes are colored in orange-red and have larger size.

The visualization further confirms the behavior of `GreedyInt`: the nodes that are selected first are nodes on the outskirts of the network. This means that the impact on $\mathbf{z}$ is bigger when moderating fringe nodes with extreme opinions, instead of central nodes. The broader implication of this is that for the ModerateInternal problem the best we can do for moderating polarization is to change the opinions one user at a time, rather than “diffusing” moderation in the network. This is in part due to the fact that the internal opinion is only one of the contributing factors to the expressed opinion of an individual, and thus its change has a limited effect.

6.5 Evaluation of the algorithms for ModerateExpressed

For the evaluation of the ModerateExpressed problem we follow the same methodology as for ModerateInternal. Figure 4 shows the $\pi(\mathbf{z} \mid T_z)$ as a function of the size of T_z for all datasets.

As expected, the `GreedyExt` is again the best-performing algorithm. However, the performance of the other algorithms changes depending on the dataset. For the *Karate*, *Books*, *Blogs* and *Hashtags P* datasets, `ExtremeNeighbors` and `Pagerank` achieve performance close to that of `GreedyExt`, especially for smaller values of k, while `ExtremeExpressed` is clearly the worst performer. As the size of the solution increases, `Pagerank` and `ExtremeNeighbors` seem to lose their effectiveness, while `ExtremeExpressed` catches up with them. These results indicate that when moderating expressed opinions, it is a good strategy to select nodes that are relatively central and express an extreme opinion. After selecting a sufficient number of influential nodes, the gains of moderating central nodes is diminished, and there is more benefit in neutralizing extreme nodes which were not affected by the influential ones, essentially, adopting the approach of moderating one node at the time.

However, we observe a very different picture in the *Elections* dataset, where `ExtremeExpressed` is almost as good as `GreedyExt`, and `Pagerank` and

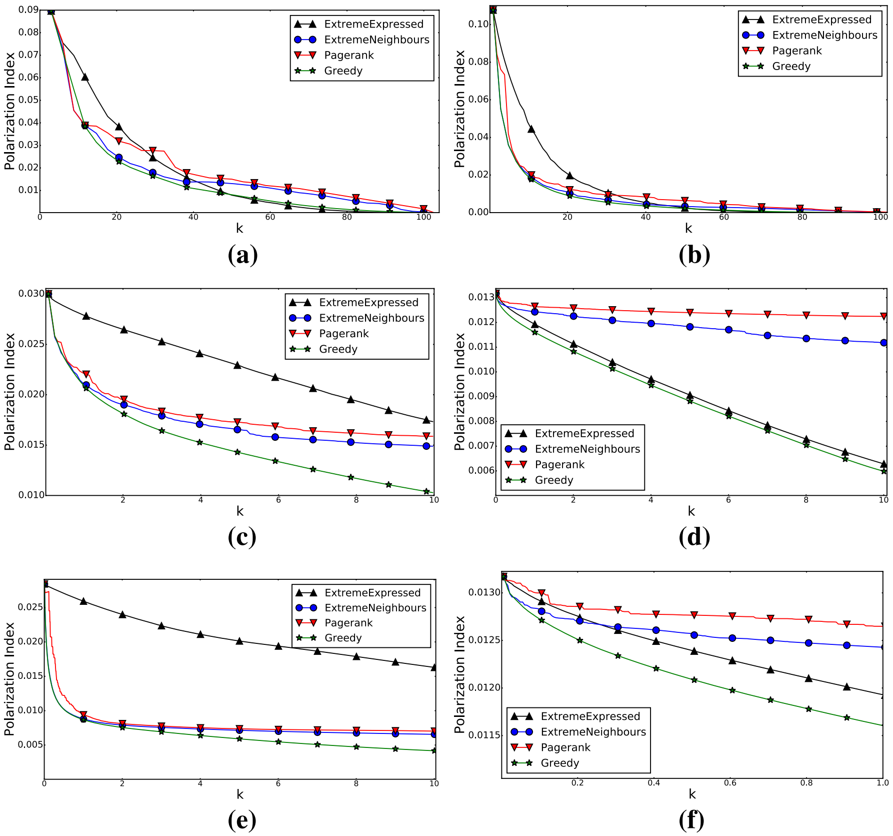
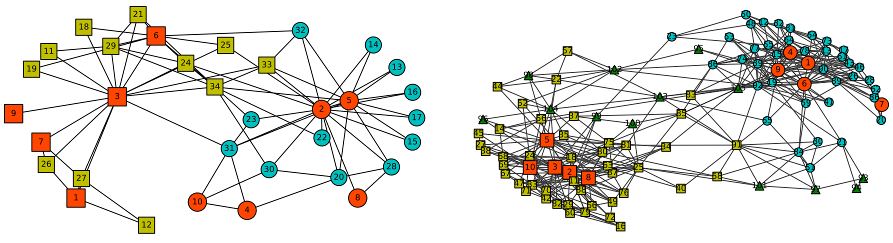


