Peralta, Antonio F.; Kertesz, Janos; Iniguez, Gerardo

Opinion formation on social networks with algorithmic bias

Published in:
Journal of Physics: Complexity

DOI:
10.1088/2632-072X/ac340f

Published: 01/12/2021

Document Version
Publisher's PDF, also known as Version of record

Opinion formation on social networks with algorithmic bias: dynamics and bias imbalance

Journal of Physics: Complexity

OPEN ACCESS

PAPER

Opinion formation on social networks with algorithmic bias: dynamics and bias imbalance

Antonio F Peralta[1,*], **János Kertész**[1,2] **and Gerardo Iñiguez**[1,3,4]

[1] Department of Network and Data Science, Central European University, A-1100 Vienna, Austria
[2] Complexity Science Hub, A-1080 Vienna, Austria
[3] Department of Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland
[4] Centro de Ciencias de la Complejidad, Universidad Nacional Autonóma de México, 04510 Ciudad de México, Mexico
[*] Author to whom any correspondence should be addressed.

E-mail: peraltaaf@ceu.edu

Keywords: opinion dynamics, algorithmic bias, dynamical systems

RECEIVED
3 August 2021

REVISED
11 October 2021

ACCEPTED FOR PUBLICATION
27 October 2021

PUBLISHED
17 November 2021

Abstract

We investigate opinion dynamics and information spreading on networks under the influence of content filtering technologies. The filtering mechanism, present in many online social platforms, reduces individuals' exposure to disagreeing opinions, producing algorithmic bias. We derive evolution equations for global opinion variables in the presence of algorithmic bias, network community structure, noise (independent behavior of individuals), and pairwise or group interactions. We consider the case where the social platform shows a predilection for one opinion over its opposite, unbalancing the dynamics in favor of that opinion. We show that if the imbalance is strong enough, it may determine the final global opinion and the dynamical behavior of the population. We find a complex phase diagram including phases of coexistence, consensus, and polarization of opinions as possible final states of the model, with phase transitions of different order between them. The fixed point structure of the equations determines the dynamics to a large extent. We focus on the time needed for convergence and conclude that this quantity varies within a wide range, showing occasionally signatures of critical slowing down and meta-stability.

1. Introduction

The collective behavior of a system made of interacting individuals can be successfully analyzed using agent-based models, as shown in many examples across various disciplines [1–3]. In these models, individuals (or agents) are often pictured as nodes in a network [4, 5], where the links represent the possible interactions between them. Each node holds a dynamical state variable whose interpretation depends on the context of the model. In opinion dynamics this opinion variable [1] can be considered as the political party preference of an individual (e.g., liberal or conservative), her inclination towards or against some regulation or initiative, etc. The usefulness of the approach lies in the simplicity of the setting, leading, nevertheless, to complex phenomena due to collective effects. The possibility of considering various elements that are hypothesized to be relevant for opinion formation (such as structural and dynamical heterogeneities) deeply improves our understanding of the underlying social phenomena.

In recent years, human communication has changed dramatically, and moved from traditional media (face-to-face, phone, or mass media like television and the press) to online social media platforms (Google, Twitter, Facebook, etc) [6, 7]. In contrast to earlier media channels, online social networks control the information that users see and send to each other by means of personalized filtering algorithms [8]. These algorithms record individual information about users' preferences and then filter incoming data accordingly [9, 10]. Hence, people tend to be exposed to opinions they already agree with, producing so-called *algorithmic bias* [11, 12]. This reinforcement feature changes, ultimately, the global behavior of the population [13–16], promoting phenomena like 'filter bubbles' or 'echo chambers', where people divide in groups with opposing views that barely interact with each other. Explaining how and when the polarization of opinion groups emerges within a population is of crucial importance [17, 18]. It is of particular interest to understand how copying or herding

processes (typical signatures of human social behavior), when coupled to algorithmic bias, may enhance or decrease polarization. These two ingredients can be implemented in the formalism of agent-based models [1] leading to a flexible mathematical framework open to analytical and numerical treatments. In our simplified version of an agent based model the only source of heterogeneity is network structure. Ultimately, the results of such models can be interpreted to help devise potential strategies to mitigate the negative effects of algorithmic bias.

Previous modeling efforts [19] have been made to consider algorithmic bias in bounded confidence models [20], where the opinion variable of individuals is a real continuous variable on an interval. The filtering algorithm requires the opinions of two individuals to be similar enough to be able to interact, and bias means that similar people with similar opinions have a greater chance to meet, leading to enhanced polarization and fragmentation in opinion space. Another class of models considers opinions to be discrete (a binary variable in the simplest case) [21, 22]. Perra and Rocha [23] have studied algorithmic bias in such a model by considering that the opinion of an individual is influenced by its neighbors in the network, and filtered in various ways. An alternative implementation of algorithmic bias in binary-state models has been proposed in [24]. In this case, the social platform records information about all the previous opinions of individuals, and then influences them to keep the opinion that has been held for the longest time, similarly to a memory or 'inertia' effect [25–27]. All these implementations of algorithmic bias in opinion dynamics modeling suggest that polarization is a consequence of both the social behavior of individuals and the content filtering algorithms constraining their actions.

In the present work, similarly to the approach of [23], we consider that a fraction of the neighbors of an individual holding disagreeing opinions are filtered, and thus interactions with those neighbors are not possible. Recently we have proposed a general formalism within the binary-state approach that includes this implementation of algorithmic bias [28]. We have extended previous theoretical tools to describe the macroscopic dynamics on networks, including mean-field and pair approximations. We have also explored modular community structures, a crucial ingredient to characterize opinion polarization, i.e. the division of the population into opinion groups. We have studied the static, asymptotic behavior of archetypal models of opinion formation in the presence of algorithmic bias and concluded that, systematically, pairwise interactions lead to polarization, while group interactions promote coexistence of opinions. Here we use the same formalism of [28], but focus on dynamical aspects of the opinion formation process. We also extend the algorithmic bias mechanism to consider situations in which the online social platform promotes one opinion over the other, thus unbalancing the dynamics in favor of the preferred opinion by the platform, in what we call *bias asymmetry*.

We use a prototypical model of social behavior, the language model [29, 30], which takes into account both pairwise and group-based (copying) interactions depending on the value of a tunable parameter α. We also include the possibility that individuals act independently of their neighbors [31–33], which we denote as noise with intensity Q. This general framework enables us to consider several interaction mechanisms and leads to various opinion formation scenarios. In [28], we have considered other archetypal models of opinion formation (including voter-like and majority-vote models), and realized that the language model essentially interpolates between dynamics with either pairwise or group interactions depending on the value of α, and is thus a good candidate to explore the effects of bias asymmetry within a single model.

The paper is organized as follows. In section 2 we define the opinion formation model, algorithmic bias, and social network with community structure we use. In section 3 we derive a set of mean-field equations that describe the global dynamics of the model, and derive its associated fixed points (stationary states of the dynamics). In section 4 we explore the local dynamics and stability of the fixed points and build the phase diagram of the model. In section 5 we present a detailed study of temporal behavior of the considered opinion dynamics model using numerical simulations and some theoretical tools, with particular emphasis on the role of initial conditions, and the behavior of the time to reach the final state. Throughout this paper we will pay special attention to the effect of algorithmic bias and its asymmetry.

2. Model and definitions

We consider the formalism of binary-state dynamics [21, 22] as basic ground for modeling opinion formation. The model system is composed of a set of $i = 1, \ldots, N$ individuals, each one holding a binary-state (opinion) variable $s_i(t) = 0, 1$ at time t (e.g., liberal or conservative in a political setting). We define the macroscopic state (global opinion) of the system as $\rho = N^{-1}\sum_{i=1}^{N} s_i \in [0, 1]$, i.e. the density of individuals in state 1. Individuals are represented by nodes of an (undirected) network, and links in the network correspond to some social relationship between them, such that the opinion of an individual can be influenced by its neighbors in the network. The state of node i changes according to rates depending on the specific dynamics and the network structure: with '*infection*' rate F_{k_i,m_i} from $s_i = 0 \to 1$, and with '*recovery*' rate R_{k_i,m_i} from $s_i = 1 \to 0$, where k_i is

the degree of the node in the network and $m_i \in [0, k_i]$ is the number of (nearest) neighbors of i in state $s_j = 1$. (Here the names of the rates refer to the analogy with epidemic spreading.)

In the following we specify the spreading dynamics, i.e. the functional form of the rates $F_{k,m}$ and $R_{k,m}$, and the network structure. As for the spreading dynamics, we incorporate algorithmic bias (representing content filtering as implemented in many online social platforms) that influences and controls the way people interact. For the network structure, we include modules or communities by dividing the population into groups with tunable connectivity.

2.1. Transition rates in the language model

As mentioned above, we focus on the *language model* [29, 30], which is able to describe both pairwise and group interactions. The transition rates are as follows:

$$F_{k,m} = Q + (1 - 2Q)\left(\frac{m}{k}\right)^{\alpha}, \tag{1}$$

$$R_{k,m} = Q + (1 - 2Q)\left(\frac{k - m}{k}\right)^{\alpha}, \tag{2}$$

with $Q \in [0, 1/2]$ and $\alpha \in (0, \infty)$ as tuning parameters. The model takes into account two mechanisms driving the dynamics: (i) noisy or idiosyncratic changes of state, with intensity Q; and (ii) herding or copying the states of neighbors with probability proportional to the fraction of neighbors in the opposite state to a power α. The rates in equations (1) and (2) were first studied in the case without noise ($Q = 0$) to model the dynamics of language death [29], but the same model has been applied to other types of social human behavior, for example in the context of opinion formation in social media [34]. The role of noise ($Q > 0$) has been extensively studied of late, as for the *non-linear noisy voter model* in [33] with α a real (continuous) number, and for the *q-voter model* [35–37] with $\alpha = q$ a positive integer. The language model encapsulates a wide variety of phenomena depending on the value of α. E.g., for the mean field (complete network) the following regions can be distinguished (i) low $0 < \alpha < 2$ (*pairwise* interactions); (ii) high $2 < \alpha < 5$, (*group* interactions); and (iii) very high $\alpha > 5$. Each of these cases displays a distinct phenomenology [28, 33] and represents a different archetypal way for humans to influence each other (either in pairs or in groups). Note that the separation between pairwise (low α) and group (high α) behaviors, mentioned throughout the paper, occurs smoothly (as a crossover and not as a sharp transition) for an intermediate value of $\alpha \approx 2$, see [28]. This distinction will help us sort out and interpret the results of the model and the possible effects of algorithmic bias.

