Effect of collaborative-filtering-based recommendation algorithms on opinion polarization

Alessandro Bellina,[1,2] Claudio Castellano,[3,2] Paul Pineau,[4] Giulio Iannelli,[2,5] and Giordano De Marzo[1,2,6,7]

[1]*Dipartimento di Fisica Università "Sapienza," P. le A. Moro, 2, I-00185 Rome, Italy*
[2]*Centro Ricerche Enrico Fermi, Piazza del Viminale, 1, I-00184 Rome, Italy*
[3]*Istituto dei Sistemi Complessi (ISC-CNR), Via dei Taurini 19, I-00185 Rome, Italy*
[4]*École Normale Supérieure Paris-Saclay, 4 Avenue des Sciences, 91190 Gif-sur-Yvette, France*
[5]*Dipartimento di Fisica, Università di Roma "Tor Vergata," 00133 Rome, Italy*
[6]*Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080, Vienna, Austria*
[7]*Sapienza School for Advanced Studies, "Sapienza," P. le A. Moro, 2, I-00185 Rome, Italy*

(Received 3 April 2023; accepted 11 September 2023; published 6 November 2023)

A central role in shaping the experience of users online is played by recommendation algorithms. On the one hand they help retrieving content that best suits users taste, but on the other hand they may give rise to the so-called "filter bubble" effect, favoring the rise of polarization. In the present paper we study how a user-user collaborative-filtering algorithm affects the behavior of a group of agents repeatedly exposed to it. By means of analytical and numerical techniques we show how the system stationary state depends on the strength of the similarity and popularity biases, quantifying respectively the weight given to the most similar users and to the best rated items. In particular, we derive a phase diagram of the model, where we observe three distinct phases: disorder, consensus, and polarization. In the last users spontaneously split into different groups, each focused on a single item. We identify, at the boundary between the disorder and polarization phases, a region where recommendations are nontrivially personalized without leading to filter bubbles. Finally, we show that our model well reproduces the behavior of users on the online music platform last.fm. This analysis paves the way to a systematic analysis of recommendation algorithms by means of statistical physics methods and opens the possibility of devising less polarizing recommendation algorithms.

DOI: 10.1103/PhysRevE.108.054304

I. INTRODUCTION

The growth of polarization and radicalization observed in recent years [1–3] is a phenomenon that can potentially undermine the functioning and stability of democratic societies. In this context, the critical role played by online platforms has been widely recognized [4–6], but the detailed mechanisms by which exposure to online content drives polarization at the population level are still to be fully clarified. Recommendation algorithms, together with more traditional media such as television [7], are believed to be among the key factors, since they strongly influence users' online experience by selecting, based on past behavior, the new information users are exposed to [8–10]. Such algorithms are fundamental for filtering and selecting the content we are interested to, a sorely needed task, given the overwhelming amount of information available online. On the other side, however, recommendation algorithms produce a feedback loop that naturally tends to bias future choices, reducing the diversity of available content and thus favoring the so-called "filter-bubble" effect and the consequent polarization of opinions [11–15]. Filter bubbles, occurring when users are mainly exposed to news and content aligned to their beliefs, are similar to the much investigated "echo chambers" [16–18]. While the latter result from homophylic interactions among users, which tend to interact with people sharing the same opinions, the former are produced by algorithmically biased recommendations on online platforms.

Recommendation algorithms are widely used by most of the web sites we visit everyday, examples being "suggested for you" posts on Facebook, recommended items on the Amazon online shop, or Google personalized PageRank. Such algorithms are designed to allow easy access to content we are expected to be interested in so as to maximize our engagement with the platform. Collaborative filtering [19,20] is a paradigmatic approach to algorithmic recommendation which, despite its simplicity, is employed by online giants such as Amazon [9,10]. The underlying principle is that past behavior of users can be exploited to determine the similarity between them or between items, which can then be used to identify new content users will most likely appreciate. The first case corresponds to "user-user" collaborative filtering, while the latter to the "item-item" one. In the following we will use interchangeably the terms "item" or "opinion" as equivalent ways to refer to a generic piece of content available on the platform.

In past years much attention has been devoted to the study of how microscopic interactions among users shape collective phenomena at the population level [21,22] and in particular to the investigation of how the polarization of opinions emerges from such interactions [23–27]. The effect of recommendation algorithms has received less attention, and only recently have scholars started to model their interplay with the dynamics of opinions. For instance, Refs. [28,29] approached the problem by endowing a voter model with an external field which represents users' interaction with their past history, thus mimicking content recommendation. A similar methodology has been proposed in Ref. [30], where the effect of recommendations based on agents' present state is considered, while in other works [31,32] the effect of the recommendation algorithm is modeled by filtering the interactions between an individual and its neighbors, depending on their state. In all cases the conclusion is that recommendation algorithms may play a crucial role in enhancing opinion polarization and fragmentation. Also the effect of link recommendations (algorithms suggesting new social connections) has been analyzed, revealing that personalized suggestions of new friends can increase polarization [33–35] and favor inequality and biases [36,37].

All the studies concerning content recommendations, despite providing useful insight on the possible effects of their implementation, consider exceedingly simple algorithms, coupled to highly stylized opinion dynamics models. As a consequence they do not shed much light on the effect of realistic recommendation algorithms adopted by online platforms. In order to fill the gap between theoretical modeling and real implementations, in this paper we present a systematic study of a model for user-user collaborative filtering. We find that, depending on two parameters (the strength of the similarity bias α and of the popularity bias β; see below for definitions), the system can be in three different phases: disorder, consensus, and polarization. In particular, when the two biases are sufficiently large, the system undergoes a spontaneous breaking of user and item symmetry, leading to the formation of polarized groups and giving rise to the filter bubble effect. Such a drawback can be avoided at the boundary between disorder and polarization, where the algorithm provides meaningful recommendations without inducing opinion polarization. Finally, we use our model to reproduce the behavior of users in the online music platform last.fm, determining, within our modeling framework, the strength of the similarity and rating biases they are subject to.

These results show that a statistical physics approach to recommendation algorithms is crucial in understanding their effect on opinion polarization, while also being a powerful tool for determining the best parameters to be used in their implementation.

II. DEFINITION OF THE MODEL

Let us consider a system composed of N users which iteratively choose (click) among M items or opinions. We denote by U (with $|U| = N$) the set of all users, while I (with $|I| = M$) is the set of items. At time t, each user u is described by a M-dimensional vector $\mathbf{r}_u(t) = \{r_{u1}(t), \dots, r_{uM}(t)\}$ whose components $r_{ui}(t)$ are given by the number of times user u has clicked on item i so far. In the following we refer to the r_{ui} as ratings and we assume that clicking on an item expresses (positive) interest in it. Initially all ratings are set equal to r_0, i.e., $r_{ui}(0) = r_0$ for all users u and items i. These initial conditions reflect the absence of any *a priori* knowledge about users' taste and mimic the so called "cold start" of recommendation algorithms [38]. Note that in real systems the number of items available to users is typically enormous. For instance there are almost 100 million tracks on Spotify and around 350 million products are available on Amazon online store. As a consequence in the following we will be interested in taking the large M limit. In order to do so, as explained in Appendix A, we have to set $r_0 \sim M^{-1}$; in the following we take $r_0 = 1/(M-1)$ unless specified otherwise.

At each time step, of duration $\delta t = 1/N$, a user u is selected at random and he or she clicks on item i with probability $R_{ui}(t)$,

$$\begin{aligned} R_{ui}(t) &= P(r_{ui}(t+\delta t) = r_{ui}(t) + 1) \\ &= \text{Prob } (u \text{ is selected at time } t \text{ and clicks on } i). \end{aligned}$$

The specific form of this normalized probability ($\sum_{u,i} R_{ui}(t) = 1$) encodes how the recommendation algorithm affects the user behavior. The idea behind the collaborative-filtering mechanism is that such a probability should be the larger the more item i is positively rated by users similar to u. More precisely, we quantify the similarity s_{uv} between users u and v as the cosine similarity of their rating vectors,

$$s_{uv} = \frac{\mathbf{r}_u \cdot \mathbf{r}_v}{\|\mathbf{r}_u\| \|\mathbf{r}_v\|} = \frac{\sum_i r_{ui} r_{vi}}{\sqrt{\sum_i r_{ui}^2}\sqrt{\sum_i r_{vi}^2}}.$$

In these terms we define the transition probability $R_{ui}(t)$ as

$$R_{ui}(t) = \frac{1}{N} \sum_{v=1}^{N} \frac{s_{uv}^{\alpha}(t)}{\sum_w s_{uw}^{\alpha}(t)} \frac{r_{vi}^{\beta}(t)}{\sum_j r_{vj}^{\beta}(t)}. \tag{1}$$

The two parameters α and β quantify the strength, respectively, of the similarity bias and of the popularity bias. Indeed, the larger α, the more users are biased toward items liked by the agents they are more similar to, while the larger β the more users are biased toward items already selected in the past. Equation (1) is a direct generalization of the standard user-user collaborative filtering expression [9,39]. The latter corresponds to $\alpha = \beta = 1$ and reads

$$\text{Score}_{ui} = \sum_{v}^{N} s_{uv} \frac{r_{vi}}{\sum_j r_{vj}}.$$

The clicking probability of a user is then obtained normalizing the scores by their sum. Note that with the present definition the probability for user u is affected by the "self-interaction" with his or her past, as the sum includes the term weighted by $s_{uu} = 1$. Because of self-interaction, when $\alpha \to \infty$ each agent interacts only with him- or herself and users are completely independent. The framework we are considering corresponds to a collaborative filtering with implicit feedback, meaning that the appreciation users give to items is not directly available, but rather it is derived by the number of times users click on items. This situation occurs, for instance, in music streaming platforms, where r_{ui} corresponds to the number of times user u listened to song (or artist) i.

III. BEHAVIOR FOR LIMIT VALUES OF THE PARAMETERS

By inspecting Eq. (1) it is easy to anticipate that, depending on the strength of the similarity and popularity biases, the system can show very different behaviors. In particular, three distinct phases can be identified by considering simple limits:

(1) $\beta = 0, \forall\alpha$: *disorder*

Without a popularity bias, Eq. (1) reduces to

$$R_{ui} = \frac{1}{N}\sum_v \frac{s_{uv}^\alpha}{\sum_w s_{uw}^\alpha}\frac{1}{M} = \frac{1}{NM},$$

and thus all users behave as random clickers. In this case all the items share the same probability of being clicked and the system is *disordered*, meaning that any user rates equally (on average) any item.

(2) $\alpha = 0$ and $\beta = \infty$: *consensus*

When the similarity bias is set to $\alpha = 0$ the transition probability simply is

$$R_{ui} = \frac{1}{N^2}\sum_v \frac{r_{vi}^\beta}{\sum_j r_{vj}^\beta}$$

and is independent of the user u. Moreover, since the popularity bias is maximal, it holds $r_{vi}^\beta/\sum_j r_{vj}^\beta = \delta_{i,i_v}$, where i_v is the most rated item by user v. As a consequence, denoting as $N_i = \sum_v \delta_{i,i_v}$ the number of users having i as the most rated item, we get

$$R_{ui} = \frac{1}{N^2}\sum_v \delta_{i,i_v} = \frac{N_i}{N^2}.$$

This expression implies that users are more likely to click on the globally most popular item, thus originating a feedback loop which for large time is expected to make such an item become the most rated for each user. This means that the system evolves toward a *consensus* phase, where all users agree on the same opinion. In this consensus phase the recommendation algorithm always suggests the same item to all users.

(3) $\alpha = \infty$ and $\beta = \infty$: *polarization*

In this case both the rating and the similarity biases are maximal and Eq. (1) becomes

$$R_{ui} = \frac{1}{N}\delta_{i,i_u},$$

where i_u is the opinion more frequently clicked on by user u. As a consequence users persistently stick to their first random choice (note that at $t = 0$ all ratings are the same, and so all items have the same probability to be chosen), giving rise to *polarization* and to the filter bubble effect. Indeed, in the polarized phase the recommendation algorithm suggests to each user just one specific item, but such an item varies from user to user.