Fig. 4 Performance of the algorithms for the ModerateExpressed problem on all datasets. **a** Karate. **b** Books. **c** Blogs. **d** Elections. **e** *Hashtags P*. **f** Elections top-1%

ExtremeNeighbors perform poorly. Note that, according to the randomization test, the *Elections* dataset is not very polarized. Therefore, there is sufficient mixing of opinions and it is not possible to moderate a large number of nodes by neutralizing an influential node. The one-node-at-the-time approach works better. In order to further investigate this claim we "zoom in" in the performance of the algorithms for k up to the top-1% of the nodes. Now, Pagerank and ExtremeNeighbors appear competitive for small k, but their performance deteriorates in comparison to ExtremeExpressed as k increases. This is in stark contrast to the *Hashtags P* dataset, which is of similar size with *Elections* but it is very polarized, where the algorithms that change influential nodes achieve a good performance.

From Fig. 4 we also observe that the reduction in $\pi(\mathbf{z})$ is significantly higher for the ModerateExpressed problem than for ModerateInternal for the same dataset, for the same k. This is expected, as the moderation of the expressed opinion has a much larger effect in the opinion of the individual, and the opinions in her social network, than the moderation of the internal opinion.



(a) **(b)**

Fig. 5 Selected nodes by `Greedy` on the *Karate* and *Books* datasets for the ModerateExpressed problem. **a** *Karate*. **b** *Books*

Table 3 Running times (secs) of all algorithms for $k = 0.1n$ in the *Elections* dataset

ModerateInternal		ModerateExpressed	
Algorithm	Running time (s)	Algorithm	Running time (s)
`BOMP`	2725		
`GreedyInt`	2930	`GreedyExt`	16,326
`ExtremeExpressed`	6	`ExtremeExpressed`	87
`ExtremeNeighbors`	106	`ExtremeNeighbors`	121
`Pagerank`	7	`Pagerank`	17

In Fig. 5, we visualize again the selected nodes by `GreedyExt` for the *Karate* and *Books* datasets. The selection is different from the one we obtained for the ModerateInternal problem (Fig. 3), and highlights the different nature of the two problems. In the solution of ModerateExpressed the nodes selected are more central in the graph. It is obvious that changing the expressed opinion of a node has a bigger impact on the opinions of the neighbors of that node. As a result, `GreedyExt` tries to pick nodes that are both central and extreme. The first selection of `GreedyExt` is the node that it is ranked first for both `Pagerank` and `ExtremeExpressed`. This combination is essential in achieving high reduction of π. As the selection process continues, the selections of `GreedyExt` alternate between central and fringe nodes as the algorithm is trying to "cover" different parts of the graph, and moderate opinions of nodes that are not easily reached by the central nodes.