Note that the original rates in equations (1) and (2) fulfill the 'up-down symmetry' condition $R_{k,m} = F_{k,k-m}$ but, as we will show next, equations (3) and (4) are only symmetric for $b_0 = b_1$, where b_0 and b_1 are formally defined in the next section 2.2. In other words, unbalanced algorithmic bias breaks the symmetry of the system and favors one opinion over the other.

2.2. Algorithmic bias

A simple implementation of algorithmic bias has been proposed by us in a previous study [28]. Here we generalize the definition of that paper by introducing two parameters characterizing the bias, instead of one. The *bias intensities* b_0 and b_1 (where the subscripts refer to state 0 or 1) take values in the interval $[0, 1]$. These parameters measure the probabilities that the online platform filters out a neighbor in the opposite state, 0 or 1, (disagreeing opinion) of an individual with a given opinion, so that further interactions with that neighbor cannot take place. This mechanism of content filtering can be implemented in the formalism of any rate governed binary-state model by considering the following effective transition rates $F^*_{k,m}$, $R^*_{k,m}$:

$$F^*_{k,m}(b_1) = \sum_{i=0}^{m} B_{m,i}(1 - b_1)F_{k-m+i,i}, \tag{3}$$

$$R^*_{k,m}(b_0) = \sum_{s=0}^{k-m} B_{k-m,s}(1 - b_0)R_{m+s,m}, \tag{4}$$

with the binomial $B_{k,m}(1 - b) = \binom{k}{m}(1 - b)^m b^{k-m}$. Equations (3) and (4) express the average rates of changing state, after removing with probability b_0 or b_1 a subset of neighbors in the opposite state (b_0 if neighbors are in state $s = 0$ and b_1 for $s = 1$). We define the *total bias intensity* $b = (b_0 + b_1)/2$, and the *bias asymmetry* $\Delta b = b_1 - b_0$, with $b_0 = b - \Delta b/2$ and $b_1 = b + \Delta b/2$. For $\Delta b > 0$ the social platform favors $s = 0$, while for $\Delta b < 0$ it favors $s = 1$. In [28] we have implemented bias in a similar way to equations (3) and (4), but with $b_0 = b_1 = b$ (i.e. $\Delta b = 0$), such that the role of algorithmic bias is symmetric, or balanced across opinions.

2.3. Modular structure

The social network of interactions between individuals is fully specified by the adjacency matrix A_{ij}, with elements equal to 1 if i and j are connected and 0 otherwise. For simplicity we consider the degree distribution P_k as the only relevant structural feature of the network, with average degree $z = \sum_k P_k k$. We use the standard configuration model [38] to produce synthetic networks in the corresponding numerical simulations.

When the network displays modular (community) structure, we may classify individuals into groups with higher connectivity to nodes inside the group than to those outside. There is no unique definition of community, and for this reason there are several algorithms of community detection in networks that may lead to different results [39]. Regardless of the details of the chosen definition, link density inside a community should be higher than between communities. In social networks, homophily (similarity between some node attributes) is one of the main tie formation mechanisms leading to communities [40–42]. Since opinion dynamics might be much faster than the homophilic processes leading to community structure, we consider such structure as static.

When the population is divided in two asymmetric groups, with different sizes and connectivity, we refer to them as the majority and minority groups. Assuming for simplicity two modules, nodes $i = 1, \ldots, N_1$ are in group 1 of size N_1, and nodes $i = N_1 + 1, \ldots, N_1 + N_2 = N$ in group 2 of size N_2 (nodes can only belong to one group). The two groups have different connectivity depending on whether links join nodes of the same or different groups. In this way, two nodes in the same group are more likely to be connected than if they belong to different groups. In order to characterize the macroscopic state accordingly, we need at least two variables: $\rho_1 = N_1^{-1}\sum_{i=1}^{N_1} s_i \in [0,1]$ and $\rho_2 = N_2^{-1}\sum_{i=N_1+1}^{N_2} s_i \in [0,1]$, the density of nodes in state 1 in groups 1 and 2, respectively, with total $\rho = \frac{N_1}{N}\rho_1 + \frac{N_2}{N}\rho_2 \in [0,1]$. Additionally, we define *polarization* as $P = |\rho_1 - \rho_2| \in [0,1]$, which measures the degree of opinion dissimilarity between groups.

We consider four average degrees, z_1, z_{12}, z_{21}, and z_2, defining the connectivity inside and between groups. Parameter z_1 (z_2) is the average degree only considering links that join nodes within group 1 (2), while z_{12} (z_{21}) is the average degree only considering links that depart from group 1 and end up in group 2 (from 2 to 1). The total number of links that go from group 1 to 2 is the same as those that go from 2 to 1, so we have the constraint $N_1 z_{12} = N_2 z_{21}$.

3. Mean-field description

We derive a set of mean-field evolution equations for the two macroscopic variables $\rho_1(t)$ and $\rho_2(t)$, of general validity in the thermodynamic ($N \to \infty$) and highly connected ($z_1, z_{12}, z_{21}, z_2 \to \infty$) limit, with constant ratios of the average degrees. Even in a finite system with high connectivity, the mean field description is a good approximation of the dynamics, and it captures the phenomenology of the model well. For large values of N, the opinion variables fluctuate slightly around their average values, i.e., $\rho_1(t) \approx \langle \rho_1(t) \rangle$, $\rho_2(t) \approx \langle \rho_2(t) \rangle$. Thus, throughout the following mean-field description, $\rho_1(t)$ and $\rho_2(t)$ refer to average values over realizations.

In order to derive the mean-field equations [28] we first define the average rate of changing state [22] as

$$f[x] \equiv \sum_k \frac{P_k k}{z} \sum_{m=0}^{k} F_{k,m} B_{k,m}(x), \tag{5}$$

where x is the probability of finding a neighbor in state 1, and $P_k k/z$ is the probability that a link connects to a node with degree k. In order to obtain a closed description of the dynamics, we must relate the probability x to the description variables ρ_1 and ρ_2. The probability x depends on the group to which the node belongs. In the case of group 1 we have

$$x_1 = \frac{N_1 z_1 \rho_1 + N_2 z_{21} \rho_2}{N_1 z_1 + N_2 z_{21}} = \frac{\rho_1 + p_1 \rho_2}{1 + p_1}, \tag{6}$$

with $p_1 = N_2 z_{21}/N_1 z_1 = z_{12}/z_1$, and, similarly, for group 2 exchanging the labels $1 \leftrightarrow 2$, with $p_2 = N_1 z_{12}/N_2 z_2 = z_{21}/z_2$. Equation (6) is the ratio of the number of links coming out of nodes in state 1 that connect to nodes in group 1, and the number of links coming out of nodes in group 1.

If we consider algorithmic bias ($b_1 > 0$), we must use the effective $F^*_{k,m}$ of equation (3) instead of $F_{k,m}$ to calculate the average rate $f^*[x]$ in the network using equation (5). Applying an argument based on the highly connected limit ($z \to \infty$) [28], the effective average rate with bias reduces to

$$f^*[x] \approx f\left[\frac{(1-b_1)x}{1-b_1 x}\right]. \tag{7}$$

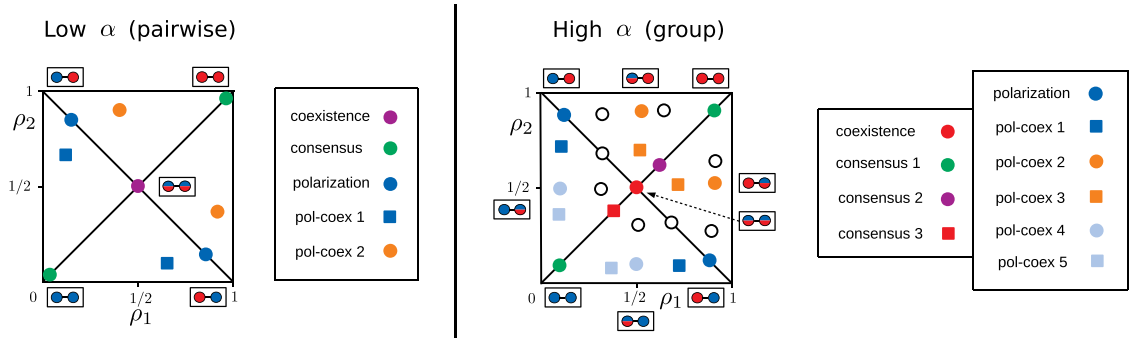

Figure 1. Schematic representation of fixed points of opinion dynamics in (ρ_1, ρ_2)-phase space. The left (right) side corresponds to the low (high) α regime of the model. Joined colored circles inside a square represent the opinion states $s = 0$ (blue) or $s = 1$ (red) of each group (e.g., two circles of the same color represent consensus, while circles of different colors represent polarization). The legend displays the color coding and names of the most relevant fixed points. Pairs of fixed points [circle (stable) and square (unstable)] have the same color if they disappear/appear together for specific values of the parameters.