In order to investigate the model, it is useful to introduce the normalized ratings $\hat{r}_{ui}$, defined as

$$\hat{r}_{ui}(t) = \frac{r_{ui}(t)}{\sum_i r_{ui}(t)} \approx \frac{r_{ui}(t)}{t + Mr_0}, \tag{2}$$

where the approximation comes from the assumption that in a time interval $\Delta t = N\delta t = 1$ each user is updated once on average. The normalized ratings satisfy, for asymptotically large times, the Martingale property, i.e., $\mathbb{E}[\hat{r}_{ui}(t + \delta t)] = \mathbb{E}[\hat{r}_{ui}(t)]$. Since they are also limited in (0,1), this ensures that these random variables converge to an asymptotic limit for large times. Thus they are the right variables to look at in order to find asymptotic stationary solutions of the system. By looking at the evolution of the normalized ratings $\hat{r}_{ui}$ for various values of the biases we can observe the different phases identified by inspecting the behavior of the model in the limit cases discussed above. First, we focus, without lack of generality, on a specific user, $u = 1$, and we study the behavior of the ratings $\hat{r}_{1i}(t)$ for such a user. Figure 1(a) shows the temporal evolution of these quantities for $M = 25$ items and $\alpha = \beta = 0.5$: all normalized ratings converge to the value $1/M$, meaning that the user under consideration equally rates all possible opinions. This corresponds to a disordered configuration in which the algorithm provides random recommendations and the users behave as random clickers. The situation radically changes when the popularity bias is increased to $\beta = 5$, as shown in Fig. 1(b). In this case one of the normalized ratings converges to one, while all the others go to zero, meaning that the user ends up always choosing the same opinion. Hence the symmetry among items, found for small values of β, breaks down when the popularity bias is increased. Analogously, we can focus on a given item (we choose the first without lack of generality) and study the rating that different users give to it. Figure 1(c) shows the evolution of the N normalized ratings $\hat{r}_{u1}(t)$, corresponding to the various items, for $\alpha = 0.5$ and $\beta = 5$. We observe consensus among users, since for all of them the normalized rating of item 1 converges to one, while all other normalized ratings (not shown) go to zero. Finally, we report in Fig. 1(d) the behavior of $\hat{r}_{u1}(t)$ when also the similarity bias α is set to 5. The symmetry among users breaks down. While for small α they all act the same, here their behavior is heterogeneous: some of them maximally rate the first item, while for others the normalized rating corresponding to such an opinion vanishes, thus giving rise to a polarized configuration in which users are divided into groups depending on the opinion they support.

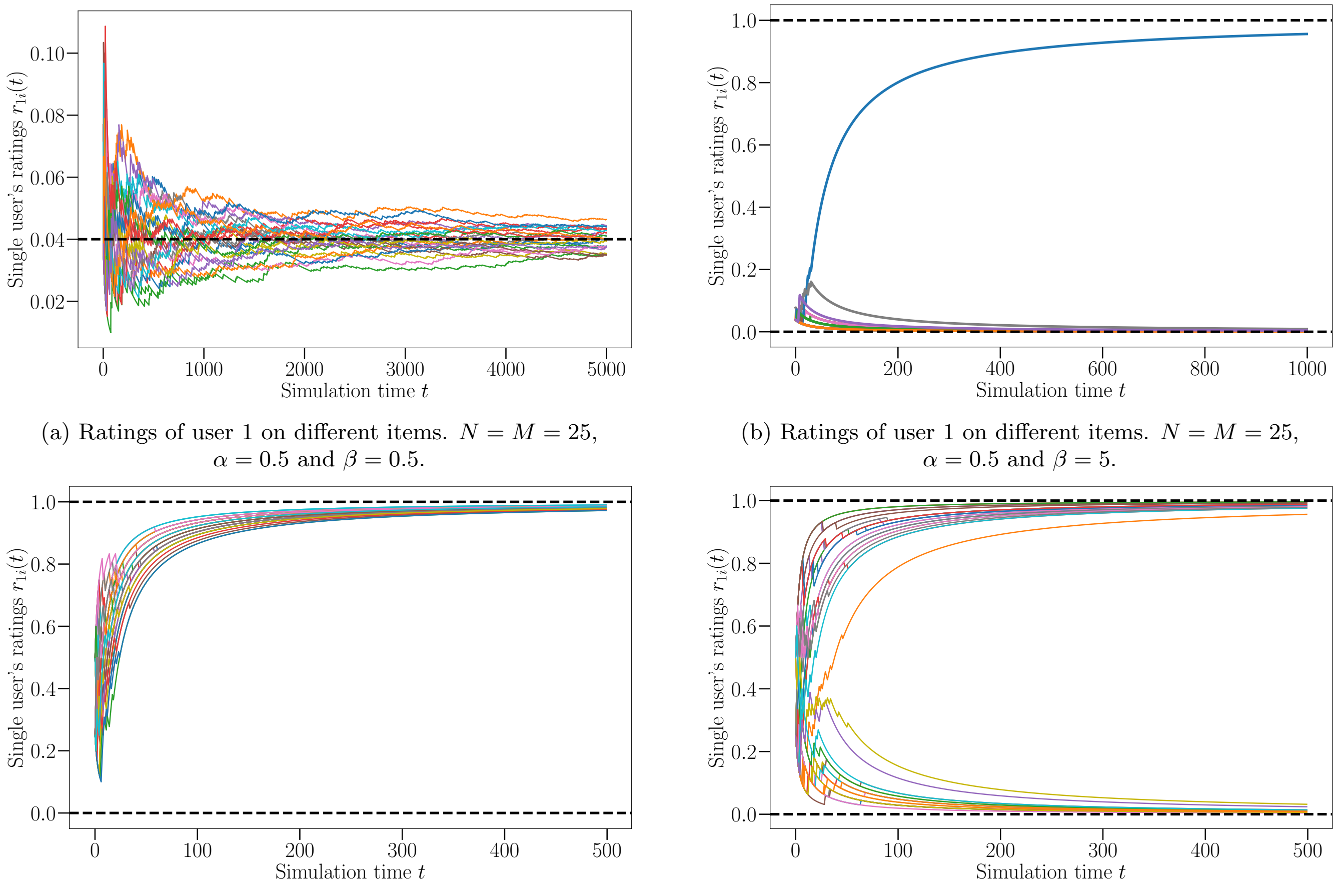
IV. PHASE DIAGRAM AND ASSOCIATED TRANSITIONS

When the control parameters α and β are varied between the limit values discussed above, phase transitions take place, associated to distinct symmetry breakings. For small values of the biases both users and items are completely symmetric and the system is in a disordered phase where all users rate all items in the same way. The popularity bias β is responsible for the item-symmetry breaking: when this parameter gets sufficiently large each user only clicks on a specific item. Analogously, an increase of the similarity bias α breaks the user symmetry, leading to heterogeneous user behavior. The polarized phase observed for large values of the biases emerges when both symmetries are broken, while consensus occurs when only the item symmetry is broken. Note that no phase where only the user symmetry is broken is possible: if all items are equally rated by each user, necessarily all users are similar. In this section we show how these considerations can be made, by analytical and numerical means, more grounded and precise.

A. Master equation

The master equation for the probability distribution of r_{ui}, $Q(r_{ui}, t)$ is

$$\frac{d}{dt}Q(r_{ui}, t) = \frac{1}{\delta t}\begin{bmatrix} Q(r_{ui} - 1, t)R_{ui}(r_{ui} - 1) \\ -R_{ui}(r_{ui})Q(r_{ui}, t)\end{bmatrix} \tag{3}$$



(a) Ratings of user 1 on different items. $N = M = 25$, $\alpha = 0.5$ and $\beta = 0.5$.

(b) Ratings of user 1 on different items. $N = M = 25$, $\alpha = 0.5$ and $\beta = 5$.

(c) Ratings for item 1 of different users. $N = M = 40$, $\alpha = 0.5$ and $\beta = 5$.

(d) Ratings for item 1 of different users. $N = M = 40$, $\alpha = 5$ and $\beta = 5$.

FIG. 1. Temporal evolution of the normalized ratings for different values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. (a) All M normalized ratings $\hat{r}_{1i}(t)$ for user $u = 1$ are shown. The configuration is disordered; all items are clicked on indifferently, and as a consequence all ratings fluctuate around the mean value $1/M$ (users are random clickers). (b) Again all M normalized ratings $\hat{r}_{1i}(t)$ for user $u = 1$ are shown. Since we are above the transition in β, the user tends to a single-item configuration, i.e., asymptotically only one item is clicked on; its normalized rating converges to 1 while the others go to 0. The comparison between panels (a) and (b) reveals the features of the *multiple-item to single-item* transition. (c) Ratings $\hat{r}_{u1}(t)$ of all users for the item $i = 1$ are shown. This is a consensus configuration; all users, which asymptotically tend to a single-item state, click on the same item $i = 1$. (d) Again the ratings $\hat{r}_{u1}(t)$ of all N users for item $i = 1$ are shown. Since $\alpha > \alpha_c$ users are polarized; some of them tend to click only on item $i = 1$, but others tend to click only on a different item i'. The comparison between panels (c) and (d) reveals the features of the *consensus-polarization* transition.

from which the drift coefficient is readily obtained,

$$\nu_{ui} = \frac{d\langle r_{ui}\rangle}{dt} = \frac{1}{\delta t}\langle R_{ui}\rangle = N\langle R_{ui}\rangle. \tag{4}$$

It follows the expression for the drift of the normalized ratings,

$$\hat{\nu}_{ui} = \frac{d\langle \hat{r}_{ui}\rangle}{dt} \approx \frac{1}{t + Mr_0}(N\langle R_{ui}\rangle - \langle \hat{r}_{ui}\rangle). \tag{5}$$

Detailed computations of Eqs. (3)–(5) are reported in Appendix B. Focusing on the long-time behavior of the system, diffusion can be neglected and the evolution of the normalized ratings can be approximated by means of the drift terms only,

$$\frac{d\hat{r}_{ui}}{dt} \approx \frac{1}{t + Mr_0}(NR_{ui} - \hat{r}_{ui}). \tag{6}$$

The stationary solutions of the dynamics are those for which the time derivative reported above is equal to zero, that is, $NR_{ui} - \hat{r}_{ui} = 0$. It is easy to show (see Appendix C) that disorder, consensus, and polarization are solutions of this equation. In particular, these solutions are defined, in terms of normalized ratings, as follows:

(1) *Disorder*: All ratings are equal, $\hat{r}_{ui} = 1/M, \forall u, \forall i$.

(2) *Consensus*: The ratings for one item are 1 for all users ($r_{ui} = 1, \forall u$), while the ratings for all other items are 0 for all users ($r_{uj} = 0, \forall u, \forall j \neq i$).

(3) *Polarization*: There are at least two groups of users, labeled by k and k', with consensus within each group ($r_{ui_k} = 1$, $r_{uj} = 0, \forall j \neq i_k, \forall u \in k$), but the selected item is different for different groups ($i_k \neq i_{k'}$). It is important to remark that in the polarized phase, the interaction induced by the recommendation algorithm is such that users sharing a different opinion do not influence each other.

As shown in Appendix C, also other configurations, where $0 < m < M$ items are equally rated, have a null drift. In order to understand which are the true stationary states of the system it is necessary to consider their stability.