6.6 Scalability

We now evaluate the scalability of our algorithms. Table 3 shows the running time for all algorithms on the *Elections* dataset for ModerateInternal and ModerateExpressed and $k = 0.1n$. All experiments were conducted on a machine with an Intel Core i7-4790 CPU and 16GB RAM. The algorithms are implemented in Python using the networkx and numpy libraries.

As expected from the theoretical analysis, for the ModerateInternal problem `ExtremeExpressed`, `ExtremeNeighbors` and `Pagerank` far outperform the

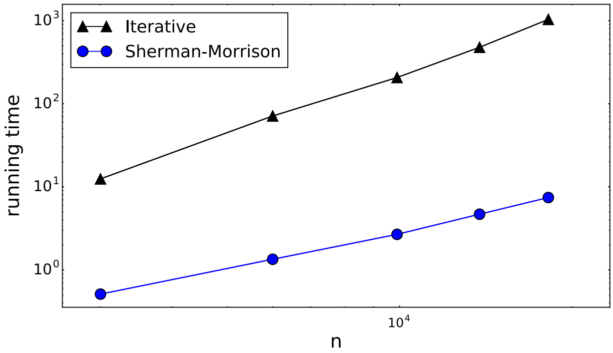


Fig. 6 Comparison of the running times (in secs) for the **Sherman–Morrison and iterative implementation of `GreedyExt`, for varying size of *n*

`GreedyInt` and `BOMP` in terms of running time. Given that `ExtremeExpressed` is matching the performance of `GreedyInt` and `BOMP`, this indicates that this is an effective heuristic for very large datasets.

We also study the effect of using the Sherman–Morrison formula in the computation of the update of the $\mathbf{z}$ vector when `GreedyExt` considers a candidate node. We consider two implementations of `GreedyExt`: one that uses the Sherman–Morisson formula (Sect. 5.3), and one that computes the $\mathbf{z}$ vector using Eq. (1) iteratively. We construct different samples from the *Elections* dataset, of size 2.8, 6, 9.9, 13.8 and 18.2 K. Figure 6 shows the comparison of the two implementations for one update of $\mathbf{z}$. The x-axis is the size of the graph and the y-axis is the running time (in secs). The plot is in log-log scale. Clearly, the Sherman–Morrison implementation is one order magnitude faster, making the algorithm scalable for larger datasets.

Our algorithms use the fundamental matrix $\mathbf{Q}$, and thus require quadratic amount of memory and time. They are applicable to medium-to-large networks, such as an ego-network, or the network induced by the users of specific hashtags, or a subset of Twitter followers, but they cannot be used for massive networks of millions of nodes. In such cases, we can use the iterative computation of the $\mathbf{z}$ vector. This computation is very similar to the computation of PageRank, and lends itself to a distributed implementation. Using existing distributed computation techniques, we can compute the polarization index for very large networks. Our algorithms can combine the efficient heuristics we described to reduce the number of candidate nodes to be considered (e.g., consider only the top nodes in terms of z_i, or PageRank value).

6.7 Case study

We conclude by taking a closer look at the characteristics of individuals that were selected by `GreedyExt` in the *Elections* dataset. For this, we pick the first 10 nodes selected by `GreedyExt` and we rank them according to three other measures: the extremity of their expressed opinion, measured by the node's $|z_i|$, their centrality, measured by their PageRank score and their degree. In the last three columns of

Table 4 Characteristics of the first ten nodes selected by `GreedyExt` in *Elections* dataset

	Opinion	z_i	$\|z_i\|$ rank	Degree rank	Pagerank rank
1	Positive	0.045	10683	16	11
2	Neutral	−0.049	10,211	34	21
3	Positive	0.03	11704	9	8
4	Negative	−0.35	32	12,473	4861
5	Negative	−0.03	12,582	52	37
6	Positive	0.02	13634	2	1
7	Negative	−0.04	10,844	114	59
8	Negative	−0.06	8366	257	157
9	Negative	−0.07	7486	601	170
10	Positive	0.04	11,269	23	19

Table 4 we show the rank of these first 10 nodes in the three rankings. In the same table we report the internal opinion of the node (-1 for Trump and $+1$ for Clinton) and their original expressed opinions z_i.