When $b_1 = 0$ we recover $f^*[x] = f[x]$. An analogous procedure can be applied to the effective recovery rate $R^*_{k,m}$ of equation (4), leading to $r^*[x]$, which in the presence of up-down symmetry ($R_{k,m} = F_{k,k-m}$) is given by

$$r^*[x] \approx f\left[\frac{(1-b_0)(1-x)}{1-b_0(1-x)}\right]. \tag{8}$$

After defining the effective average rates $f^*[x]$, $r^*[x]$ and the probabilities $x_{1,2}$, we obtain a system of two differential (mean-field) equations for the dynamics of the state variables $\vec{\rho}(t)$ with components $\rho_1(t)$ and $\rho_2(t)$:

$$\frac{\mathrm{d}\rho_1}{\mathrm{d}t} = (1-\rho_1)f\left[\frac{(1-b_1)(\rho_1+p_1\rho_2)}{1+p_1-b_1(\rho_1+p_1\rho_2)}\right] - \rho_1 f\left[\frac{(1-b_0)(1-\rho_1+p_1(1-\rho_2))}{1+p_1-b_0(1-\rho_1+p_1(1-\rho_2))}\right] \equiv \mu_1[\rho_1,\rho_2], \tag{9}$$

$$\frac{\mathrm{d}\rho_2}{\mathrm{d}t} = (1-\rho_2)f\left[\frac{(1-b_1)(\rho_2+p_2\rho_1)}{1+p_2-b_1(\rho_2+p_2\rho_1)}\right] - \rho_2 f\left[\frac{(1-b_0)(1-\rho_2+p_2(1-\rho_1))}{1+p_2-b_0(1-\rho_2+p_2(1-\rho_1))}\right] \equiv \mu_2[\rho_1,\rho_2]. \tag{10}$$

This mean-field description of the opinion formation model has the social behavioral parameters (Q, α), the algorithmic bias parameters $(b, \Delta b)$, and the group connectivity parameters (p_1, p_2), together with the initial condition $\rho_1(0), \rho_2(0)$.

3.1. Fixed point structure and stationary solutions

The stationary results of equations (9) and (10), i.e. the infinite time limit $\rho_1(\infty)$ and $\rho_2(\infty)$, correspond to the stable fixed points. The fixed points ρ_1^{st}, ρ_2^{st} are obtained from the condition

$$\mu_1[\rho_1^{\mathrm{st}}, \rho_2^{\mathrm{st}}] = 0, \tag{11}$$

$$\mu_2[\rho_1^{\mathrm{st}}, \rho_2^{\mathrm{st}}] = 0. \tag{12}$$

The study of the solutions of equations (11) and (12) as a function of the parameters is a first step in understanding the dynamics and general behavior of the model. In figure 1 we show the positions in (ρ_1, ρ_2)-phase space of all possible fixed points in the model, together with color and name coding to identify and refer to them easily in the following sections and figures. Note the presence of the collective opinion states most relevant to our discussion: *consensus* ($\rho_1^{\mathrm{st}} = \rho_2^{\mathrm{st}} \approx 0$ or $\rho_1^{\mathrm{st}} = \rho_2^{\mathrm{st}} \approx 1$), *coexistence* ($\rho_1^{\mathrm{st}} = \rho_2^{\mathrm{st}} \approx 1/2$), and *polarization* ($\rho_1^{\mathrm{st}} \approx 0, \rho_2^{\mathrm{st}} \approx 1$ or $\rho_1^{\mathrm{st}} \approx 1, \rho_2^{\mathrm{st}} \approx 0$). All these states are possible in the low (pair) and high (group) α regimes for certain values of the model parameters $(Q, \alpha, b, \Delta b, p_1, p_2)$.

4. Local dynamics and stability

A basic step in understanding the dynamics of the model is to explore the vector fields of equations (9) and (10) and the associated trajectories close to the fixed points $\rho_1(t) \approx \rho_1^{\mathrm{st}}, \rho_2(t) \approx \rho_2^{\mathrm{st}}$. This can be done by means of a linearization of the dynamical equations. The linearization process leads to an exponential solution,

$$\rho_1(t) \approx \rho_1^{\mathrm{st}} + C_1 v_{11}\, \mathrm{e}^{-\lambda_1 t} + C_2 v_{21}\, \mathrm{e}^{-\lambda_2 t}, \tag{13}$$

$$\rho_2(t) \approx \rho_2^{\mathrm{st}} + C_1 v_{12}\, \mathrm{e}^{-\lambda_1 t} + C_2 v_{22}\, \mathrm{e}^{-\lambda_2 t}, \tag{14}$$

where $\vec{v}_1 = (v_{11}, v_{12})$ and $\vec{v}_2 = (v_{21}, v_{22})$ are the eigenvectors with associated eigenvalues λ_1 and λ_2 of the Jacobian matrix $J_{ij} = \frac{\partial \mu_i}{\partial \rho_j}$, evaluated at the fixed point $\rho_1^{\rm st}$, $\rho_2^{\rm st}$. C_1 and C_2 can be calculated from the initial condition as

$$C_1 = \frac{v_{21}\rho_2^* - v_{22}\rho_1^*}{v_{12}v_{21} - v_{11}v_{22}}, \tag{15}$$

$$C_2 = \frac{v_{12}\rho_1^* - v_{11}\rho_2^*}{v_{12}v_{21} - v_{11}v_{22}}, \tag{16}$$

with $\rho_1^* = \rho_1(0) - \rho_1^{\rm st}$ and $\rho_2^* = \rho_2(0) - \rho_2^{\rm st}$.

The stability of a fixed point is determined by the sign of (the real part of) the associated eigenvalues: for $\lambda_{1,2} < 0$ it is stable, for $\lambda_{1,2} > 0$ it is unstable, and for $\lambda_1 < 0$, $\lambda_2 > 0$ or $\lambda_1 > 0$, $\lambda_2 < 0$ it is a saddle point. Only when the fixed point is stable, we expect trajectories to converge to the fixed point in the long time limit $[\rho_1(t) \to \rho_1^{\rm st}, \rho_2(t) \to \rho_2^{\rm st}]$.

The eigenvalues can be calculated as

$$\lambda_{1,2} = \frac{1}{2}\left(\tau \pm \sqrt{\tau^2 - 4\Delta}\right), \quad \Delta = \lambda_1\lambda_2, \quad \tau = \lambda_1 + \lambda_2, \tag{17}$$

where $\tau = J_{11} + J_{22}$ is the trace and $\Delta = J_{11}J_{22} - J_{12}J_{21}$ the determinant of the Jacobian matrix evaluated at the fixed point $\rho_1^{\rm st}$, $\rho_2^{\rm st}$. If $\tau^2 > 4\Delta$ the eigenvalues only have a real part (the case for all fixed points in figure 1). The condition for a transition (bifurcation) is that one of the eigenvalues becomes zero (a so-called marginal stability), or equivalently $\Delta[\rho_1^{\rm st}, \rho_2^{\rm st}] = 0$. This condition together with equations (11) and (12) determines the transition lines and the phase diagrams. At a transition we expect some fixed points to appear or disappear (usually in couples).

In figure 2 we show the phase diagram and vectors fields in the prototypical scenario of pairwise interactions (low α) as a function of bias asymmetry Δb. The phase diagrams in figures 2(a) and (b) correspond to the well-known *cusp catastrophe* [43], where transitions are saddle node bifurcations in which two fixed points (stable and saddle point or saddle point and unstable) merge and disappear for high enough bias asymmetry. In the case of two equal groups ($p_1 = p_2$, figure 2(a)), when tuning bias asymmetry, polarization is destroyed favoring the consensus states. After that, for a specific high value of the asymmetry, one of the two consensus states disappears (figure 2(c)) and the only remaining state is $\rho_1^{\rm st} = \rho_2^{\rm st} \approx 0$ for $\Delta b > 0$, or $\rho_1^{\rm st} = \rho_2^{\rm st} \approx 1$ for $\Delta b < 0$. Every time a pair of fixed points merge, there is a region in phase space where the dynamics $\rho_1(t)$, $\rho_2(t)$ becomes very slow and meta-stable states appear (in section 5 we discuss this in more detail). Note that in the symmetric version of the model ($\Delta b = 0$), when more than one stable fixed point exists, initial conditions determine the final state. However, when the exogenous ingredient of bias asymmetry is introduced by the social media ($\Delta b \neq 0$), it is possible to 'select and control' the final opinion of the system.

Another relevant scenario is that of asymmetric groups $p_1 \neq p_2$, where one of them is either better connected and/or bigger in size, i.e. for $p_1 < p_2$ group 1 has more nodes or links than group 2, and the other way around for $p_1 > p_2$. The two polarized states ($\rho_1^{\rm st} \approx 0$, $\rho_2^{\rm st} \approx 1$ and $\rho_1^{\rm st} \approx 1$, $\rho_2^{\rm st} \approx 0$) are not symmetric, depending on which is the opinion of the majority and minority groups. For this reason, there are two transition lines in the phase diagram of figure 2(b) with cusps at different positions, one for $\Delta b > 0$ and the other for $\Delta b < 0$. Thus, bias asymmetry promotes (instead of destroying) polarization in the region between the two cusps. This result has a clear interpretation: if the social platform favors the opinion of the minority group, polarization will become stronger as it will be harder for the majority group to convince the other, leading the system towards consensus. The value of Δb at the cusp is the 'optimal' one if we wish to balance such majority-minority scenario (i.e., asymmetry in group sizes and connections) by using an exogenous algorithmic bias.

In figure 3 we plot the eigenvalues of all fixed points as a function of bias asymmetry in the same case specified in figure 2. The eigenvalues provide us with a lot of information about the nature of the fixed points and the dynamics close to them. The sign of the eigenvalues in figure 3 agree with the schematic representation of the vector fields in figure 2(c), and it determines the stability analysis of the fixed points. Every time an eigenvalue becomes zero, a pair of fixed points disappears, defining a transition or bifurcation.