B. Stability analysis

1. The transition from multiple item to single item

Let us consider the case $\alpha = \infty$. From the phenomenological considerations presented above we expect the disordered solution occurring for $\beta = 0$ to be stable also for small β, while users should stick to a single item for larger values of the popularity bias. In the disordered phase all items are equally likely to be clicked on. A disordered (multiple-item) configuration is then described by all normalized ratings equal to $1/M$ with small fluctuations ϵ_{ui}:

$$\forall (u, i), \quad \hat{r}_{ui} = \frac{1}{M} + \epsilon_{ui}, \quad |\epsilon_{ui}| \ll \frac{1}{M}. \tag{7}$$

The fluctuations can be either positive or negative, and they are constrained by the normalization of ratings,

$$\sum_i \hat{r}_{uj} = \sum_i \left(\frac{1}{M} + \epsilon_{uj} \right) = 1 \Rightarrow \sum_j \epsilon_{uj} = 0. \tag{8}$$

By plugging Eq. (7) into Eq. (6), recalling that $r_{ui}/\sum r_{ui} = \hat{r}_{ui}/\sum \hat{r}_{ui}$ and expanding for small ϵ_{ui} it is easy to show (see Appendix D) that

$$\frac{d\epsilon_{ui}}{dt} \propto (\beta - 1)\epsilon_{ui}. \tag{9}$$

If $\beta < 1$, fluctuations ϵ_{ui} are exponentially suppressed. The system is then always driven back to the disordered solution, which is therefore stable. When $\beta > 1$ fluctuations are instead amplified and the multiple-item solution is unstable. By a similar argument (see Appendix D) it is possible to show that the single-item solution, for which in the large time limit $\hat{r}_{ui} = 1$ and $\hat{r}_{uj} = 0 \ \forall \ j \neq i$ is stable when $\beta > 1$, while it is unstable if $\beta < 1$. Moreover, solutions characterized by users equally rating more than 1 but less than M items are found to be unstable both for $\beta < 1$ and $\beta > 1$, showing the nonexistence of a fourth stable phase. We thus conclude that at $\beta_c = 1$ a transition between the multiple-item and the single-item solution occurs, associated with the breaking of symmetry among items. In Appendix D we show that the same picture applies in the general case $\alpha < \infty$.

2. The transition from consensus to polarization

For the transition in β, we can study the transition in α by looking at the stability of the consensus and polarization solutions. Since we already know that for $\beta < 1$ the single-item solution is never observed, we can assume $\beta > 1$; for simplicity here we set $\beta = \infty$, while we refer to Appendix E for the general case. We assume users close to the single-item solution,

$$\begin{aligned} \hat{r}_{ui} &= 1 - \epsilon \\ \hat{r}_{uj} &= \frac{\epsilon}{M-1} \qquad \forall j \neq i, \end{aligned} \tag{10}$$

with $0 < \epsilon \ll 1$. Note that here we made the assumption that ϵ does not depend on u and i, but the same results can be obtained also considering the more general case of fluctuations of the form ϵ_{ui}.

In the consensus configuration all users are aligned along the same item i, and so all similarities are approximately equal to 1. As a consequence, since $\beta = \infty$, we can write Eq. (6) as

$$\begin{aligned} \frac{d\hat{r}_{ui}}{dt} &\approx \frac{1}{t + Mr_0} \left(\frac{\sum_v s_{uv}^{\alpha} \delta_{i,i_v}}{\sum_w s_{uw}^{\alpha}} - \hat{r}_{ui} \right) \\ &= \frac{1}{t + Mr_0} (1 - (1 - \epsilon)) = \frac{\epsilon}{t + Mr_0} > 0, \end{aligned}$$

where we used the fact that all similarities are equal to 1 and $i_v = i$ for all users. Since this quantity is always positive regardless of α, the normalized rating along i keeps increasing, asymptotically converging to the consensus solution $\hat{r}_{ui} = 1$. Considering instead an item $j \neq i$, we get for the drift

$$\frac{d\hat{r}_{uj}}{dt} = \frac{1}{t + Mr_0} \left(0 - \frac{\epsilon}{M-1} \right) < 0,$$

meaning that the normalized ratings of the other items consistently go to zero. These results imply that the consensus solution is an attractor of the dynamics for any value of the similarity bias α.

We can then turn to the study of the polarized state. In this case the similarity between two users polarized on the same item is again 1, while for users polarized on different items it holds

$$s_{uv} = \frac{2(M-1)\epsilon - M\epsilon^2}{1 - 2\epsilon + M\epsilon^2} = O(\epsilon) \quad \text{if } i_u \neq i_v,$$

where i_u is the item user u is polarized on. Assuming for simplicity $K < N/2$ distinct users polarized on item i_u, and the remaining $N - K$ polarized along i_v, Eq. (6) reads

$$\frac{d\hat{r}_{ui_u}}{dt} \approx \frac{1}{t + Mr_0} \left[\frac{K}{K + (N-K)\epsilon^{\alpha}} - (1 - \epsilon) \right].$$

This quantity is negative for $\alpha < 1$, while it is positive for $\alpha > 1$. Analogously, the drift of the rating of user u along item i_v is

$$\frac{d\hat{r}_{ui_v}}{dt} \approx \frac{1}{t + Mr_0} \left[\frac{K\epsilon^{\alpha}}{(N-K) + K\epsilon^{\alpha}} - \frac{\epsilon}{M-1} \right],$$

which is positive for $\alpha < 1$ and negative for $\alpha > 1$. These results imply that for $\alpha < \alpha_c = 1$ the K users polarized along i_u will not remain polarized as the system evolves, finally polarizing along the item i_v shared by the majority of agents. Conversely, for $\alpha > \alpha_c$ the drift reinforces the minority and a polarized state consisting of two different group of users emerges. The analysis thus indicates that for $\alpha < 1$ only the consensus state is stable, while we expect to observe both consensus and polarization above the critical value.

The overall phase diagram of the model is summarized in Fig. 2.

3. The simplest case: N = M = 2

We can easily visualize the transitions in the simplest possible case $N = M = 2$. In this situation the state of the system is described in terms of two variables only. Indeed, from the normalization of the ratings it follows that we can focus on the ratings users give to the first item, i.e., $\hat{r}_{11}$ and $\hat{r}_{21}$. This allows us to visualize the drift given by Eq. (6) through a stream plot in the square ($0 \leqslant \hat{r}_{11} \leqslant 1$,

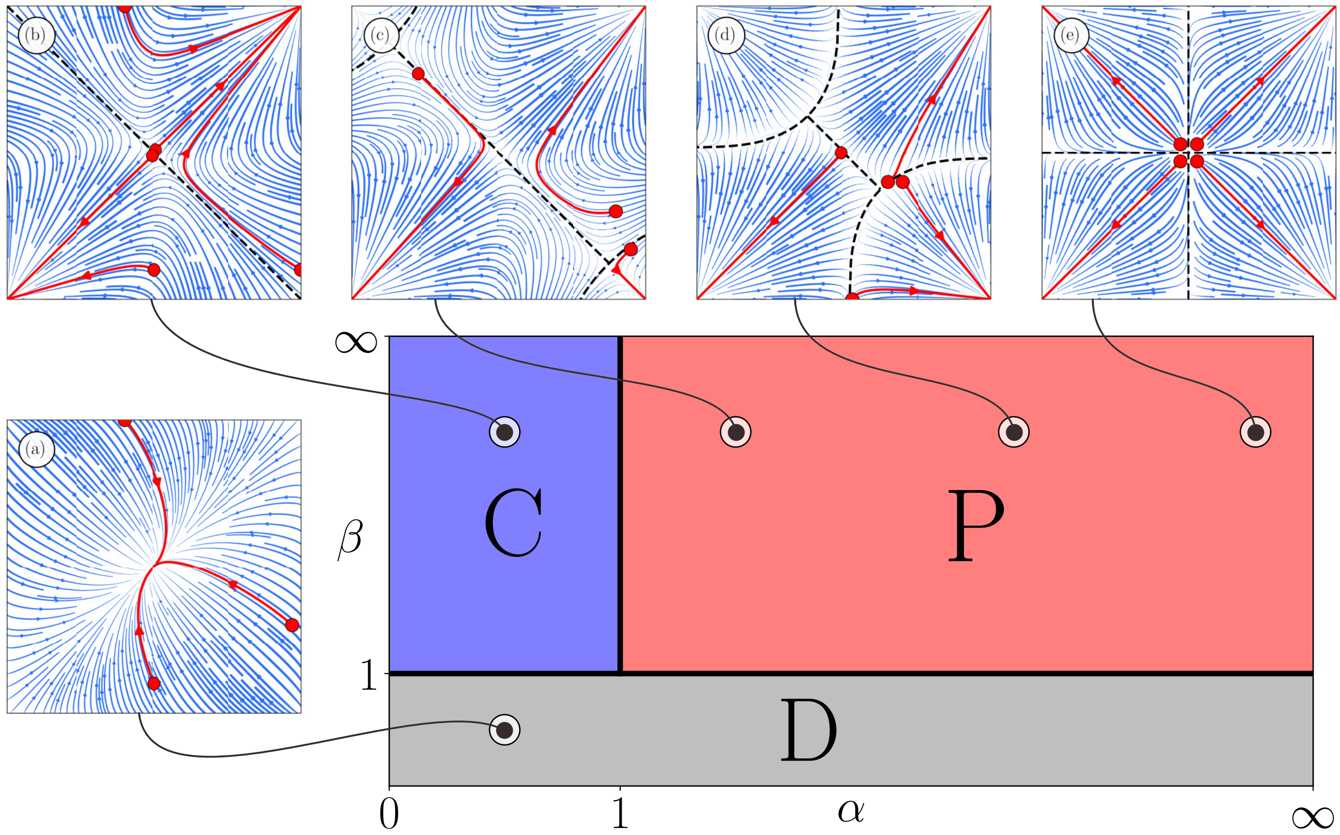


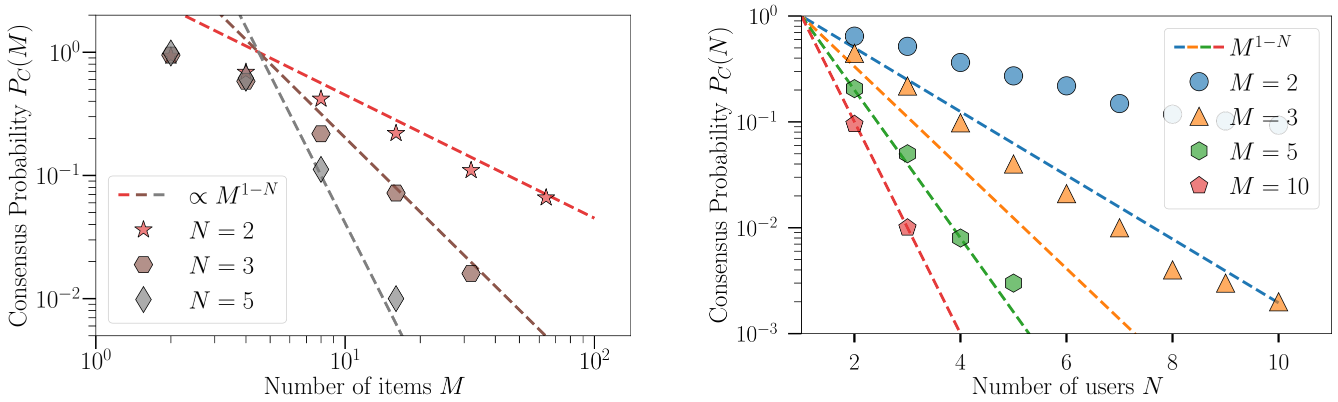
FIG. 2. Phase diagram of the model. Graphical representation of the phase diagram. The (α, β) plane is split up into three regions according to the distinct observed phases. The $\beta = 1$ line divides the multiple-item (disorder) regime from the single-item one. This is in turn split up into the consensus phase, where all users agree on the same item, and the polarization phase where users stick to different items. The line at $\alpha = 1$ is exact only in the $N \rightarrow \infty$ limit, while for finite-size systems the transition (crossover) is observed for larger α. The panels around the diagram show the ratings streamlines for $N = 2,\ M = 2$ at different (α, β) values. The red streamlines represent the evolution of the ratings when starting from the respective red dots, showing the fixed points of the dynamics. The basins of attractions are separated by black dashed lines. In the disorder phase the only fixed point is the point $(\hat{r}_{1,1}, \hat{r}_{2,1}) = (0.5, 0.5)$, as in panel (a). In the consensus phase $(\hat{r}_{1,1}, \hat{r}_{2,1}) \in \{(0, 0), (1, 1)\}$, as in panel (b). Above $(\alpha, \beta) = (1, 1)$ two new attraction basins arise which are related to the emergence of polarization states, $(\hat{r}_{1,1}, \hat{r}_{2,1}) \in \{(1, 0), (0, 1)\}$, as in panels (c) and (d). Eventually, when $\alpha \rightarrow \infty$ (e) the phase space is equally divided into consensus and polarization.