We observe that `GreedyExt` mainly selects central nodes, but also selects a node with very extreme expressed opinion high in the list (4-th pick). Nine out of the top-10 nodes are clearly very central in the network as they are ranked high both by their degree and their PageRank scores. This is in line with the previous observation that `GreedyExt` initially tries to diffuse as much neutrality as possible and then tries to cover individuals that were not reached. We also note in the top-10 selected nodes we have five Trump followers, four Clinton followers, and a neutral user. Note that this matches relatively closely the proportions of Trump, Clinton and Neutral followers in the full dataset, which agrees with our previous observations on *Karate* and *Books*, and indicates that it is a good strategy to take a balanced approach to moderating opinions.

7 Conclusions

In this paper we considered the problem of polarization in online social networks. Using a popular opinion formation model, we proposed the *polarization index*, a novel measure for quantifying the degree of polarization in the network that takes into account both the network structure and the existing opinions of users. We then considered the problem of identifying a small set of individuals, such that, if we convince them to adopt a moderate opinion, this will minimize the polarization index. We defined two variants of the problem, and showed that both variants are NP-hard. We proposed efficient algorithms by exploiting the mathematical properties of the opinion formation model. Experiments with real data demonstrate the validity of our model, and the effectiveness of our algorithms in reducing polarization. Our experiments also highlight the properties and the differences of the two problems we considered.

In our work we assumed that the opinions are given as input for the computation of the polarization index. An interesting future direction for our work is to use opinion mining techniques to derive the opinions of the users in the social network. Such techniques can be used as the first step in our pipeline. Alternatively, we could integrate ideas from opinion and sentiment mining into the computation of a polarization metric, or in the moderation algorithms.

Furthermore, our approach to moderation is to set the opinions of the users to zero. An alternative approach would be to set the user opinions to values other than zero, so as to minimize polarization. In the case of the internal opinions, there are interesting connections of this problem with the algebraic properties of the fundamental matrix **Q** that are worth exploring in future work.

Acknowledgements This work was supported by the Marie Curie Reintegration Grant projects titled JMUGCS which has received research funding from the European Union, and the National Science Foundation grants: IIS 1320542, IIS 1421759, CAREER 1253393, as well as a gift from Microsoft. We would also like to thank Evaggelia Pitoura for useful comments and discussions on early drafts of the paper.

Appendix A: Theorem 1 proof

Theorem 1 *The* ModerateInternal *problem is NP-hard.*

Proof Our proof uses a reduction from the m-SubsetSum problem, where given a set of N positive integer numbers $v_1, \ldots, v_N$, a value m, and a target value b, we ask if there is a set of numbers B of size m, such that $\sum_{v_i \in B} v_i = b$.

Given an instance of the m-SubsetSum problem, we construct an instance of ModerateInternal as follows. The graph is a star with $N + 1$ nodes: we have a central node u_0, and a spoke node u_i for each integer v_i. For the center of the star (node u_0) we have that $w_{00} = t$, for an appropriately selected value of t (we will discuss this below), and $s_0 = -1$. The weight of the edge (u_0, u_i) from the center to node u_i is $w_{0i} = v_i$, and the weight of node u_i to its internal opinion is also $w_{ii} = v_i$. The opinion of all spoke nodes is $s_i = 1$. We set $k = N - m$, and we ask for a set of nodes T_s, $|T_s| = k$, such that, when setting $s_i = 0$ for $u_i \in T_s$ $\pi(\mathbf{z} \mid T_z) = \|\mathbf{z}\|^2$ is minimized.