In figure 4 we show the phase diagram and vector field in the case of group interactions (high α) as a function of bias asymmetry Δb. Figures 4(a) and (b) correspond to the so-called *butterfly catastrophe* [43]. We observe some differences with respect to the pair interaction case, besides the rich phenomenology of additional fixed points. The first difference is that the coexistence and consensus states can be both stable for some parameter values, at odds with results in figure 2. With respect to the dependence of the consensus states on bias asymmetry, for low noise we have a similar behavior as for pair interactions, while for high noise it is possible that a consensus state, which is not observed for $\Delta b = 0$, appears for some value $\Delta b \neq 0$. Standard polarization is also destroyed for a critical value of the asymmetry in the group interaction case. The difference is that new stable polarized states are possible in the group case (pol-coex 2 and pol-coex 4 in figure 1), whose

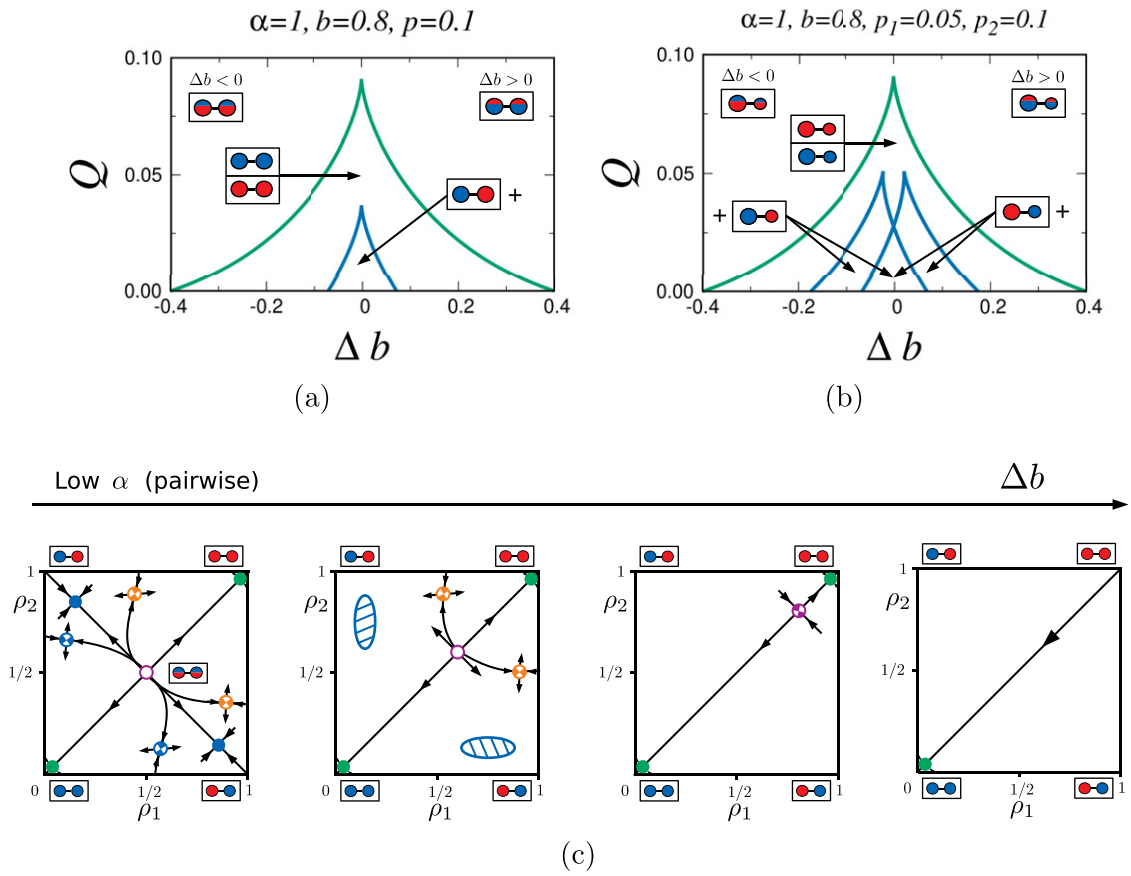

Figure 2. Phase diagrams for (a) $p_1 = p_2 = p = 0.1$, and (b) $p_1 = 0.05$, $p_2 = 0.1$, and vector fields (c) for fixed values of the model parameters $\alpha = 1$ (linear or pairwise regime) and $b = 0.8$. In the phase diagrams (a) and (b) the varying parameters are $(Q, \Delta b)$, i.e. the noise and bias asymmetry. The transition lines (green and blue) delimit the parameter regions where the different possible fixed points are stable. Circles inside a square indicate the corresponding fixed point following the scheme of figure 1. The phase diagram (a) corresponds to the case of two equal groups, while in (b) there is a majority (big circle) and minority (small circle) group. In the region below the green line both consensus states are stable, and above only one of them remains, while below the blue line polarization is stable. In panel (c), the left vector field is a typical situation below the blue line of the phase diagram (a) for $\Delta b = 0$. The other vector fields (from left to right) show how this changes as we increase Δb and cross the various transition lines. The elliptical striped zones are regions where the dynamics is very slow and meta-stable states are possible (see section 5).

behavior with respect to bias asymmetry is non-trivial. Similarly to the consensus states, these new polarized states may appear for a particular value of the asymmetry, even though they are not present in the symmetric case.

In figure 5 we show the eigenvalues of the fixed points of figures 1 and 4(c) as a function of Δb for different values of the noise Q. These figures provide us with information about the stability, dynamics and transitions that are possible in this case. Note that, depending on the value of Q, we may find some of the fixed points or not, and that different transitions happen for specific values of Δb, in accordance with the phase diagram in figure 4(b). The vector field scheme in figure 4(c) is a low-noise scenario where all fixed points are present (figure 5(c)). For other values of the noise Q (in figures 5(a) and (b)), not all the fixed of figure 4(c) are possible, and the transitions may occur in different orders as we increase Δb.

5. Global dynamics, convergence times and meta-stable states

The global dynamics $\rho_1(t)$, $\rho_2(t)$ of the system has a non-trivial dependence on the initial condition $\rho_1(0)$, $\rho_2(0)$ and the model parameters $(Q, \alpha, b, \Delta b, p_1, p_2)$, besides time t. The determination of the fixed points, the stability, and the local (linearized) dynamics are a good guideline to predict and understand, at least qualitatively, the dynamical behavior of the system. There are other aspects of the dynamics that cannot be fully explained by the fixed points and the linear dynamics approach. Among these, we study with particular attention what we call *meta-stable states*, where the dynamics slows down strongly and stays for a long time around a determined value of the state variables. This phenomenon is observed in the model, especially when two fixed points merge and disappear (the elliptical striped zones in figures 2(c) and 4(c)). As there are no fixed points

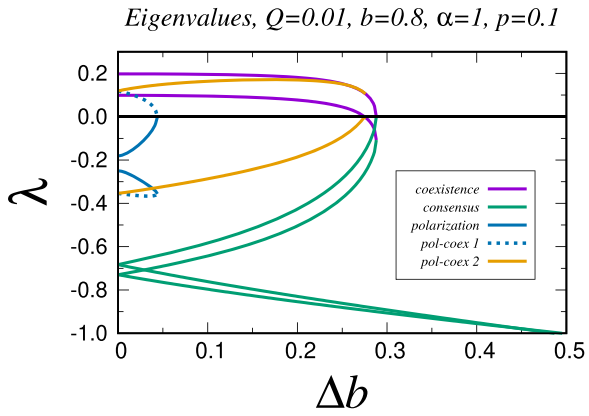

Figure 3. Eigenvalues of the various fixed points of figures 1 and 2(c) for parameter values $\alpha = 1$ (linear or pairwise regime), $b = 0.8$, $p_1 = p_2 = p = 0.1$ and $Q = 0.01$, as a function of the bias asymmetry Δb. The color and name coding in the legends is equivalent to that of figure 1. The lines of the same color (blue) correspond to a pair of polarized fixed points, dashed (saddle point) and solid (stable), that merge together and disappear for a particular value of Δb.

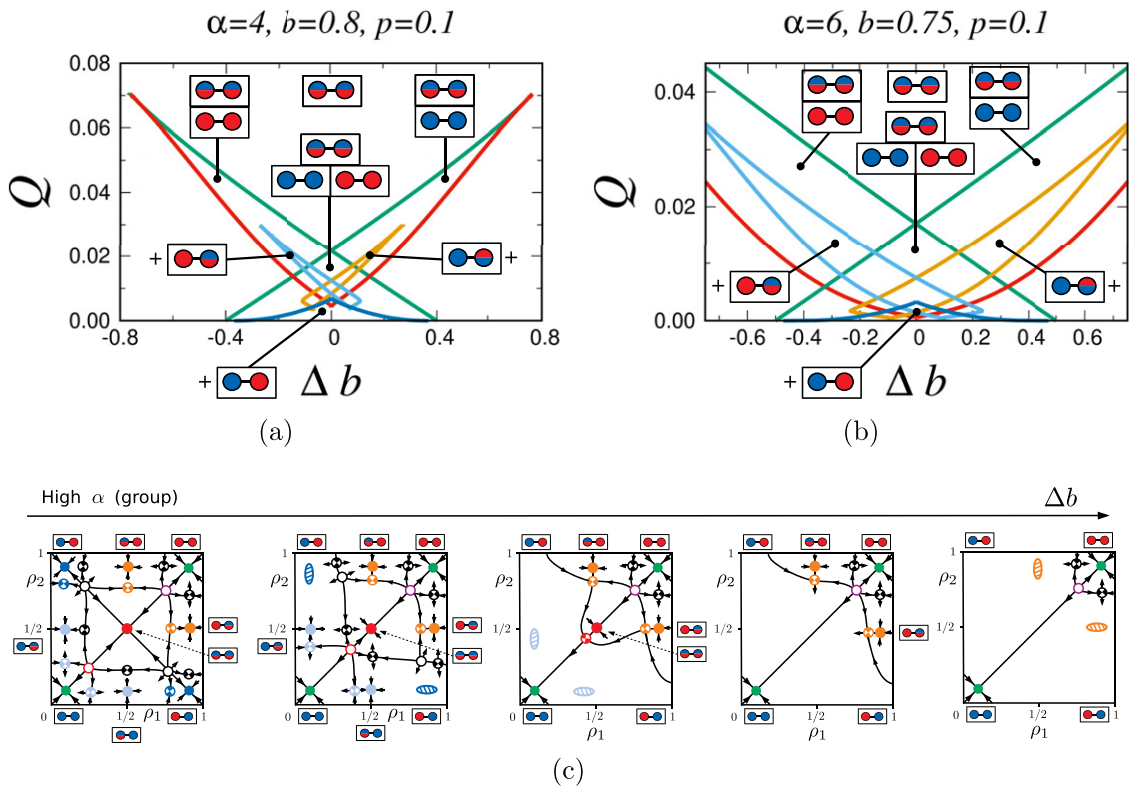

Figure 4. Phase diagrams for (a) $\alpha = 4$, $b = 0.8$, and (b) $\alpha = 6$, $b = 0.75$, and vector fields (c) for fixed connectivity $p_1 = p_2 = p = 0.1$. In the phase diagrams (a) and (b) the varying parameters are $(Q, \Delta b)$, i.e., the noise and bias asymmetry. The transition lines (green, red, dark and light blue, and yellow) delimit the parameter regions where fixed points are stable. Circles inside a square indicate the corresponding fixed points following the scheme of figure 1. In panel (c), the left vector field is a typical situation below the blue line (and above the red, light blue and yellow lines) of the phase diagram (b) for $\Delta b = 0$. The other vector fields from left to right show how this changes as we increase Δb and cross the various transition lines. The elliptical striped zones are regions where the dynamics is very slow and meta-stable states are possible (see section 5).

around these zones, it is not possible to evaluate the eigenvalues and explore the local dynamics. Thus, we need to use a different theoretical method to characterize the meta-stable states. We also explore the time needed to reach the final states and the dependence on the initial conditions, theoretically and by means of Monte Carlo simulations.