$0 \leqslant \hat{r}_{21} \leqslant 1$). The disordered state D corresponds to the point ($\hat{r}_{11} = \hat{r}_{21} = 0.5$), consensus C to the two points ($\hat{r}_{11} = \hat{r}_{21} = 1,\ \hat{r}_{11} = \hat{r}_{21} = 0$), while polarization P corresponds to the points ($\hat{r}_{11} = 1,\ \hat{r}_{21} = 0$) and ($\hat{r}_{11} = 0,\ \hat{r}_{21} = 1$). Figure 2(a) shows the vector field $\hat{\nu} = (\hat{\nu}_{11}, \hat{\nu}_{21})$ for $\alpha = 0.5$ and $\beta = 0.5$. As discussed above, all stream lines point toward the disordered state D, which is the only attractor of the dynamics. This is also shown by three different trajectories (in red), which all end up in the central point of the stream plot. The situation changes by increasing the popularity bias above the critical value $\beta_c = 1$ to $\beta = 1.5$, as shown in Fig. 2(b). Disorder stops being an attractor, and all the stream lines point toward the two consensus configurations C. The transition is abrupt, since as soon as the popularity bias exceeds β_c, the attraction basin of the disordered state disappears, while that of consensus occupies all the phase space.

In the same way we can investigate the transition driven by the similarity bias. Starting from Fig. 2(b), where consensus is the only attractor, we increase the similarity bias to $\alpha = 1.5$ [Fig. 2(c)], a value above the critical value $\alpha_c = 1$. This makes the polarized state P emerge, although its basin of attraction still remains small compared to that of consensus. For larger values of α the basin of attraction of polarization grows [Fig. 2(d)], reaching the same size of the attraction basin of Consensus in the limit of infinite similarity bias [Fig. 2(e)]: the phase space splits into four quadrants, two belonging to the Consensus attractor, the other two to the Polarization one. Starting from a fully disordered initial condition (the center of the square) and neglecting diffusion, in the $\alpha = \infty$ limit we expect to reach consensus in half of the realizations of the process. For smaller values of $\alpha > 1$ this probability will be larger than 1/2.

C. Numerical investigation of the phase transitions

The phase diagram deduced in the previous subsections and sketched in Fig. 2 can be validated by means of numerical simulations of the model behavior. In particular, it is possible to define two order parameters, related to the variance of normalized ratings, whose values mark the two transitions, associated to the breaking of user or item symmetries. See Appendix F for details. While the transition controlled by



(a)

(b)

FIG. 3. Probability of consensus as a function of M and N. Fraction of 1000 runs going to consensus. Panel (a) shows that, for fixed $\alpha = 10$, P_C tends to 0 as N is increased. Larger values of M imply a better agreement with Eq. (11), derived for $\alpha = \infty$. Panel (b) shows that, already for $\alpha = 2$, the power-law dependence of P_C on M with exponent $1 - N$ [see Eq. (11)] is obeyed for sufficiently large M, with an N-dependent prefactor.

β occurs as expected, with an abrupt jump of the order parameter around $\beta = 1$, the analysis of the consensus-to-polarization transition, controlled by α, requires more care, since above $\alpha_c = 1$, both consensus and polarization solutions are stable.

In the $N = M = 2$ case, in the limit of large α, the phase space splits into four equally sized regions, two belonging to the basin of attraction of consensus, the others to the basin of attraction of polarization. The initial condition we adopt, with all ratings equal, lies in the center of the phase space, corresponding to a perfectly disordered configuration. This means that, neglecting diffusion, the very first random click completely determines whether the system evolves toward consensus or polarization. Since the first click corresponds, in the stream plot, to a step along one of the diagonals with equal probability, we expect consensus to be reached with probability $P_C = 1/2$.

In the case of generic N and M and $\alpha = \infty$ the picture is similar. Since every user is completely independent from the others he or she gets polarized along one item uniformly selected at random among the M possible values. Global consensus is reached only if, by chance, the selected item is the same for all users. All other configurations correspond to polarization. The probability P_C that all users randomly choose the same item is

$$P_C = \frac{M}{M^N} = M^{-N+1}. \tag{11}$$

In the limit of large M (or large N) this probability of consensus P_C vanishes: polarization is the only stationary state actually reached by the dynamics. Consensus gets harder and harder to be observed due to entropic effects. This can be also seen by noticing that the consensus corners form a countable set, while the polarization corners are an uncountable one, since the latter can be also seen as the set of all infinite binary sequences.

Equation (11) implies that, exactly as for $N = M = 2$, for generic N and M and $\alpha = \infty$ the phase space is divided in M^N quadrants of which only M lead to consensus. By analogy with the $N = M = 2$ case we expect that, also in the generic case, the basin of attraction of polarization starts gradually growing from the "corners" of the phase space as α becomes larger than 1, eventually invading the whole quadrants. Note that the system is always initialized in the center of the phase space, because all ratings are initially equal. Then the first random click moves it toward the periphery of the phase space, where the polarization basin lies. Since in the large system limit there are infinitely more polarized corners than consensus ones, we expect random fluctuations to lead the system to a polarized state very easily and thus the consensus probability to go to zero very rapidly also for values of α close to the critical point $\alpha_c = 1$. The scenario just described is confirmed by numerical simulations. Figures 3(a) and 3(b) show a comparison between Eq. (11) and the probability of consensus measured in numerical simulations. For fixed and large enough M, the exponential decay in N, M^{-N+1}, is perfectly recovered in simulations, while for small M we observe strong discrepancies. This behavior derives from the fact that, as discussed in Appendix F, by increasing N only, the transition gets sharper, but occurs at values of α larger than 1. Conversely, when M is increased, the transition moves toward $\alpha_c = 1$ while also getting sharper. Keeping instead N fixed and looking at the consensus probability as a function of M we observe a good agreement for what concerns the scaling exponent, $1 - N$, although with a much larger prefactor. In any case, we can conclude that, for N and M sufficiently large, entropic effects make consensus very hard to be reached as soon as $\alpha > \alpha_c = 1$. Hence in large systems the phase diagram is composed of three pure phases: disorder, consensus, and polarization.

V. CRITICAL RECOMMENDATIONS

The analytical approach and the numerical simulations show that the collaborative-filtering model is characterized by three distinct phases: disorder, consensus, and polarization. Only in the latter does the algorithm really provide personalized recommendations. Indeed, in the disordered phase users get completely random recommendations, while in the consensus phase there is no personalization, as all users receive exactly the same suggestion. Conversely, in the polarization regime users spontaneously split into groups, each characterized by a different recommended item. Thus, in this phase the algorithm provides to each user personalized recommendations perfectly in line with his or her past choices. Note, however, that in the long run each user is exposed only to a single item. Users are trapped into a filter bubble preventing them from being exposed to all other items.

Understanding if and how personalized recommendations can be obtained without users being completely stuck in a filter bubble is of crucial importance. With this goal we focus on the transition between multiple item and single item, where an intermediate behavior is expected to be found. This transition corresponds to the line $\beta = \beta_c = 1$. For such value of the popularity bias, the probability for user u to click on item i reads, from Eq. (1),

$$\begin{aligned} R_{ui} &= \frac{1}{N}\sum_{v=1}^{N}\frac{s_{uv}^{\alpha}}{\sum_w s_{uw}^{\alpha}}\frac{r_{vi}}{\sum_j^M r_{vj}} \\ &\approx \frac{1}{N}\sum_{v=1}^{N}\frac{s_{uv}^{\alpha}}{\sum_w s_{uw}^{\alpha}}\frac{r_{vi}}{t+Mr_0}. \end{aligned}$$

In the disordered phase all normalized ratings $\hat{r}_{ui}$ for a given user u are concentrated around the value $1/M$, and, for large t, they are described by a δ distribution. In the polarized phase they are instead described by the superposition of two δ functions, one in zero and the other in one, corresponding to the winning opinion. We want to understand if, on the critical line, the distribution of ratings assumes a nontrivial form between these two limits. We first consider the limit $\alpha \to \infty$, where each agent is coupled only to his or her past, different agents being completely independent. We can focus on just one user and set $N = 1$. In this way the transition rates become

$$R_{ui} = R_i = \frac{r_i}{t+Mr_0}, \qquad (12)$$

where we indicate for simplicity with r_i the ratings of the user under consideration. Equation (12) can be seen as the transition rate for a Polya Urn model with balls of M distinct colors [40]. Indeed, we can interpret r_i as the number of balls of color i inside the urn and R_i is the probability of randomly extracting a ball of such a color.[1] Since at each time step exactly one item is clicked on and the corresponding rating is increased by a unit, there is a perfect mapping between the collaborative-filtering model for $\alpha = \infty$ and $\beta = 1$ and a Polya Urn with reinforcement parameter $S = 1$. This implies that the probability $P(\hat{\mathbf{r}})$ of observing a normalized rating vector $\hat{\mathbf{r}} = (\hat{r}_1, \dots, \hat{r}_M)$ is given by

$$P(\hat{\mathbf{r}}) = \frac{\prod_i^M r_i^{\mathbf{r_0}/S-1}}{\mathcal{D}(\mathbf{r_0}/S)}, \qquad (13)$$

where $S = 1$ is the reinforcement parameter of the Polya Urn, $\mathbf{r_0} = (r_0, \dots, r_0)$ is the vector of initial conditions and $\mathcal{D}(\cdot)$ is the multivariate β function [40,41].

The distribution of a single normalized rating $P(\hat{r}_i)$ can be obtained by marginalizing Eq. (13) over the remaining $M-1$ ratings, obtaining

$$P(\hat{r}_i) = \frac{\hat{r}_i^{r_0-1}(1-\hat{r}_i)^{(M-1)r_0-1}}{B(r_0,(M-1)r_0)},$$

where $B(x, y)$ is the Euler β function. Depending on the value of r_0 this distribution has different shapes (see Appendix G for details). In particular, for $r_0 = 1/(M-1)$, it is a pure power law with exponent $-(M-2)/(M-1)$. This means that on the critical line $\beta = 1$, for $\alpha = \infty$ users are neither completely polarized nor behaving randomly. Rather they show a nontrivial distribution of the ratings, thus conciliating personalized recommendations with the exploration of the whole item space. Figure 4(a) confirms this prediction also for values of the similarity bias smaller than infinity: for $\alpha \approx 10$ users behave as if they were independent. We can see this also by looking at the popularity $w_i = \sum_u r_{ui}$, whose distribution quantifies how much different users agree on the same items. When users are practically independent, we expect such a distribution to be peaked around its mean value. Indeed, different users select different items as their favorite, and thus summing on all users gives, for all items, approximately the same popularity value. This behavior is shown in the inset of Fig. 4(a), where we report the complementary cumulative distribution of the popularity in the large α regime.

What happens when users cannot be considered effectively independent? For $\alpha = 0$ an approximate mapping to a Polya Urn (see Appendix G) yields a distribution of normalized ratings perfectly analogous to Eq. (13), with only r_0 replaced by r_0/N. As a consequence, $P(\hat{r}_i)$ is a β distribution (see Appendix G), which for initial condition $r_0 = 1/[N(M-1)]$, decays as a power law $\hat{r}_i^{-1}$. This behavior is checked in Fig. 4(b), where we show the distribution of the ratings for various values of α and $r_0 = 1/[N(M-1)]$. For $\alpha = 0$ we observe deviations from the predicted behavior, but as α is increased, the system more closely follows a power-law distribution. Also in this case we show in the inset the complementary cumulative distribution of the popularity, which is broader than the one of a Polya Urn for small α, while it becomes more and more similar to it as α increases.