The intuition of the proof is that the expressed opinion of the center node z_0 determines $\pi(\mathbf{z})$. The value of z_0 is determined by the weight t of the internal opinion of u_0, and the weights of the edges of nodes whose opinion is not set to zero. If we select t appropriately, we can guarantee that $\|\mathbf{z}\|^2$ is minimized when the nodes whose opinion is not set to zero sums to the value b.

Formally, assume that we have selected the set T_s, $|T_s| = k$. Assume that $u_0 \notin T_s$. Also let $R = V \backslash T_s \cup \{u_0\}$ denote the set of spoke nodes whose opinion was *not* set to 0. According to the opinion formation model, the equations for the expressed opinions of the spoke nodes are as follows. For every node $u_i \in R$, $z_i = \frac{z_0}{2} + \frac{1}{2}$. while for every node $u_i \in T_s$, $z_i = \frac{z_0}{2}$.

We can thus write:

$$\begin{aligned}\pi(\mathbf{z} \mid T_s) = \|\mathbf{z}\|^2 &= z_0^2 + k\frac{1}{4}z_0^2 + (N-k)\frac{1}{4}(z_0^2 + 2z_0 + 1) \\ &= \frac{N+4}{4}z_0^2 + \frac{N-k}{2}z_0 + \frac{N-k}{4}.\end{aligned}$$

Recall that we want to minimize $\pi(\mathbf{z} \mid T_s)$. To find the value of z_0 that minimizes $\pi(\mathbf{z} \mid T_s)$, we take the derivative of the expression above, we set it zero, and solve for z_0. We get that the value of z_0 that minimizes $\pi(\mathbf{z})$ is:

$$z_0^* = \frac{k - N}{N + 4}.$$

It follows that the minimum value of $\pi(\mathbf{z} \mid T_s)$ is

$$\pi^* = \frac{(N - k)(k + 4)}{4(N + 4)}.$$

We now set the value of t such that if the set of numbers in R sums to the value of b, then z_0 achieves the z_0^* value. First we compute the value of z_0 as a function of t. In the following we set $W = \sum_{i=1}^{N} v_i$. We have that:

$$\begin{aligned} z_0 &= \sum_{i=1}^{N} \frac{v_i z_i}{W + t} - \frac{t}{W + t} = \sum_{u_i \in T_s} \frac{v_i z_0}{2(W + t)} + \sum_{u_i \in R} \frac{v_i (z_0 + 1)}{2(W + t)} - \frac{t}{W + t} \\ &= \frac{\sum_{i=1}^{N} v_i}{2(W + t)} z_0 + \frac{\sum_{u_i \in R} v_i}{2(W + t)} - \frac{t}{W + t} = \frac{W}{2(W + t)} z_0 + \frac{\sum_{u_i \in R} v_i - 2t}{2(W + t)} \end{aligned}$$

Solving for z_0 we get:

$$z_0 = \frac{\sum_{u_i \in R} v_i - 2t}{W + 2t}.$$

We want the minimum to be achieved when $\sum_{u_i \in R} v_i = b$. Setting $z_0 = z_0^*$ we get:

$$\frac{b - 2t}{W + 2t} = \frac{K - N}{N + 4}$$

Solving for t we get:

$$t = \frac{(N + 4)b + (N - k)W}{2(k + 4)}.$$

Now, we want to prove the following. There is a set B of m numbers such that $\sum_{v_i \in B} v_i = b$, if and only if there is a set of nodes T_s of size $k = N - m$ such that when setting their internal opinion to zero, $\pi(\mathbf{z} \mid T_s) < \pi^* + \epsilon$ for some appropriate value of ϵ.

The forward direction is easy. If there exists this set B, then there is a set T_s such that when setting their opinions to zero, for the set R we have that

$$z_0 = \frac{\sum_{u_i \in R} v_i - 2t}{W + 2t} = \frac{b - 2t}{W + 2t} = \frac{k - N}{N + 4},$$

and therefore $\pi(\mathbf{z} \mid T_s) = \pi^*$.