5.1. Numerical simulations

Before introducing the theoretical description of the meta-stable states, we analyze the results coming from Monte Carlo simulations. Implementing the rules of the model (section 2), we obtain stochastic trajectories

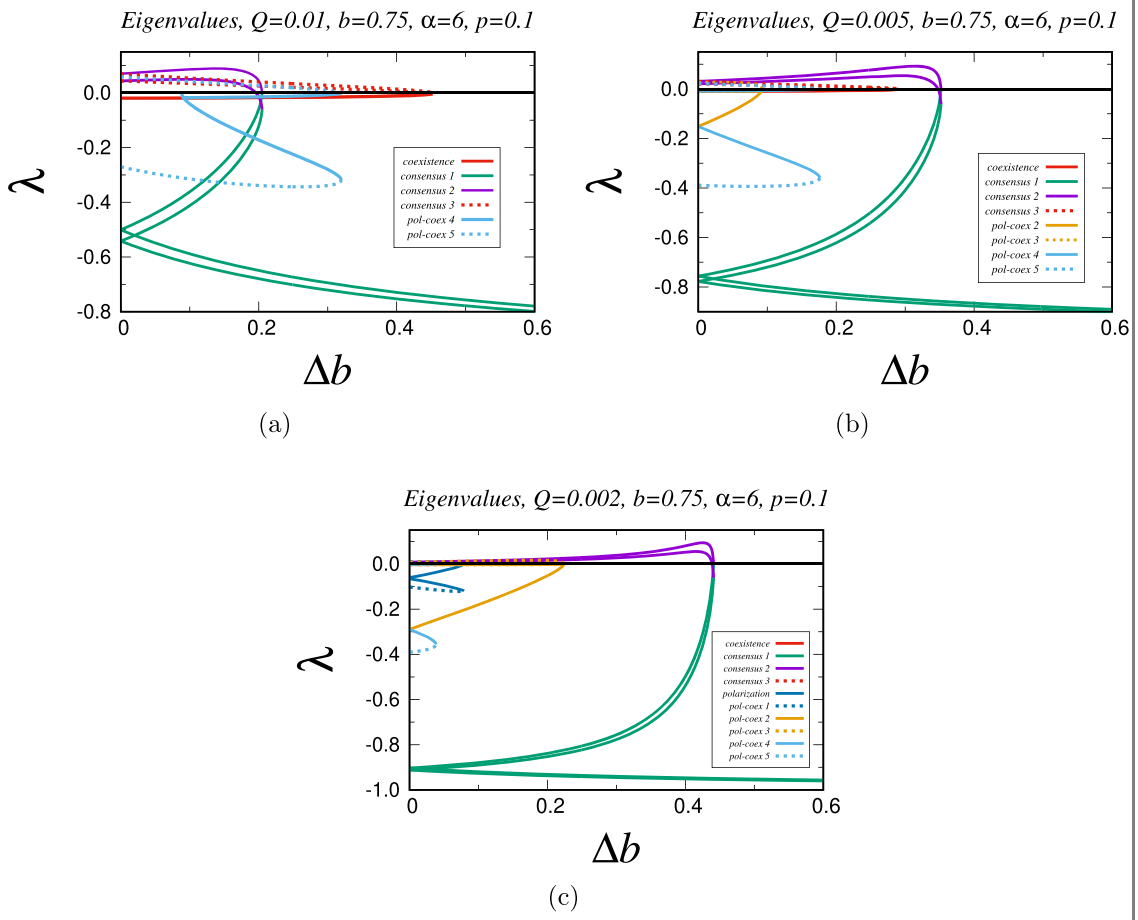

Figure 5. Eigenvalues of the various fixed points of figures 1 and 4(c) for parameter values $\alpha = 6$ (group regime), $b = 0.75$, $p_1 = p_2 = p = 0.1$ and $Q = 0.01, 0.005, 0.002$, as a function of bias asymmetry Δb. The color and name coding in the legends is equivalent to that of figure 1. Lines of the same color (light and dark blue, red and yellow) correspond to a pair of fixed points, dashed (saddle point or unstable) and solid (stable), that merge together and disappear for a particular value of Δb.

$\rho_1(t)$, $\rho_2(t)$, from which we calculate the average global state of the system $\langle \rho(t) \rangle$ and polarization $\langle P(t) \rangle$ that characterize the dynamics. In figures 6 and 7 we show numerical results for pair interactions ($\alpha = 1$) on a large ($N = 20\,000$) $z-$regular type of network with modular structure. In order to compare the simulations with the phenomenology coming from the theory, we use the same parameter values as in figures 2(a) and 3, i.e. $Q = 0.01$, $b = 0.8$ and $p_1 = p_2 = 0.1$.

In figure 6 we show the dynamics coming from numerical simulations of the model for various homogeneous initial conditions and bias asymmetries. In the top panels, from left to right, one of the consensus states becomes unstable for a determined value of the bias asymmetry and then, independently on the initial condition, all trajectories evolve towards the remaining stable consensus state. This corresponds to crossing (horizontally) the green transition line in the phase diagram of figure 2(a). Note that before the transition, when the two consensus states are possible, and depending on the initial condition $\rho_1(0) = \rho_2(0) = \rho(0)$, the dynamics evolves towards one state or the other. We can thus define a threshold initial condition ρ_0 that separates the basin of attraction of the consensus states. This threshold depends on the bias asymmetry $\rho_0(\Delta b)$: for no asymmetry ($\Delta b = 0$) it is $\rho_0 = 0.5$, it increases for $\Delta b > 0$ (decreases for $\Delta b < 0$), and it is not defined above the transition point as only one consensus state is stable. In figure 7, we show the dynamics coming from numerical simulations of the model for different polarized initial conditions and bias asymmetries. In the top panels, from left to right, the polarized state becomes unstable for a determined value of the bias asymmetry and then all trajectories evolve towards a non-polarized ($P = 0$) consensus state. This corresponds to crossing (horizontally) the blue transition line in the phase diagram of figure 2(a). Note that the dynamical results shown in figures 6 and 7 are in good agreement with the qualitative description of the vector fields in figure 2(c).

A significant dynamic phenomenon observed in figures 6 and 7 (specially in panels 6(b) and 7(b)–(d)) is the presence of trajectories that get trapped for a long period of time in some state, but eventually get released and end in one of the possible final (stable) states. In figure 7(d) we see that such a meta-stable state appears above the transition point Δb^*, and that its duration decreases as we increase the asymmetry. We represent

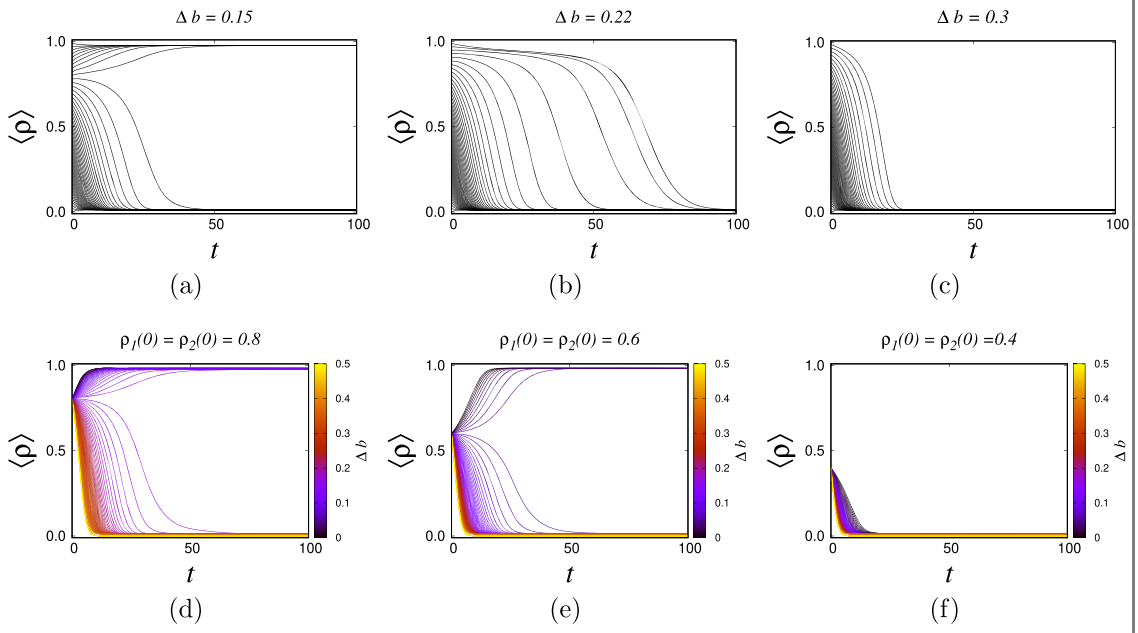

Figure 6. Average density $\langle \rho \rangle$ of nodes in state 1 as a function of time t coming from numerical simulations of the model, starting from a homogeneous initial condition $\rho_1(0) = \rho_2(0) = \rho(0)$. The model and connectivity parameters are $\alpha = 1$, $b = 0.8$, $Q = 0.01$, $p_1 = p_2 = 0.1$, $z_1 = z_2 = 18$, $z_{12} = z_{21} = 2$, and the size of the network is $N = 20\,000$, divided in two groups (communities) of equal size ($N_1 = N_2 = N/2$), averaged over 1000 realizations of the dynamics. In the top panels (a)–(c), trajectories correspond to various initial conditions $\rho(0)$ for a fixed value of the bias asymmetry Δb (specified in the title). In the bottom panels (d)–(f), a color gradient is used for bias asymmetry Δb in the range $[0, 0.5]$ for a fixed initial condition $\rho_1(0)$, $\rho_2(0)$ (specified in the title).