In conclusion, even if for intermediate values of α an exact mapping to a Polya Urn is not possible, what we observe numerically is that the behavior of the system for $\beta = 1$ and generic α is well described either by the $\alpha = 0$ limit or the $\alpha = \infty$. In all cases we find broad rating distributions. Note that the behavior we observe on the critical line depends on the point where we cross it, i.e., if we move from disorder to consensus or from disorder to polarization. An analogous power-law scaling is found only using a different r_0 in the two cases. In particular, when α is small, users shows a stronger degree of collective consensus, while for larger α they behave more individualistically. It is also important to remark that when M, N are large enough, the transition occurs at $\beta_c = 1$ independently of the number of users or items. This implies that the critical recommendation regime is stable when new

[1] Since r_0 is noninteger, the interpretation in terms of balls in an urn does not strictly apply here. However, the theory for Polya Urns works also for real r_i.

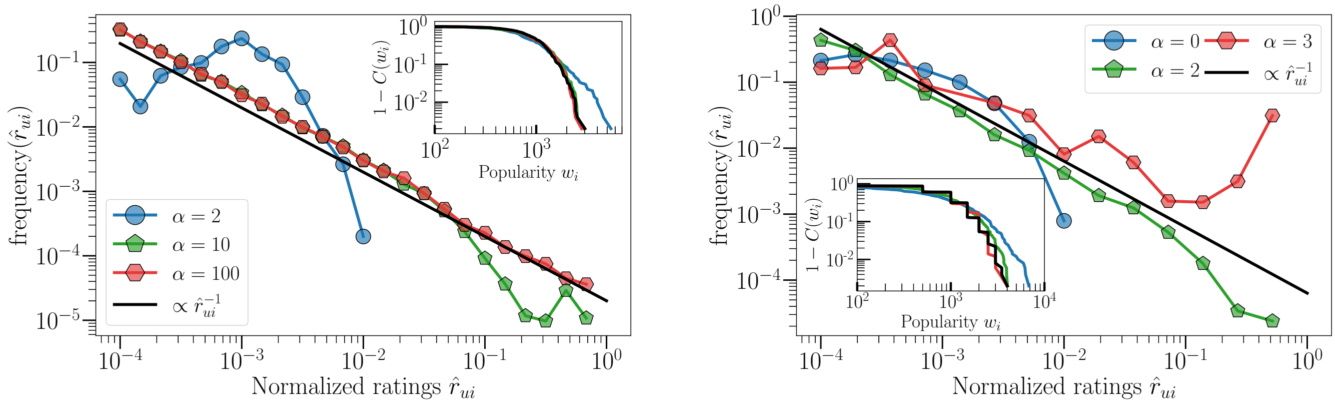


FIG. 4. Distributions of the normalized ratings for $\beta = 1$. (a) Distribution of normalized ratings for $N = 1000$, $M = 500$, $\beta = 1$, and different values of α. The system is initialized with initial conditions $r_0^{(1)} = 1/(M-1)$. In this case we expect a power-law distribution with exponent -1 for large values of α (straight line), which we observe already for $\alpha = 10$. When $\alpha = 2$, the approximation of independent users breaks, and the distribution is no more approximated by a power law. The inset shows the complementary cumulative distribution of the popularity. For large α it coincides with that of a Polya Urn (black line) since users are independent, while as α decreases it becomes broader, due to the emergence of correlations among users. (b) As in the left panel, but initializing the system with initial conditions $r_0^{(2)} = 1/N(M-1)$. In this case the model approximately follows a Polya Urn dynamics for $\alpha = 0$; as expected, the distribution is close to a power law with exponent -1 (straight line). For larger values of α, when users are mostly independent, a peak at $\hat{r_{ui}} \approx 1$ appears in the power law distribution. Also in this case, the inset shows the complementary cumulative distribution of the popularity, that becomes more and more similar to that obtained from independent Polya Urns (black line) as α increases.

users or items enter the system, a crucial requirement for applying the recommendation algorithm in realistic scenarios. Moreover, the standard implementation of the user-user collaborative filtering, corresponding to $\alpha = \beta = 1$, lies on the critical line, and it is thus in the optimal region of the algorithm's phase space.

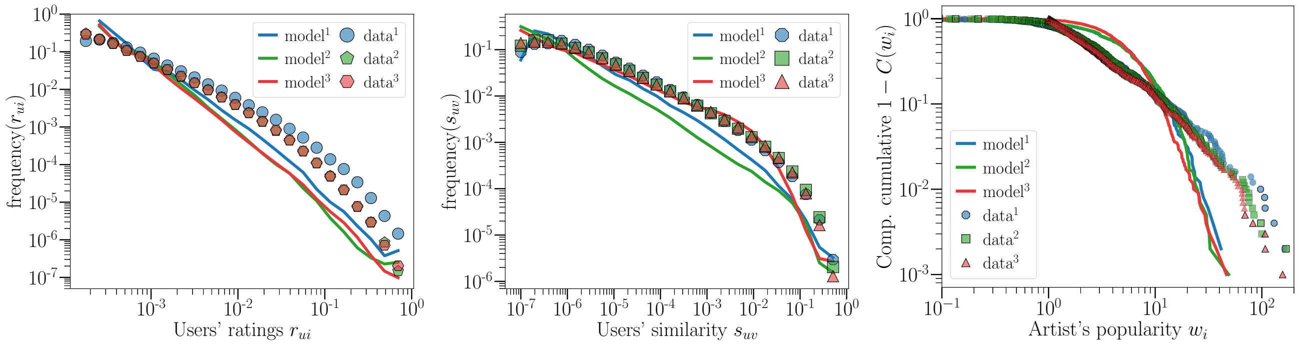
VI. MODELING MUSIC RECOMMENDATIONS

The model we consider describes a recommendation algorithm with implicit feedback, where the ratings are computed from users' behavior and not directly from their votes. This is the typical situation in online music platforms; in such a context the number of times a user plays a song, i.e., what we call rating, is a proxy of how much the user likes that song. Thus, it is very natural to compare the model behavior with data coming from an online music platform. The popular website last.fm is a suitable platform for such a task, as it provides full listening histories of a large amount of its users. These data have been already analyzed in a number of studies [42–44], and they represent a sort of standard in the music recommendation system. In particular, we focus on the Music Listening Histories Dataset, which contains more than 27 billion time-stamped logs extracted from last.fm [45], using it as used in the paper. In order to build a rating matrix out of the dataset, we selected N random users (with $N = 1000, 2000, 5000$) and the top M most popular artists (with $M = 500, 1000$). We then defined the entry r_{ui} as the number of times the uth user listened to the ith artist.

The empirical distributions of the ratings for different combinations of M and N are displayed in Fig. 5(a), while the distributions of the similarities among users are reported in Fig. 5(b). Both distributions turn out to be very broad. Also the popularity of individual artists, $w_i = \sum_u r_{ui}$, is broadly distributed [see Fig. 5(c)]. This is a clear indication that users tend to give high ratings to the same set of artists, i.e., they do not behave independently. If users were completely independent, each one would prefer a different artist, and thus the popularity distribution would tend to be peaked.

The broad distribution of ratings suggests that to reproduce last.fm empirical data we should consider our recommender model at the border between multiple item and single item, i.e., for $\beta = 1$. We perform numerical simulations of the model dynamics for $\beta = 1$ and determine the value of the similarity bias α that best fits the empirical distributions. In this way we quantify the level of interaction among users. Since we know that the popularity is broadly distributed, we set $r_0 = 1/[N(M-1)]$, a value that for $\alpha = 0$ gives a broad popularity distribution. We then perform numerical simulations for various values of α, determining the one best reproducing the data. The model dynamics is run for a time equal to the average number of plays per user in the last.fm dataset. It is worth pointing out that even if we set the average values equal, in the real system the number of plays largely fluctuates from user to user, while in our model each user clicks more or less the same number of times.

Results of numerical simulations are compared with the empirical evidence in Fig. 5, and a partial agreement is observed; more details about the adherence of the model to real data are reported in Appendix G. We selected the α values such that the distributions most closely reproduce empirical data. These values are close to 2 in all the cases considered, suggesting that a non-negligible amount of interaction underlies users behavior in last.fm. These results indicate that our model is capable of reproducing the main features of users' behavior on the online platform last.fm and allow us to gauge the strength of the collaborative filtering they are exposed to, even if other mechanisms we are neglecting also may play



(a) Ratings distributions.

(b) Similarities distributions.

(c) Popularities distributions.

FIG. 5. Empirical data from last.fm (a) Histogram of the empirical ratings from last.fm (symbols) for three combinations of N, M: $(N, M)^1 = (1000, 500)$, $(N, M)^2 = (2000, 1000)$, $(N, M)^3 = (5000, 1000)$. Each rating represents how many times a given user listened to a particular artist. Distributions obtained by simulating the model with a fixed parameter α: $\alpha^{\text{fix},1} = 1.74$, $\alpha^{\text{fix},2} = 1.78$, $\alpha^{\text{fix},3} = 1.98$ are shown as solid lines. (b) Histogram of the cosine similarity for the same set of parameters of the previous panel. Empirical data are plotted as symbols; results from simulations of the model are shown as solid lines. (c) Complementary cumulative distribution of the popularity defined as $w_i = \sum_u r_{ui}$. Empirical data show a wide distribution of the popularity, a feature which cannot be fully recovered by our model.

a role. Despite our model making significant improvement in modeling recommendation algorithms, we observe some discrepancies between numerical simulations and empirical data. These may arise from several features not considered in our schematic representation. For instance, last.fm, like many other online platforms, features a social network structure that influences the content to which users are exposed. Additionally, real users demonstrate temporal correlations in their behavior, fluctuations in music consumption, and personal preferences.

VII. CONCLUSIONS

Understanding the effects of recommendation algorithms in social phenomena and their role in the polarization of opinions is a central problem to be tackled in order to prevent the rise of radicalization. The feedback loop these algorithms tend to establish represents a serious threat to our society, as users are trapped in filter bubbles where they are exposed only to content and news confirming their past beliefs. In this paper we addressed this issue by considering the effects of user-user collaborative filtering, a paradigmatic approach to algorithmic recommendations, on a group of N users allowed to repeatedly choose among M different items or opinions. Depending on the strength of the similarity bias α, which sets the importance the algorithm gives to choices made by similar users, and on the magnitude of the popularity bias β, which gauges the weight given to items with high ratings, the phase diagram of the system is characterized by three phases:

(1) A *disordered* phase for $\beta < 1$ and any α, where all users rate items in a completely random manner

(2) A *consensus* phase for $\beta > 1$ and $\alpha < 1$, where all users share the same opinion

(3) A *polarized* phase for $\beta > 1$ and $\alpha > 1$, where each user sticks to a given item, the system being split into various communities corresponding to different selected items.

None of these phases correspond to viable recommendations. Indeed, in the disordered phase the algorithm recommends just random items; in the consensus phase it treats all users in the same way; and in the polarized phase the filter bubble problem emerges, since each user is exposed to a single opinion. However, by looking at the transition line between disorder and polarization we showed that it is possible to operate the recommendation algorithm in an "ideal" regime, featuring both personalized and diverse recommendations, without the onset of filter bubbles. On this critical line users explore more than just a single opinion, as confirmed by a broad distribution of the normalized ratings. We believe the approach we introduced represents a step toward the development of a general theory of collaborative-filtering-based recommendation algorithms, allowing us to understand their optimal regimes and their potential drawbacks.

We then compared our model with empirical data coming from the last.fm music platform. We observe broad distributions of the user ratings, of the similarity among users, and of the artists popularity. In the framework of our model, these results can be interpreted as the outcome of a recommendation system operating a collaborative-filtering algorithm on the critical line $\beta = 1$. By fitting our model to the data, we were able to qualitatively recover a broad distribution of the ratings, of the popularity, and of the similarity. In particular, we find values of the similarity bias $\alpha \approx 2$, which show the presence of a non-negligible effective interaction among users.