For the backwards direction, if no such set of numbers exists, then it is not possible to find a set of nodes T_s such the nodes in R give $z_0 = \frac{K-N}{N+4}$ that minimizes $\pi(\mathbf{z} \mid T_s)$. Therefore, there must be an ϵ such that $\pi(\mathbf{z} \mid T_s) \geq \pi^* + \epsilon$.

To set ϵ note that for any $z_0 \neq z_0^*$

$$\left|z_0 - z_0^*\right| = \left|\frac{\sum_{u_i \in R} v_i - b}{W + 2t}\right| \geq \frac{1}{W + 2t} = \frac{k+4}{(N+4)(W+b)},$$

where the inequality follows from the fact that the values $v_1, \ldots, v_N, b$ are integers and their difference is at least one. Now, let $\mathbf{z}^*$ be the vector with z_0^* that achieves the minimum value π^*. For any other $\mathbf{z}$ we have

$$\begin{aligned}
\pi(\mathbf{z}) - \pi^* &= \frac{N+4}{4}\left(z_0^2 - (z_0^*)^2\right) + \frac{N-k}{2}(z_0 - z_0^*) \\
&= \left(z_0 - z_0^*\right)\left(\frac{N+4}{4} z_0 + \frac{N+4}{4} z_0^* - \frac{2(N+4)}{4}\frac{k-N}{N+4}\right) \\
&= \left(z_0 - z_0^*\right)\left(\frac{N+4}{4} z_0 - \frac{N+4}{4} z_0^*\right) = \frac{N+4}{4}\left(z_0 - z_0^*\right)^2 \\
&\geq \frac{N+4}{4}\left(\frac{1}{W+2t}\right)^2 = \frac{(k+4)^2}{4(N+4)(W+b)^2}.
\end{aligned}$$

So it suffices to set $\epsilon < \frac{(k+4)^2}{4(N+4)(W+b)^2}$.

Finally, in our computations so far we have assumed that our set T_s does not contain node u_0. This is not a restrictive assumption. Consider a solution T_s, where $u_0 \in T_s$, and $s_0 = 0$. Then, since s_0 is the only negative opinion value in our instance, it follows that $z_0 \geq 0$, and for any node $u_i \in R$ we have that $z_i = \frac{1}{2} z_0 + \frac{1}{2} \geq \frac{1}{2}$. There are $N + 1 - k$ nodes in R. Therefore,

$$\pi(\mathbf{z} \mid T_s) \geq \frac{N+1-k}{2}.$$

Note that $\pi^* = (N-k)(k+4)/4(N+4) \leq (N-k)/4$, since $k \leq N$. Therefore, $\pi(\mathbf{z}) \geq 2\pi^* + 1/4$. Selecting $\epsilon < \pi^* + \frac{1}{4}$ guarantees that $\pi(\mathbf{z}|T_s) > \pi^* + \epsilon$. Thus, if there is a set T_s such that $\pi(\mathbf{z}|T_s)$ is minimized, it cannot contain u_0. □

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 u.s. election: Divided they blog. In: International workshop on link discovery, LinkKDD

Akoglu L (2014) Quantifying political polarity based on bipartite opinion networks. In: International conference on weblogs and social media, ICWSM

Amelkin V, Singh AK, Bogdanov P (2015) A distance measure for the analysis of polar opinion dynamics in social networks. arXiv:1510.05058

Bakshy E, Messing S, Adamic L (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239):1130–1132

Bessi A, Zollo F, Vicario MD, Puliga M, Scala A, Caldarelli G, Uzzi B, Quattrociocchi W (2016) Users polarization on Facebook and Youtube. PLoS ONE 11(8):e0159641

Bindel D, Kleinberg JM, Oren S (2015) How bad is forming your own opinion? Games Econ Behav 92:248–265

Cambria E, Poria S, Bisio F, Bajpai R, Chaturvedi I (2015) The CLSA model: a novel framework for concept-level sentiment analysis. Springer International Publishing, Cham. doi:10.1007/978-3-319-18117-2_1