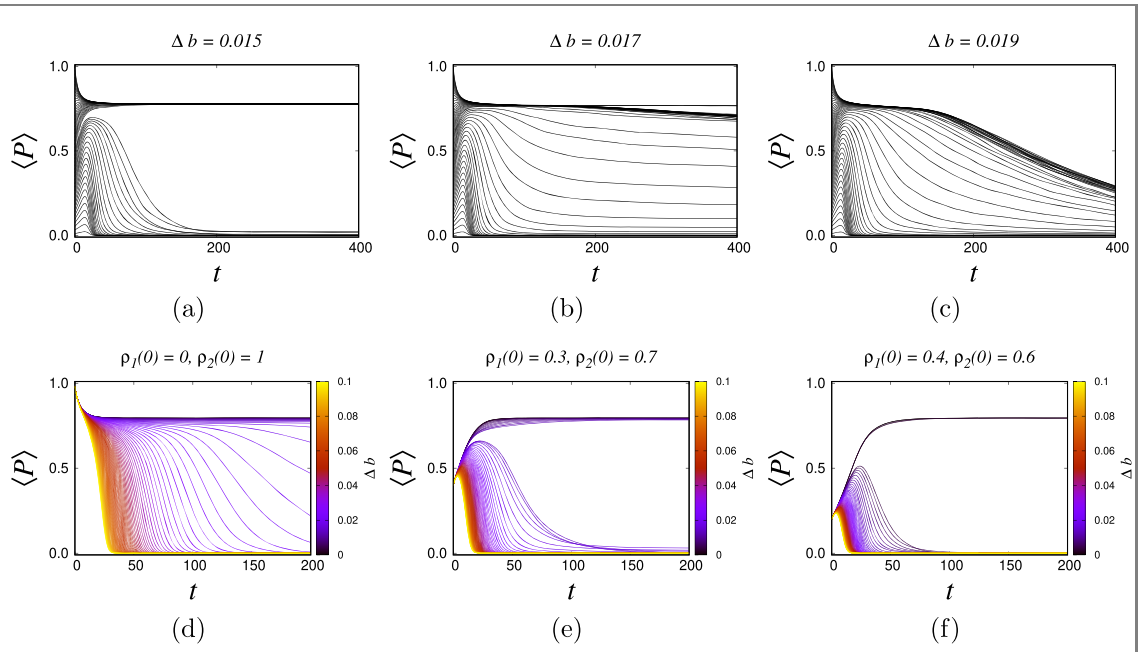

Figure 7. Average polarization $\langle P \rangle = \langle |\rho_1 - \rho_2| \rangle$ as a function of time t coming from numerical simulations of the model, starting from a polarized initial condition $\rho_2(0) = 1 - \rho_1(0)$. All model, connectivity and network parameters are the same as in figure 6. In the top panels (a)–(c) a fixed bias asymmetry is used for different initial conditions, while in the bottom panels (d)–(f) a fixed initial condition is used for different bias asymmetries (colors).

this meta-stable state as an elliptical stripped zone in the vector fields figure 2(c), and we characterize them theoretically in what follows (section 5.2).

5.2. Saddle node bifurcations and meta-stability

Close to a critical (bifurcation) point, it is possible to obtain the normal form of the dynamics [44], which goes beyond the linearization of equations (13) and (14). Assume that we have a fixed point ρ_1^{st} and ρ_2^{st} with an

eigenvalue equal to zero $\lambda_1 = 0$, for a determined value of a tuning parameter, e.g. $\Delta b = \Delta b^*$ (bifurcation or critical point). We perform a change of variables to the eigenvector basis $\vec{u}(t) = P^{-1}\vec{\rho}(t)$, where the columns of the matrix P are the eigenvectors $\vec{v}_{1,2}$, that is,

$$P = \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{pmatrix}. \tag{18}$$

The evolution of the transformed variables $\vec{u}(t)$ becomes

$$\frac{\mathrm{d}\vec{\rho}}{\mathrm{d}t} = \vec{\mu}(\Delta b, \vec{\rho}) \to \frac{\mathrm{d}\vec{u}}{\mathrm{d}t} = \vec{U}(\Delta b, \vec{u}) \equiv P^{-1}\vec{\mu}(\Delta b, P\vec{u}). \tag{19}$$

The transformed vector field fulfills $\vec{U}(\Delta b^*, \vec{u}_{\mathrm{st}}) = 0$ (here $\vec{u}_{\mathrm{st}}$ refers only to the fixed point at the bifurcation point $\Delta b = \Delta b^*$) and $\partial U_i/\partial u_j = -\lambda_i \delta_{ij}$ (again at Δb^* and $\vec{u}_{\mathrm{st}}$), i.e., the linear part is uncoupled. In this case, when $\lambda_1 = 0$, there is a center manifold $\vec{u}(t) = \vec{h}(\Delta b, u_1(t))$, i.e. a special trajectory where the time dependence of all variables is governed by the slow $u_1(t)$. This satisfies $\vec{u}_{\mathrm{st}} = \vec{h}(\Delta b^*, u_1^{\mathrm{st}})$ and the orthogonality condition $\partial h_i/\partial u_1 = 0$. The center manifold can be obtained from equation (19) as a series expansion. Once the functions $\vec{h}(\Delta b, u_1)$ have been determined, we obtain a single equation for $u_1(t)$,

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= U_1(\Delta b, \vec{h}(\Delta b, u_1)) \\ &= \beta^{(10)}(\Delta b - \Delta b^*) + \beta^{(11)}(\Delta b - \Delta b^*)(u_1 - u_1^{\mathrm{st}}) + \beta^{(02)}(u_1 - u_1^{\mathrm{st}})^2 \\ &\quad + \beta^{(03)}(u_1 - u_1^{\mathrm{st}})^3 + \cdots, \end{aligned} \tag{20}$$

where $\beta^{(10)}$, $\beta^{(11)}$, $\beta^{(02)}$, $\beta^{(03)}$ are coefficients whose expressions can be derived[5] from the series expansion of the center manifold [45].

We consider the most common bifurcation found as a function of the bias asymmetry Δb (see figures 2(c) and 4(c)), i.e. the saddle node bifurcation with $\beta^{(10)} \neq 0$ and $\beta^{(02)} \neq 0$. In its simplified form we have

$$\frac{\mathrm{d}u_1}{\mathrm{d}t} = \beta^{(10)}(\Delta b - \Delta b^*) + \beta^{(02)}(u_1 - u_1^{\mathrm{st}})^2, \tag{21}$$

where the higher order terms can be disregarded. Assuming positive coefficients $\beta^{(10)} > 0$, $\beta^{(02)} > 0$, and for $\Delta b < \Delta b^*$, the solution of equation (21) is

$$u_1(t) = u_1^{\mathrm{st}} - \sqrt{\frac{\beta^{(10)}(\Delta b^* - \Delta b)}{\beta^{(02)}}}\, \tanh\left[\sqrt{\beta^{(10)}\beta^{(02)}(\Delta b^* - \Delta b)}\, t + \mathcal{C}\right], \tag{22}$$

while for $\Delta b > \Delta b^*$ it is

$$u_1(t) = u_1^{\mathrm{st}} + \sqrt{\frac{\beta^{(10)}(\Delta b - \Delta b^*)}{\beta^{(02)}}}\, \tan\left[\sqrt{\beta^{(10)}\beta^{(02)}(\Delta b - \Delta b^*)}\, t + \mathcal{C}\right], \tag{23}$$

where $\mathcal{C}$ is determined from the initial condition $u_1(0)$. Note how, for $\Delta b < \Delta b^*$ in equation (22), the system goes to a stable fixed point in the infinite time limit as $\tanh(\infty) = 1$ and $u_1(\infty)$ takes a finite value, while for $\Delta b > \Delta b^*$ in equation (23) the system slows down close to $u_1(t) \approx u_1^{\mathrm{st}}$ and then diverges (corresponding to a meta-stable state). According to equation (23), close to the bifurcation point $\Delta b \gtrsim \Delta b^*$, the solution scales as $u_1(t) - u_1^{\mathrm{st}} \sim (\Delta b - \Delta b^*)^{1/2}$ and time as $t \sim (\Delta b - \Delta b^*)^{-1/2}$.

From the previous derivation we infer that, as we increase bias asymmetry Δb, we find a critical value Δb^* (see figures 2(c) and 4(c)) where two fixed points, a saddle and a stable point, merge and disappear at a saddle node bifurcation. At the critical point we observe a meta-stable (slow) region in both variables and time. The size of this region scales with the distance to the critical point as $(\Delta b - \Delta b^*)^{1/2}$ in the variables and as $(\Delta b - \Delta b^*)^{-1/2}$ in time. The zone is centered at the position where the two fixed points merge, i.e. the fixed point at the critical point $\Delta b = \Delta b^*$, and it is elongated along the slow eigendirection $\vec{v}_1$.

5.3. Convergence times

An important quantity to analyze the global dynamical behavior of the model is the convergence time needed to reach the final (stationary) state. Often, when modeling the opinion dynamics of a population, we are not

[5] For example, for the polarization transition in figure 3 we obtain $\beta^{(10)} = 0.226\,376$, $\beta^{(02)} = 0.672\,844$, $\beta^{(11)} = 0.789\,372$, and $\beta^{(03)} = -1.553\,56$.