Clearly many realistic ingredients have not been taken into account in the present work and deserve to be investigated. In particular, we assumed that users have an infinite memory, so that the initial condition is never completely forgotten and influences the properties of the steady state. The consideration of a system where only ratings expressed in a finite time window in the past determine the algorithmic recommendation is an extremely interesting avenue for future research. Additional important ingredients we neglected are the heterogeneities in the rate of clicking, as observed in last.fm, or possible different initial conditions for users. Understanding whether these modifications change the overall phenomenology is an interesting further development. Also, for a closer comparison with empirical data from real recommender systems it would be interesting to analyze not only the properties of the stationary state of the system, but also the timescales needed to reach it. Finally, while this study intentionally excluded network structure to maintain model simplicity, many platforms, including last.fm [46], are characterized by the presence of a social network, which is an important factor in the enhancement of polarization [31,47]. Future research will delve into the social interaction aspect that has been set aside in this initial model, adding depth to the understanding of the influence of these algorithms.

Despite the heterogeneity of users and artists and the limitations we discussed, our model with small similarity bias reproduces the phenomenology of last.fm users. These results point to the presence of a collaborative-filtering-based recommendation algorithm in the online music platform we considered and show that a relatively simple model can capture its main features and allow one to assess the relevance of its rating and similarity biases.

APPENDIX A: SCALING OF INITIAL CONDITIONS

Let us consider a specific item that has been clicked on k times at time t, i.e., $r_{ui}(t) = r_0 + k$. The variation of the normalized rating $\delta\hat{r} = \hat{r}_{ui}(t+1) - \hat{r}_{ui}(t)$, assuming it is clicked once in a time unit, is

$$\delta\hat{r} = \frac{(M-1)r_0 + t - k}{(Mr_0 + t + 1)(Mr_0 + t)} \sim \frac{1}{Mr_0},$$

where the limit of $t \ll Mr_0$ has been taken, implying that we are dealing with the very first moments of the dynamics. Depending on the scaling of r_0 with M we can distinguish three possibilities:

(1) If $r_0 \sim M^{-a}$ with $a > 1$, $\delta\hat{r}$ vanishes when $M \to \infty$; this means that in the large M limit the system is not able to move in the phase space

(2) If $r_0 \sim M^{-a}$ with $a < 1$, $\delta\hat{r}$ increases with M, being upper bounded by 1; this means that in the large M limit the system is able to arrive with the first click on the border of the phase space not being able to explore it all

(3) If instead $r_0 \sim M^{-1}$, then $\delta\hat{r}$ remains constant to a value smaller than 1 in the large M limit; this is the right scaling we are looking for to preserve the phenomenology of the system regardless of the value of M.

The conclusion in that in order to take the large M limit the initial condition must scale as $r_0 \sim 1/M$.

APPENDIX B: MASTER EQUATION FOR THE RATINGS

Denoting by $Q(r_{ui}, t)$ the probability distribution of r_{ui} at time t, its temporal evolution is

$$Q(r_{ui}, t+\delta t) = Q(r_{ui}-1, t)R_{ui}(r_{ui}-1) + [1 - R_{ui}(r_{ui})]Q(r_{ui}, t),$$

reflecting the nondecreasing evolution of the r_{ui}.

Expanding the l.h.s. in the continuous time limit $\delta t = 1/N \to 0$ one obtains Eq. (3)

$$\frac{d}{dt}Q(r_{ui}, t) = \frac{1}{\delta t}\left[\begin{matrix} Q(r_{ui}-1, t)R_{ui}(r_{ui}-1) \\ -R_{ui}(r_{ui})Q(r_{ui}, t) \end{matrix}\right].$$

The drift coefficient, i.e., the time derivative of the average value $\langle r_{ui}(t)\rangle$, can be derived by writing

$$\frac{d\langle r_{ui}(t)\rangle}{dt} = \sum_{r_{ui}=r_0}^{\infty} r_{ui}\frac{d}{dt}Q(r_{ui}, t)$$

and inserting Eq. (3) into it,

$$\begin{aligned}\frac{d\langle r_{ui}(t)\rangle}{dt} &= \frac{1}{\delta t}\sum_{r_{ui}=r_0}^{\infty} r_{ui}[Q(r_{ui}-1, t)R_{ui}(r_{ui}-1) \\ &\quad - Q(r_{ui}, t)R_{ui}(r_{ui}), t)] \\ &= \frac{1}{\delta t}\sum_{r_{ui}=r_0}^{\infty}[(r_{ui}+1) - r_{ui}]Q(r_{ui}, t)R_{ui}(r_{ui}) \\ &= \frac{1}{\delta t}\sum_{r_{ui}=r_0}^{\infty} Q(r_{ui}, t)R(r_{ui}) = \frac{1}{\delta t}\langle R_{ui}\rangle \\ &= N\langle R_{ui}\rangle,\end{aligned}$$

where we have set $Q(r_0 - 1, t)R(r_0 - 1) = 0$.

So far we have considered the values of the ratings. If we want to consider the normalized ratings $\hat{r}_{ui}(t) \approx r_{ui}(t)/(t + Mr_0)$, the drift is given by

$$\begin{aligned}\hat{v}_{ui} &= \frac{d\langle\hat{r}_{ui}\rangle}{dt} \approx \frac{d}{dt}\left(\frac{\langle r_{ui}\rangle}{t + Mr_0}\right) \\ &= \frac{1}{t + Mr_0}\frac{d\langle r_{ui}\rangle}{dt} - \frac{\langle r_{ui}\rangle}{(t + Mr_0)^2}.\end{aligned}$$

APPENDIX C: STATIONARY SOLUTIONS

The system of differential equations (6) at stationarity reads

$$\hat{r}_{ui} = NR_{ui} = \sum_{v=1}^{N} \frac{s_{uv}^{\alpha}(t)}{\sum_w s_{uw}^{\alpha}(t)}\frac{\hat{r}_{vi}^{\beta}(t)}{\sum_j \hat{r}_{vj}^{\beta}(t)}. \tag{C1}$$

The solution of these $N \times M$ equations gives the stationary states of the model.

The configurations fulfilling this requirement are the following.

Disorder solution. Under this condition all normalized ratings are equal to $\hat{r}_{ui} = 1/M$ for large times, since all items are rated on average the same number of times. Similarities are then all equal to 1, thus Eq. (C1) is fulfilled for any value of u and i:

$$\frac{1}{M} = \sum_{v=1}^{N}\frac{1}{N}\frac{M^{-\beta}}{MM^{-\beta}} = \frac{1}{M}.$$

In this phase the distribution of normalized ratings is a δ function centered around $\hat{r}_{ui} = 1/M$, while the distribution of similarities is a δ function centered around $s_{uv} = 1/M$.

Consensus solution. Only one item is clicked on in the limit $t \gg 1$, then $\hat{r}_{ui} = 1$ for some i, and $\hat{r}_{uj} = 0$, $\forall j \neq i$, and this holds for any user. Since also in this case all similarities are equal to 1, Eq. (C1) reads, for item i,

$$1 = \sum_{v=1}^{N}\frac{1}{N}\frac{1^{\beta}}{1^{\beta} + (M-1)0^{\beta}} = 1,$$

while for the items $j \neq i$

$$0 = \sum_{v=1}^{N} \frac{1}{N} \frac{0^{\beta}}{1^{\beta} + (M-1)0^{\beta}} = 0.$$

Hence the stationarity condition is satisfied for any u and i. In this phase the distribution of the normalized ratings is the superposition of two δ functions, one centered in 0 (weight $1 - 1/M$) and one centered in 1 (weight $1/M$). The distribution of similarities is a δ function centered in 1.

Polarization solution. Only one item is clicked on by each user in the limit $t \gg 1$, but it is not the same for all users. Let us assume that for K users $\hat{r}_{ui_1} = 1$ and $\hat{r}_{uj} = 0$ for $j \neq i_1$, while for the other $N - K$ users $\hat{r}_{ui_2} = 1$ and $\hat{r}_{uj} = 0$ for $j \neq i_2$. In this case the similarity is 1 for users rating the same item, while it is 0 otherwise. It is simple to check that Eq. (C1) is satisfied. It is also easy to check that the stationarity condition is satisfied by any possible polarized configuration (more than two groups, of any size). In this phase, the distribution of normalized ratings is again the superposition of two δ functions, centered in 0 and 1, as for the consensus solution. Also the distribution of similarities is given by the superposition of a δ function centered in 0 and one centered in 1, with weights depending on the distribution of group size.

Other solutions. In principle, there exist also other stationary solutions. These are all configurations where multiple items are rated by a single user exactly in the same proportion. For example, $m < M$ items are equally rated by user u, i.e., $\hat{r}_{ui_1} = \cdots = \hat{r}_{ui_m} = 1/m$ for m items, while $r_{uj} = 0$ for $j \neq i_1, \ldots, i_m$. These solutions are stationary both if the m items are the same for all users and if they differ for different users. This holds for any value of m (for $m = 1$ we have single-item solutions, while for $m = M$ we have disorder). In the next Appendix it is shown that these solutions are unstable for any values of the parameters β and α.

APPENDIX D: THE TRANSITION FROM MULTIPLE ITEM TO SINGLE ITEM

In this Appendix we present explicit calculations about the *multiple-item* to *single-item* transition occurring as a function of β, for any value of α.

1. The multiple-item solution

Let us start from the case $\alpha = \infty$. Plugging the disorder solution Eq. (7) into the expression Eq. (6) we obtain, apart from the factor $\frac{1}{t+Mr_0}$,

$$\frac{d\hat{r}_{ui}}{dt} = \frac{d\epsilon_{ui}}{dt} \propto \frac{\left(\frac{1}{M} + \epsilon_{ui}\right)^{\beta}}{\sum_j \left(\frac{1}{M} + \epsilon_{uj}\right)^{\beta}} - \left(\frac{1}{M} + \epsilon_{ui}\right) \tag{D1}$$

$$\propto \frac{(1 + M\epsilon_{ui})^{\beta}}{\sum_j (1 + M\epsilon_{uj})^{\beta}} - \left(\frac{1}{M} + \epsilon_{ui}\right). \tag{D2}$$

Expanding for $M\epsilon_{ui} \ll 1$ leads to

$$\frac{d\epsilon_{ui}}{dt} \propto \frac{(1 + \beta M\epsilon_{ui})}{\sum_j (1 + \beta M\epsilon_{uj})} - \left(\frac{1}{M} + \epsilon_{ui}\right).$$

Since $\sum_j \epsilon_{uj} = 0$ [see Eq. (8)] then $\sum_j (1 + \beta M\epsilon_{uj}) = M$ and hence

$$\frac{d\epsilon_{ui}}{dt} \propto \frac{1}{M}(1 + \beta M\epsilon_{ui}) - \frac{1}{M}(1 + M\epsilon_{ui}) = (\beta - 1)\epsilon_{ui},$$

that is, Eq. (9).

If α is finite, the similarities do not cancel from Eq. (6), and we should consider all of them. We now show that similarity terms are equal to 1 up to corrections of the second order in ϵ_{ui}, which we can safely neglect for any $\alpha > 0$. In this case the situation is substantially equivalent to the case $\alpha = 0$, where all similarities are exactly equal to 1.