Cambria E, Poria S, Bajpai R, Schuller BW (2016) SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: 26th International conference on computational linguistics (COLING 2016), Proceedings of the conference: Technical Papers, Osaka, Japan, December 11–16, 2016, pp. 2666–2677

Chen T, Xu R, He Y, Xia Y, Wang X (2016) Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Comp Int Mag 11(3):34–44. doi:10.1109/MCI.2016.2572539

Conover M, Ratkiewicz J, Francisco MR, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. In: International conference on weblogs and social media ICWSM

Dandekar P, Goel A, Lee DT (2013) Biased assimilation, homophily, and the dynamics of polarization. Proc Natl Acad Sci 110(15):5791–5796

Davis G, Mallat S, Zhang Z (1994) Adaptive time-frequency decompositions with matching pursuits. Opt Eng 33(7):2183–2191

Del Vicario M, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2017) Modeling confirmation bias and polarization. Sci Rep 7:40391. doi:10.1038/srep40391

Feige U (2003) Vertex cover is hardest to approximate on regular graphs. Technical report MCS03-15 of the Weizmann Institute

Friedkin NE, Johnsen E (1990) Social influence and opinions. J Math Soc 15(3–4):193–206

Garimella K, Morales GDF, Gionis A, Mathioudakis M (2016) Quantifying controversy in social media. In: ACM international conference on web search and data mining, WSDM, pp 33–42

Garimella VRK, Morales GDF, Gionis A, Mathioudakis M (2017) Reducing controversy by connecting opposing views. In: ACM WISDOM international conference on web search and data mining

Garrett RK (2009) Echo chambers online? Politically motivated selective exposure among internet news users1. J Comput Mediat Commun 14(2):265–285. doi:10.1111/j.1083-6101.2009.01440.x

Gionis A, Terzi E, Tsaparas P (2013) Opinion maximization in social networks. In: SIAM international conference on data mining, pp 387–395

Guerra PHC, Jr, WM, Cardie C, Kleinberg R (2013) A measure of polarization on social media networks based on community boundaries. In: International conference on weblogs and social media, ICWSM

Hager WW (1989) Updating the inverse of a matrix. SIAM Rev 31(2):221–239

Isenberg DJ (1986) Group polarization: a critical review and meta-analysis. J Personal Soc Psychol 50(6):1141–1151

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: ACM SIGKDD international conference on knowledge discovery and data mining, pp 137–146

Lappas T, Crovella M, Terzi E (2012) Selecting a characteristic set of reviews. In: ACM SIGKDD international conference on knowledge discovery and data mining, pp 832–840

Lawrence P, Sergey B, Motwani R, Winograd T (1998) The pagerank citation ranking: bringing order to the web. Technical report, Stanford University

Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167

Mallat S (2008) A wavelet tour of signal processing, third edition: the sparse way, 3rd edn. Academic Press, Cambridge

Munson SA, Lee SY, Resnick P (2013) Encouraging reading of diverse political viewpoints with a browser widget. In: International conference on weblogs and social media, ICWSM

Munson SA, Resnick P (2010) Presenting diverse political opinions: how and how much. In: International conference on human factors in computing systems, CHI, pp 1457–1466

Natarajan BK (1995) Sparse approximate solutions to linear systems. SIAM J Comput 24(2):227–234

Pariser E (2011) The filter bubble: what the internet is hiding from you. The Penguin Group

Poria S, Cambria E, Gelbukh A (2016) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl Based Syst 108(C):42–49. doi:10.1016/j.knosys.2016.06.009

Sunstein CR (2002) The law of group polarization. J Polit Philos 10(2):175–195

Vicario MD, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) Modeling confirmation bias and polarization. arXiv:1607.00022

Vydiswaran V, Zhai C, Roth D, Pirolli P (2015) Overcoming bias to learn about controversial topics. J Assoc Inf Sci Technol 66(8):1655–1672