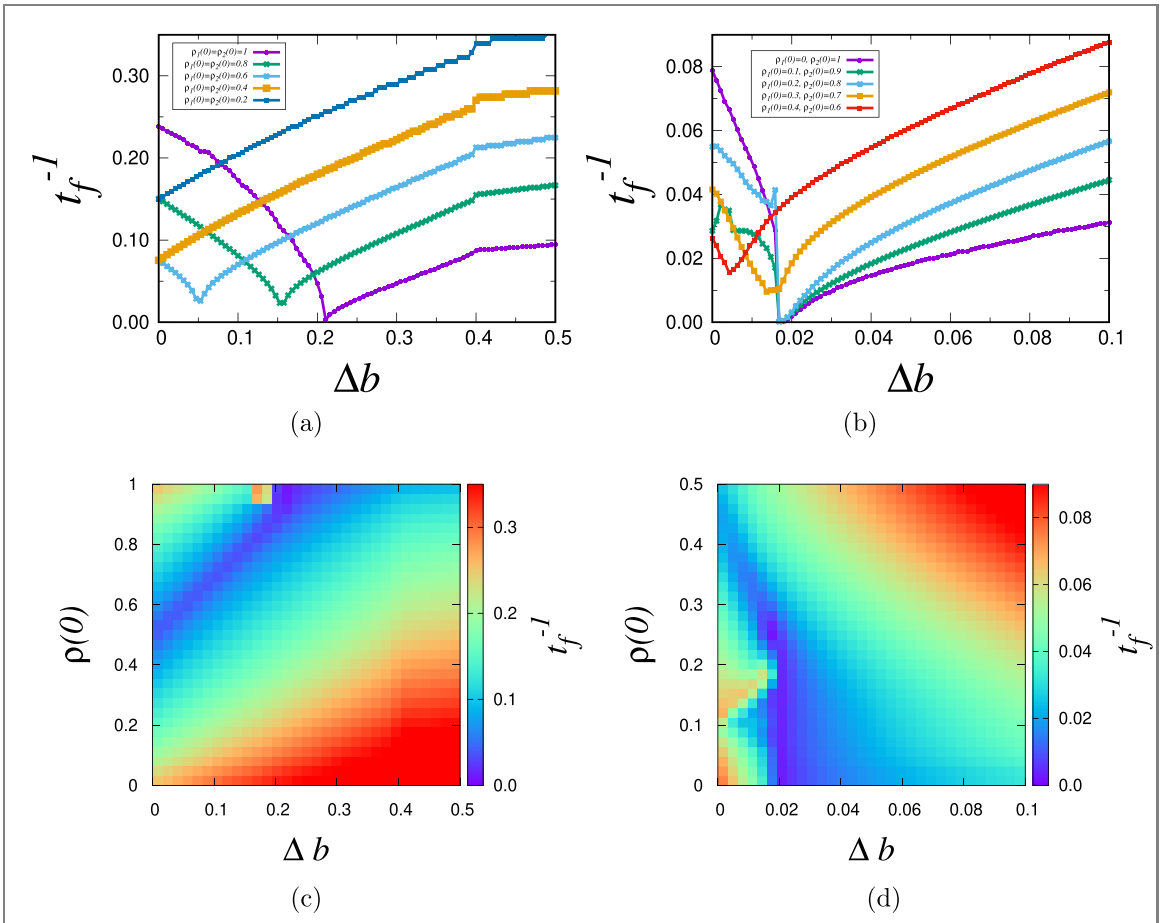

Figure 8. Inverse of convergence time (t_f^{-1}) as a function of bias asymmetry Δb and initial condition $\rho(0)$, coming from numerical simulations of the model. We use the same parameters of the model ($Q = 0.01$, $b = 0.8$, $\alpha = 1$ and $p = 0.1$), networks and simulation details as in figures 6 and 7. A percentage of %VAL = 10% was used in equations (26) and (27) to determine the time t_f. Top panels show the dependence $t_f^{-1}(\Delta b)$ for fixed initial conditions specified in the keys: (a) homogeneous $\rho_1(0) = \rho_2(0) = \rho(0)$, and (b) polarized $\rho_2(0) = 1 - \rho_1(0) = \rho(0)$. Bottom panels show the full dependence $t_f^{-1}(\Delta b, \rho(0))$ as a colormap, for (c) homogeneous and (d) polarized initial conditions.

only interested in the possible final states of the system, but also in this convergence time and its dependence on the parameters of the model.

For reason of convenience, we write the time dependence of the global quantities as follows:

$$\rho_1(t) = \rho_1^{\mathrm{st}} + (\rho_1(0) - \rho_1^{\mathrm{st}})g_1(t), \tag{24}$$

$$\rho_2(t) = \rho_2^{\mathrm{st}} + (\rho_2(0) - \rho_2^{\mathrm{st}})g_2(t), \tag{25}$$

such that $g_{1,2}(0) = 1$ and $g_{1,2}(\infty) = 0$. In the linear regime of equations (13) and (14), $g_{1,2}(t)$ are a sum of two exponential functions ($\mathrm{e}^{-\lambda_1 t}$ and $\mathrm{e}^{-\lambda_2 t}$) with amplitudes depending on the eigenvectors and initial conditions. The quantities λ_1^{-1} and λ_2^{-1} estimate well the two time scales involved in the time evolution close to the fixed point. For the general global dynamics $g_{1,2}(t)$, the exponential form is not valid and we have a complicated time-dependence. We measure the time scale t_f of the global dynamics as the smallest of the solutions $t_f^{(1)}$ and $t_f^{(2)}$ fulfilling the following conditions:

$$\frac{\rho_1(t_f^{(1)}) - \rho_1^{\mathrm{st}}}{\rho_1(0) - \rho_1^{\mathrm{st}}} = g_1(t_f^{(1)}) = \%\mathrm{VAL}, \tag{26}$$

$$\frac{\rho_2(t_f^{(2)}) - \rho_2^{\mathrm{st}}}{\rho_2(0) - \rho_2^{\mathrm{st}}} = g_2(t_f^{(2)}) = \%\mathrm{VAL}, \tag{27}$$

where %VAL is an arbitrary percentage measuring how close the system is to the final state when the convergence time t_f is reached.

In figure 8 we show the dependence of the convergence time t_f, extracted from numerical simulations, on bias asymmetry and initial conditions in the case of pair interactions, for the same parameters as in figures 6 and 7. Note that the system reaches different final states depending on the values of Δb, $\rho_1(0)$ and $\rho_2(0)$. This

can be clearly seen in figures 8(c) and (d), where a dark blue line separates two zones where the system reaches different final states, and where the behavior of the convergence times changes (in figures 8(a) and (b) it corresponds to the minimum of the curves). The line that separates the two behaviors is what we call a threshold initial condition $\rho_0(\Delta b)$ in section 5.1, i.e. the limit of the basin of attraction of the possible (consensus and polarization) final states. The dependence of the inverse of the convergence time (t_f^{-1}) with bias asymmetry Δb, increasing or decreasing, shows a clear correspondence with the eigenvalues of the final (stable) state (polarization or consensus) in figure 3. When the final state is polarization ($\Delta b < 0.02$ in figures 8(b) and (d)) it seems that the trend, increasing or decreasing, as a function of Δb is not that clear. This can be also understood from figure 3, where there is one eigenvalue increasing and another decreasing with Δb. Generally, the smallest eigenvalue dominates the dynamics, unless the initial condition is aligned with the fast eigendirection. That is the reason why we mainly observe a decreasing behavior in the simulations, while an increasing trend is also possible in some situations.

In figures 8(a) and (b) we can identify the characteristic scaling behavior $t_f^{-1} \sim |\Delta b - \Delta b^*|^{1/2}$, of the saddle-node bifurcation. Before the transition point ($\Delta b \lesssim \Delta b^*$), this is directly related to the eigenvalue of the final state $\lambda \sim (\Delta b^* - \Delta b)^{1/2}$ and can be considered as critical slowing down, while after the transition ($\Delta b \gtrsim \Delta b^*$), the meta-stable state dominates the dynamics (see figure 7(d)) with an equivalent time scaling as in equation (23). Note that there are two saddle-node bifurcations, figures 8(b) (consensus) and figure 8(a) (polarization), with different transition points Δb^*.

6. Summary and conclusions

In this paper we have studied the role of algorithmic bias and community structure in the potential rise of polarization of opinions in online social networks. We have devoted special attention to the temporal behavior of an archetypal two-state opinion-formation model, the language model, as well as to the role of the bias asymmetry Δb, i.e. the possibility that the online platform favors one opinion over the other. We have derived a pair of mean-field differential equations for the relevant variables of the dynamics, the density $\rho_1(t)[\rho_2(t)]$ of nodes in group 1 (2) holding opinion 1. This theoretical description accurately captures the phenomenology of the model and shows a good fit with numerical simulations.

The possible final opinion states reproduced by the model are: *consensus* ($\rho_1 = \rho_2 \approx 0$ or $\rho_1 = \rho_2 \approx 1$), *coexistence* ($\rho_1 = \rho_2 \approx 1/2$), and *polarization* ($\rho_1 \approx 0$, $\rho_2 \approx 1$ or $\rho_1 \approx 1$, $\rho_2 \approx 0$). All states are found in the whole spectrum between pair and group interactions displayed by the language model as the α parameter is changed. For some parameter values in the group interaction case, we also find additional polarized (polarization-coexistence) states with $\rho_1 \approx 0$ and $\rho_2 \approx 1/2$ (and the three equivalent states exchanging the groups $1 \leftrightarrow 2$ and states $0 \leftrightarrow 1$). Using linear stability analysis, we have determined the phase diagrams for pair and group interactions. In general, we find that sufficiently strong asymmetry in the bias is capable to destroy first the stability of the polarized states, and then one of the consensus states via saddle-node bifurcations. The phase diagram of the consensus states corresponds to the well-known cusp catastrophe for pair interactions and butterfly catastrophe for group interactions. Thus, bias asymmetry is a means to 'select' final states of the dynamics, controlling the global behavior of the system.