Let us recall the expression of the cosine similarity

$$s_{uv} = \frac{\sum_i \hat{r}_{ui}\hat{r}_{vi}}{\sqrt{\sum_i \hat{r}_{ui}^2}\sqrt{\sum_i \hat{r}_{vi}^2}}$$

and plug into it the ratings corresponding to the disorder solution Eq. (7),

$$s_{uv} = \frac{\sum_i \left(\frac{1}{M} + \epsilon_{ui}\right)\left(\frac{1}{M} + \epsilon_{vi}\right)}{\sqrt{\sum_i \left(\frac{1}{M} + \epsilon_{ui}\right)^2}\sqrt{\sum_i \left(\frac{1}{M} + \epsilon_{vi}\right)^2}}.$$

Multiplying by M^2 both the numerator and the denominator and expanding the terms in the square roots at denominator [which is equivalent to neglect terms of order $O(\epsilon^2)$] we can rewrite

$$s_{uv} \approx \frac{\sum_i (1 + M\epsilon_{ui})(1 + M\epsilon_{vi})}{\sqrt{\sum_i (1 + 2M\epsilon_{ui})}\sqrt{\sum_i (1 + 2M\epsilon_{vi})}}$$

$$= \frac{\sum_i (1 + M\epsilon_{ui} + M\epsilon_{vi} + M^2\epsilon_{ui}\epsilon_{vi})}{\sqrt{\sum_i (1 + 2M\epsilon_{ui})}\sqrt{\sum_i (1 + 2M\epsilon_{vi})}}.$$

Exploiting once again the fact that $\sum_i \epsilon_{ui} = \sum_i \epsilon_{vi} = 0$ we finally obtain that

$$s_{uv} \approx \frac{M + M^2 \sum_i \epsilon_{ui}\epsilon_{vi}}{M} = 1 + M \sum_i \epsilon_{ui}\epsilon_{vi},$$

thus confirming the fact that $s_{uv} = 1 - O(M^2\epsilon^2)$; we specify the minus sign to remind that similarities cannot be larger than 1. Note that, for this reason, the term $\sum_i \epsilon_{ui}\epsilon_{vi}$ is always negative.

To show what happens if similarities can be put equal to 1, for simplicity let us consider directly the case $\alpha = 0$ where similarities are exactly equal to 1. In this case expression (6) reduces to

$$\frac{d\hat{r}_{ui}}{dt} \approx \frac{1}{t + Mr_0}\left(\frac{1}{N}\sum_v \frac{\hat{r}_{vi}^{\beta}}{\sum_j \hat{r}_{vj}^{\beta}} - \hat{r}_{ui}\right)$$

$$\approx \frac{1}{t + Mr_0}\left[\frac{1}{N}\sum_v \frac{\left(\frac{1}{M} + \epsilon_{ui}\right)^{\beta}}{\sum_j \left(\frac{1}{M} + \epsilon_{uj}\right)^{\beta}} - \left(\frac{1}{M} + \epsilon_{ui}\right)\right].$$

Following the same procedure as in the case $\alpha = \infty$, we can simplify this expression obtaining

$$\frac{d\hat{r}_{ui}}{dt} = \frac{d\epsilon_{ui}}{dt} \propto \frac{1}{N}\beta \sum_v \epsilon_{vi} - \epsilon_{ui} = \beta\langle\epsilon_{vi}\rangle - \epsilon_{ui},$$

where $\langle\cdot\rangle = \frac{1}{N}\sum_v$.

Reapplying the average over users we obtain

$$\frac{d\langle\epsilon_{ui}\rangle}{dt} = \langle\beta\langle\epsilon_{vi}\rangle - \epsilon_{ui}\rangle = (\beta-1)\langle\epsilon_{vi}\rangle.$$

Similarly to the case $\alpha = \infty$, in which we had only one user, this expression means that when $\beta < 1$, the items which are on average underrated would tend to increase, while items which are on average overrated would tend to decrease, ensuring the stability of the disorder solution. When $\beta > 1$ the situation is reversed, and thus the disorder solution is unstable.

2. The single-item solution

Let us now consider a configuration close to a single-item solution, i.e., a situation in which the normalized ratings are as in Eq. (10).

Let us start from the case $\alpha = \infty$ for simplicity. Plugging Eq. (10) into Eq. (6) yields

$$-\frac{d\epsilon}{dt} = \frac{1}{t+Mr_0}\left[\frac{(1-\epsilon)^\beta}{(1-\epsilon)^\beta + (M-1)\frac{\epsilon^\beta}{(M-1)^\beta}} - (1-\epsilon)\right]$$
$$= \frac{1}{t+Mr_0}\left[\frac{(1-\epsilon)^\beta - (1-\epsilon)^{\beta+1} - \dfrac{\epsilon^\beta(1-\epsilon)}{(M-1)^{\beta-1}}}{(1-\epsilon)^\beta + \dfrac{\epsilon^\beta}{(M-1)^{\beta-1}}}\right].$$

Neglecting the prefactor and the denominator, which are positive, stability requires the numerator to be positive,

$$(1-\epsilon)^\beta - (1-\epsilon)^{\beta+1} - \frac{\epsilon^\beta(1-\epsilon)}{(M-1)^{\beta-1}} > 0.$$

Some straightforward algebra leads to

$$\left[(M-1)\left(\frac{1}{\epsilon}-1\right)\right]^{\beta-1} > 1.$$

Raising to the power $1/(\beta-1)$, the sign of the inequality is conserved if $\beta > 1$, and thus

$$(M-1)\left(\frac{1}{\epsilon}-1\right) > 1 \Rightarrow \epsilon < \frac{M-1}{M}.$$

Hence for small ϵ the single-item solution is stable if $\beta > 1$.

On the other hand, if $\beta < 1$, the sign of the inequality is reversed, and thus we obtain

$$(M-1)\left(\frac{1}{\epsilon}-1\right) < 1 \iff \epsilon > \frac{M-1}{M}.$$

In such a case for small ϵ the single-item solution is unstable.

A similar computation can be carried out for the drift of the least rated items $j \neq i$; it is easy to obtain that $d\hat{r}_{uj}/dt < 0$ for $\beta > 1$ as long as $\epsilon < (M-1)/M$, while $d\hat{r}_{uj}/dt > 0$ for $\beta < 1$ under the same assumption. All these results indicate that the multiple-item solution is stable for $\beta < 1$, while the single-item solution is the stable one for $\beta > 1$.

Considering the case with finite α does not change much the reasoning. The only difference would be to consider explicitly the similarity terms in Eq. (6); we can get rid of them considering a consensus single-item configuration, i.e., where all users have similarity very close to 1; the computations then trivially reduce to the case of considering only one user when $\alpha = \infty$.

Finally, we look also at the stability of other possible solutions of Eq. (C1), those in which $m < M$ items are rated in the same proportion:

$$\begin{array}{ll} \hat{r}_{ui} \sim \dfrac{1}{m} + \epsilon_{ui} & \text{for } m \text{ items} \\ \hat{r}_{uj} \sim \epsilon_{uj} & \text{for } M-m \text{ items} \end{array}.$$

Performing computations similar to those previously explained, it is possible to realize that these solutions are stable neither for $\beta < 1$, because the large ratings would decrease while the small ones would increase (deviations are suppressed), nor for $\beta > 1$, because the large ones would increase while the small ones would decrease (deviations are amplified).

APPENDIX E: THE TRANSITION FROM CONSENSUS TO POLARIZATION

In this Appendix we present explicit calculations about the transition between consensus and polarization occurring, as a function of α, for $\beta > 1$.

Let us consider two users u and v who are close to the single-item state. Each of them is described by Eq. (10), but with different selected items: $i_u \neq i_v$. The similarity between these two users is, in the limit $M\epsilon \ll 1$,

$$s_{uv} = \frac{2(M-1)\epsilon - M\epsilon^2}{1 - 2\epsilon + M\epsilon^2} \approx \frac{2(M-1)\epsilon}{1-2\epsilon} = O(M\epsilon). \tag{E1}$$

Let us consider $\beta = \infty$. If K users are close to the single-item state with item i_u selected, while the remaining $N-K$ users have selected item i_v, the expression of the drift is

$$\frac{d\hat{r}_{ui}}{dt} = \frac{1}{t+Mr_0}\left(\frac{\sum_v s_{uv}^\alpha \delta_{i,i_u}}{\sum_w s_{uw}^\alpha} - \hat{r}_{ui}\right). \tag{E2}$$

Since similarities between users in the same group are $s_{uv} \approx 1$, while similarities between users of different groups are $s_{uv} \approx M\epsilon$, we can write

$$\frac{d\hat{r}_{ui_u}}{dt} \approx \frac{1}{t+Mr_0}\left[\frac{K}{K+(N-K)(M\epsilon)^\alpha} - (1-\epsilon)\right].$$

The sign of the drift coefficient is thus related to the sign of the following expression:

$$K - (1-\epsilon)[K + (N-K)(M\epsilon)^\alpha]$$
$$= -(N-K)(M\epsilon)^\alpha + K\epsilon + (N-K)M^\alpha\epsilon^{\alpha+1}. \tag{E3}$$

If $\alpha \in [0, 1)$, taking the limit $\epsilon \to 0$, the dominant term is the one of order ϵ^α, and $d\hat{r}_{ui_u}/dt < 0$. So in this case the rating of the most rated item would tend to decrease, and the polarized solution is not stable. On the contrary, if $\alpha > 1$, the dominant term is the one of order ϵ, so that $d\hat{r}_{ui_u}/dt > 0$ and the polarized solution is stable.

At the same time, if one considers $j \neq i_u$, an analogous computation gives, in the limit of small ϵ, $d\hat{r}_{uj}/dt > 0$ for $\alpha < 1$ (instability of the polarized solution) $d\hat{r}_{uj}/dt < 0$ for $\alpha > 1$ (stability of the polarized solution). In particular, for $\alpha = 1$ the drift is still negative since it is proportional to

$$-(N-K)\epsilon + K\epsilon + (N-K)\epsilon^2 = -N\epsilon + (N-K)\epsilon^2,$$

which is a negative quantity if $\epsilon < 1$. Thus we can conclude that at the critical point $\alpha = 1$, the basin of attraction of the polarized solutions is still null. A very rough estimate of the amplitude of the basin of attraction can be obtained from the expression (E3) by neglecting the higher order term $(N - K)\epsilon^{\alpha+1}$, thus obtaining a solvable inequality which states that the drift points toward the polarized solutions at least until

$$\epsilon < \left(\frac{K}{N-K}\right)^{\frac{1}{\alpha-1}}.$$

The same argument applies to any possible polarized solution, i.e., every value of K between 1 and $N - 1$ and an arbitrary number of groups.

For what concerns the case of generic $\beta > 1$, the only difference with Eq. (E2) is that the terms $\hat{r}_{ui}^{\beta}/\sum_j \hat{r}_{uj}^{\beta}$ must be considered explicitly; but in the limit of small ϵ we have

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\hat{r}_{ui_u}^{\beta}}{\sum_j \hat{r}_{uj}^{\beta}} &= \lim_{\epsilon\to 0}\frac{(1-\epsilon)^{\beta}}{(1-\epsilon)^{\beta} + (M-1)\frac{\epsilon^{\beta}}{(M-1)^{\beta}}}\\ &= \lim_{\epsilon\to 0} 1 - O(\epsilon) = 1\\ \lim_{\epsilon\to 0}\frac{\hat{r}_{ui\neq i_u}^{\beta}}{\sum_j \hat{r}_{uj}^{\beta}} &= \lim_{\epsilon\to 0}\frac{\frac{\epsilon^{\beta}}{(M-1)^{\beta}}}{(1-\epsilon)^{\beta} + (M-1)\frac{\epsilon^{\beta}}{(M-1)^{\beta}}}\\ &= \lim_{\epsilon\to 0} O(\epsilon) = 0;\end{aligned}$$

thus we are led back to Eq. (E2). As a consequence, for every $\beta > 1$, we obtain that the critical value of the transition is $\alpha = 1$.

APPENDIX F: THE ORDER PARAMETERS

We define the item fluctuations V_i as

$$V_i = \frac{M^2}{M-1}\frac{1}{N}\sum_u\left[\frac{1}{M}\sum_i \hat{r}_{ui}^2 - \left(\tfrac{1}{M}\sum_i \hat{r}_{ui}\right)^2\right], \qquad \text{(F1)}$$

which is the variance of users' normalized rating vectors $\hat{\mathbf{r}}_u = (\hat{r}_{u1}, \hat{r}_{u2}, \dots, \hat{r}_{uM})$ averaged over all users. The factor $M^2/(M-1)$ ensures that V_i varies in the interval [0, 1]. In the disordered phase it holds $\hat{r}_{ui} = 1/M$, and so we have $V_i = 0$. Conversely, when users stick to a single item, the normalized ratings satisfy $\hat{r}_{ui} = 1$ and $\hat{r}_{uj} = 0$ for $j \neq i$, thus giving $V_i = 1$.