When the population is divided in two asymmetric groups in terms of size or connectivity (which can be thought of as a majority and minority groups), a somewhat different situation relating to the polarized states is produced by the model. In a range of values of the bias asymmetry $(-\Delta b, \Delta b)$, polarization is not necessarily suppressed but also favored, while above that range it is only destroyed. The two polarized states are not equivalent, depending on which is the opinion of the majority and minority groups. If the social platform benefits the opinion of the minority group then polarization is promoted, while in the opposite case polarization is suppressed in favor of consensus. The values of the bias asymmetry that delimit this behavior can be understood as the case where the structural asymmetry of the network (in group size and connectivity) is compensated by the 'favoritism' of algorithmic bias.

Results of numerical simulations confirm the possible final states predicted by the mean-field theory and the role of bias asymmetry. We have found that the convergence time (time to reach a stationary final state) depends on bias asymmetry, initial conditions, and the other parameters of the model in non-trivial ways. By means of the eigenvalues of the linearized dynamics, we were able to characterize this dependence close to the final states, gaining a better understanding of the type of transitions we have, and what happens to the dynamics when one of the final states loses its stability. An unavoidable phenomenon we observe when one of the stable solutions disappears is the presence of meta-stable states. This means that the system becomes trapped for a long period of time in a region of the phase space. Using the normal form of the bifurcations we have derived the scaling relations of the meta-stable zone as a function of the bias asymmetry. This shows a double square root law: $\rho_{1,2}(t) \sim (t_f)^{-1} \sim (\Delta b - \Delta b^*)^{1/2}$.

In conclusion, we have explored a simplified opinion formation model including some of the arguably most relevant features driving opinion dynamics on online social networks: spreading processes (pair or group copying together with independent behavior of individuals), algorithmic bias, and an underlying networked community structure. The joint effect of these ingredients produces a complex phase space of collective social behavior including coexistence, consensus, and polarization of opinions. We used this formalism as testing ground to study the influence of algorithmic bias on online communication dynamics. We showed that bias imbalance can have a crucial effect on the final opinion state and the dynamics in general. Polarization and consensus states are destroyed for high enough bias asymmetry at different transitions, while just after the transition these destroyed final states become meta-stable. We characterized all possible transitions via phase diagrams. For the local dynamics (close to the final states) we used a linearization of the dynamical equations, while for the global dynamics the normal form of the bifurcation allowed us to detect meta-stable states. Finally, we calculated convergence times both from the theoretical description and by means of numerical simulations.

The aim of such simplified modelling cannot be the quantitative reproduction of some empirical observations, e.g., related to elections. Still, we think that the richness of the behavior of the model, the relatively large number of relevant parameters, and the non-triviality of the temporal evolution of opinions are all features with strong relation to real-world systems. For example, the effect of the parameters on polarization is sometimes counter-intuitive: how algorithmic bias influences the outcome of the dynamics depends strongly on the type of interaction. In our model we can tune interactions from pairwise to group—in reality both are present and highly heterogeneous. Another relevant observation is the frequent appearance of meta-stable states. As real-world phenomena evolve over finite times (e.g., opinions matter on election day), meta-stable states may be crucial in forecasting efforts. In the future we would like to combine modeling with empirical data analysis, partly from observational data and partly from controlled social experiments, in order to further understand the interplay between human collective action and online algorithms.

Acknowledgments

We thank Matteo Neri for his help in the initial stages of this project. We acknowledge support from AFOSR (Grant No. FA8655-20-1-7020). JK is grateful for support by project ERC DYNASNET Synergy (Grant No. 810115). JK and GI acknowledge support from projects EU H2020 Humane AI-net (Grant No. 952026) and SAI enabled by FWF within the EU CHIST-ERA project SAI (Grant No. FWF I 5205-N).

Data availability statement

No new data were created or analysed in this study.

References

[1] Castellano C, Fortunato S and Loreto V 2009 Statistical physics of social dynamics *Rev. Mod. Phys.* **81** 591
[2] Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 Epidemic processes in complex networks *Rev. Mod. Phys.* **87** 925
[3] Porter M A and Gleeson J P 2016 Dynamical systems on networks *Front. Appl. Dyn. Syst.* **4** 49–51
[4] Newman M E J 2003 The structure and function of complex networks *SIAM Rev.* **45** 167
[5] Lambiotte R and Schaub M 2021 *Modularity and Dynamics on Complex Networks* (Cambridge: Cambridge University Press)
[6] Lazer D *et al* 2009 Computational social science *Science* **323** 721
[7] Conte R *et al* 2012 Manifesto of computational social science *Eur. Phys. J. Spec. Top.* **214** 325
[8] Nikolov D, Lalmas M, Flammini A and Menczer F 2018 Quantifying biases in online information exposure *J. Assoc. Inf. Sci. Technol.* **70** 218
[9] Bozdag E 2013 Bias in algorithmic filtering and personalization *Ethics Inf. Technol.* **15** 209
[10] Möller J, Trilling D, Helberger N and van Es B 2018 Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity *Inf. Commun. Soc.* **21** 959
[11] Pariser E 2011 *The Filter Bubble: What the Internet Is Hiding from You* (London: Penguin Books Limited)
[12] Bakshy E, Messing S and Adamic L A 2015 Exposure to ideologically diverse news and opinion on facebook *Science* **348** 1130
[13] Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G and Quattrociocchi W 2016 Echo chambers: emotional contagion and group polarization on facebook *Sci. Rep.* **6** 37825
[14] Bail C A *et al* 2018 Exposure to opposing views on social media can increase political polarization *Proc. Natl Acad. Sci. USA* **115** 9216
[15] Ciampaglia G L, Nematzadeh A, Menczer F and Flammini A 2018 How algorithmic popularity bias hinders or promotes quality *Sci. Rep.* **8** 15951
[16] Blex C and Yasseri T 2020 Positive algorithmic bias cannot stop fragmentation in homophilic networks *J. Math. Sociol.* 1–18
[17] Baumann F, Lorenz-Spreen P, Sokolov I M and Starnini M 2020 Modeling echo chambers and polarization dynamics in social networks *Phys. Rev. Lett.* **124** 048301

[18] Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W and Starnini M 2021 The echo chamber effect on social media *Proc. Natl Acad. Sci. USA* **118** e2023301118
[19] Sîrbu A, Pedreschi D, Giannotti F and Kertész J 2019 Algorithmic bias amplifies opinion fragmentation and polarization: a bounded confidence model *PLoS One* **14** e0213246
[20] Deffuant G, Neau D, Amblard F and Weisbuch G 2000 Mixing beliefs among interacting agents *Advs. Complex Syst.* **03** 87
[21] Gleeson J P 2011 High-accuracy approximation of binary-state dynamics on networks *Phys. Rev. Lett.* **107** 068701
[22] Gleeson J P 2013 Binary-state dynamics on complex networks: pair approximation and beyond *Phys. Rev.* X **3** 021004
[23] Perra N and Rocha L E C 2019 Modelling opinion dynamics in the age of algorithmic personalisation *Sci. Rep.* **9** 7261
[24] De Marzo G, Zaccaria A and Castellano C 2020 Emergence of polarization in a voter model with personalized information *Phys. Rev. Res.* **2** 043117
[25] Stark H-U, Tessone C J and Schweitzer F 2008 Decelerating microdynamics can accelerate macrodynamics in the voter model *Phys. Rev. Lett.* **101** 018701
[26] Peralta A F, Khalil N and Toral R 2019 Ordering dynamics in the voter model with aging *Physica* A **552** 122475
[27] Peralta A F, Khalil N and Toral R 2020 Reduction from non-Markovian to Markovian dynamics: the case of aging in the noisy-voter model *J. Stat. Mech.* 024004
[28] Peralta A F, Neri M, Kertész J and Iñiguez G 2021 Effect of algorithmic bias and network structure on coexistence, consensus, and polarization of opinions *Phys. Rev.* E **104** 044312
[29] Abrams D M and Strogatz S H 2003 Modelling the dynamics of language death *Nature* **424** 900
[30] Vazquez F, Castelló X and Miguel M S 2010 Agent based models of language competition: macroscopic descriptions and order-disorder transitions *J. Stat. Mech.* P04007
[31] Ants A K 1993 Rationality, and recruitment *Q. J. Econ.* **108** 137
[32] Granovsky B L and Madras N 1995 The noisy voter model *Stoch. Process. Appl.* **55** 23
[33] Peralta A F, Carro A, Miguel M S and Toral R 2018 Analytical and numerical study of the non-linear noisy voter model on complex networks *Chaos* **28** 075516
[34] Xiong F and Liu Y 2014 Opinion formation on social media: an empirical approach *Chaos* **24** 013130
[35] Castellano C, Muñoz M A and Pastor-Satorras R 2009 Nonlinear q-voter model *Phys. Rev.* E **80** 041129
[36] Nyczka P, Sznajd-Weron K and Cisło J 2012 Phase transitions in the q-voter model with two types of stochastic driving *Phys. Rev.* E **86** 011105
[37] Jędrzejewski A 2017 Pair approximation for the q-voter model with independence on complex networks *Phys. Rev.* E **95** 012307
[38] Catanzaro M, Boguña M and Pastor-Satorras R 2005 Generation of uncorrelated random scale-free networks *Phys. Rev.* E **71** 027103
[39] Yang Z, Algesheimer R and Tessone C J 2016 A comparative analysis of community detection algorithms on artificial networks *Sci. Rep.* **6** 30750
[40] McPherson M, Smith-Lovin L and Cook J M 2001 Birds of a feather: homophily in social networks *Annu. Rev. Sociol.* **27** 415
[41] Asikainen A, Iñiguez G, Ureña-Carrión J, Kaski K and Kivelä M 2020 Cumulative effects of triadic closure and homophily in social networks *Sci. Adv.* **6** eaax7310
[42] Peixoto T P 2021 Disentangling homophily, community structure and triadic closure in networks (arXiv:2101.02510)
[43] Stewart I 1982 Catastrophe theory in physics *Rep. Prog. Phys.* **45** 185
[44] Guckenheimer J and Holmes P 2002 *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (*Applied Mathematical Sciences*) (New York: Springer)
[45] Peralta A F and Toral R 2020 Binary-state dynamics on complex networks: stochastic pair approximation and beyond *Phys. Rev. Res.* **2** 043370