Analogously, we introduce the user fluctuations V_u as

$$V_u = \frac{M^{2(N-1)}}{M^{N-1}-1}\frac{1}{M}\sum_i\left[\frac{1}{N}\sum_u \hat{r}_{ui}^2 - \left(\tfrac{1}{N}\sum_u \hat{r}_{ui}\right)^2\right], \qquad \text{(F2)}$$

which is the variance of the items' normalized rating vectors $\hat{\mathbf{r}}_i = (\hat{r}_{1i}, \hat{r}_{2i}, \dots, \hat{r}_{Ni})$ averaged over all users and normalized to be in [0, 1]. It is easy to see that both in disorder and consensus it holds $V_u = 0$, while in polarized states $V_u > 0$, the precise value depending on the size distribution of the groups. In terms of these order parameters the three phases are identified by the following:

Disorder: $V_i = 0$ and $V_u = 0$
Consensus: $V_i > 0$ and $V_u = 0$
Polarization: $V_i > 0$ and $V_u > 0$.

In order to validate the phase diagram discussed above, we perform computer simulations of the model dynamics, focusing on the order parameters just defined. We run the simulations for a time equal to $T = 1000M$.

First, we look at the multiple-item to single-item transition, occurring as a function of β. We show in Figs. 6(a) and 6(b) the order parameter V_i as a function of β for $\alpha = 5$ and different combinations of M and N. As expected we observe a transition in $\beta_c = 1$ that becomes increasingly sharper as the system size grows.

Figure 6(c) shows how the order parameter V_u depends on α for fixed N. As M is increased, the transition becomes sharper and sharper and moves to the critical point α_c. For fixed M the transition is instead observed at $\alpha > \alpha_c = 1$ and becomes sharper with growing N [Fig. 6(d)]. This indicates that our predictions are accurate when M is sufficiently large, a situation always occurring in real recommendation systems.

APPENDIX G: THE POLYA URN MODEL

The Polya Urn model considers an urn where initially there are $\boldsymbol{r_0} = \{r_{10}, \dots, r_{M0}\}$ balls of M different colors; at each time step a ball is randomly extracted from the urn, then it is reinserted with S other balls of the same color.

From the general results of the Polya Urn model [40] it can be obtained that the distribution of the normalized ratings follows a multivariate β distribution:

$$P(\hat{\mathbf{r}}) = \frac{\prod_i r_i^{\frac{r_i^{(0)}}{S}-1}}{\mathcal{D}\left(\frac{\mathbf{r_0}}{S}\right)} = \frac{\prod_i \hat{r}_i^{r_0-1}}{\mathcal{D}(\mathbf{r_0})},$$

where $\mathcal{D}$ is the multivariate β function, $S = 1$ since each rating is increased by one subsequently to a click, and we focus on the case of uniform initial conditions $r_i^{(0)} = r_0$ for any i. This distribution is defined on the standard $(M - 1)$ simplex by the constraint of the normalization of the ratings.

These initial conditions also imply that each rating $\hat{r}_i$ is statistically equivalent to the others; thus it is sufficient to obtain the marginal distribution of a single rating to have the complete statistics of the system.

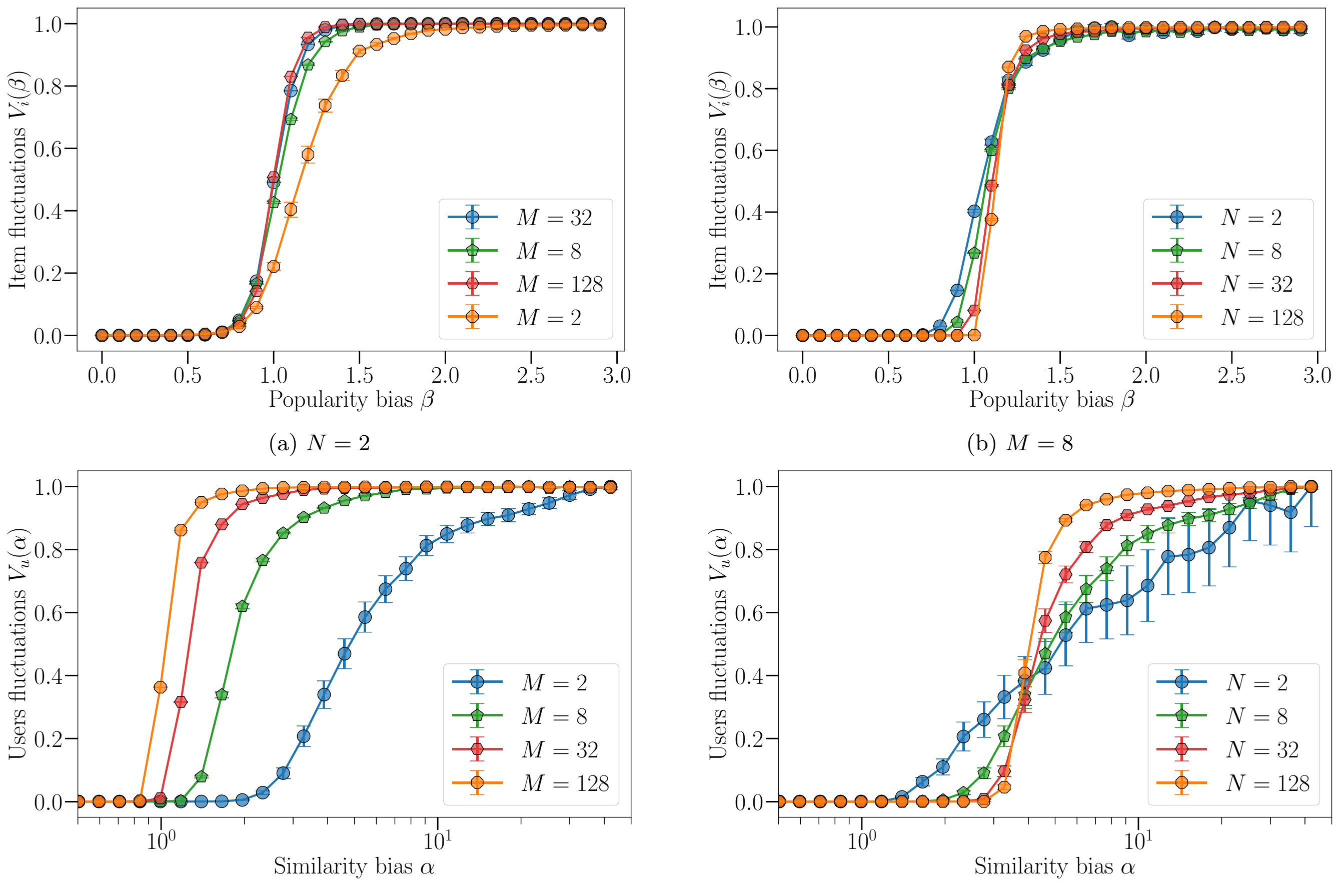
In particular, the marginal one-dimensional distribution of a multivariate β function gives a β distribution $\beta(x_i; a_i, \sum_{j\neq i} a_j)$, which in our case reads

$$P(\hat{r}_i) = \beta[\hat{r}_i; r_0, (M-1)r_0] = \frac{\hat{r}_i^{r_0-1}(1-\hat{r}_i)^{(M-1)r_0-1}}{B[r_0, (M-1)r_0]},$$

where $B(x, y)$ is the Euler β function.

Depending on the value of r_0 we thus have different behaviors. For $r_0 = 1/(M - 1)$ the distribution of the single ratings shows an exact power-law decay:

$$P(\hat{r}_i) = \frac{\Gamma\left(\frac{1}{M-1}\right)}{\Gamma\left(\frac{M}{M-1}\right)}\,\hat{r}_i^{-(M-2)/(M-1)}. \qquad \text{(G1)}$$



(a) $N = 2$

(b) $M = 8$

(c) $N = 8$

(d) $M = 2$

FIG. 6. Order parameters and phase transitions. Panels (a) and (b) show the order parameter V_i for the multiple-item to single-item transition as a function of β, for $\alpha = 5$. Curves are computed averaging over 1000 realizations of the dynamics. The increase of M or N has no effect on the critical value β_c but simply sharpens the transition. Panels (c) and (d) report the behavior of the order parameter V_u for the consensus-polarization transition. It turns out that increasing M moves the transition towards $\alpha_c = 1$. Increasing N for fixed M has little effect with regard to the location of the transition while it makes the transition sharper. This indicates that at fixed M, also in the limit $N \to \infty$ we can see the transition only after a value $\alpha_c^{\text{eff}} > 1$.

If $r_0 < 1/(M-1)$, the power-law decay for small $\hat{r}_i$ is followed by a divergence in $\hat{r}_i = 1$ so that the distribution is bimodal. If instead $r_0 > 1/(M-1)$, then $P(\hat{r}_i)$ is truncated by an exponential cutoff.

The mapping to a Polya Urn can be used also for the case $\alpha = 0$, although in this case the matching is only approximate. Indeed, for $\alpha = 0$ Eq. (1) becomes

$$R_{ui} = \frac{1}{N^2}\sum_v \frac{r_{vi}}{\sum_j r_{vj}},$$

and since on average each user is updated once at each time step we can write

$$R_{ui} \approx \frac{1}{N^2}\sum_v \frac{r_{vi}}{t+Mr_0} \approx \frac{1}{N}\frac{r_{ui}}{t+Mr_0},$$

where we have assumed that, since $\alpha = 0$, all users behave the same so that $1/N\sum_v r_{vi} \approx r_{ui}$. Defining the variable $w_{ui} = r_{ui}/N$ we get

$$R_{ui} \approx \frac{w_{ui}}{t+Mr_0},$$

which has the same form of Eq. (12). As a consequence also the dynamics of the variable w_{ui} is described by a Polya Urn, but now the reinforcement parameter is $S = 1/N$. Since the normalized variables $\hat{r}_{ui}$ and $\hat{w}_{ui}$ coincide, we can then write

$$P(\hat{r}_{ui}) = P(\hat{w}_{ui})$$
$$= \frac{\hat{r}_i^{r_0N-1}(1-\hat{r}_i)^{(M-1)r_0N-1}}{B(r_0N,(M-1)r_0N)}.$$

If we set $r_0 = 1/[N(M-1)]$, this expression implies a power-law decay for $\hat{r}_{ui}$, with exponent $-(M-2)/(M-1)$. As in the case $\alpha = \infty$ we considered above, other values of r_0 give different broad distributions.

APPENDIX H: ADJUSTED R^2 FOR REAL DATA FITS

In this Appendix, we present the outcomes of the adjusted R^2 analysis for the distributions of ratings, similarity, and popularity. The $\overline{R}^2$ is a measure of the extent to which the variation in the dependent variable is predictable from the independent variable. This value aids in evaluating the efficacy of a model's prediction against actual observed data.

TABLE I. Adjusted R^2 for the distribution of the Ratings, Similarity, and Popularity. In this last case, real data are also compared with the results from a standard Polya Urn model.

N	M	$\overline{R}^2_{\text{rat}}$	$\overline{R}^2_{\text{sim}}$	$\overline{R}^2_{\text{pop,CF}}$	$\overline{R}^2_{\text{pop,urn}}$
1000	500	0.880	0.964	0.639	0.411
2000	1000	0.909	0.933	0.726	0.504
5000	1000	0.932	0.963	0.585	0.292

Given the real data $(x_1, \dots, x_n)$ and their relative predictions $(z_1, \dots, z_n)$, then we define the R^2 as

$$R^2 = 1 - \frac{\Sigma_{\text{res}}}{\Sigma_{tot}}$$

where $\Sigma_{\text{res}} = \sum_i (x_i - z_i)^2$ is the sum of squares of the residuals, while $\Sigma_{tot} = \sum_i (x_i - \bar{x})^2$, where $\bar{x}$ is the average of the data, is the total sum of squares (proportional to the variance of the data).

The value of R^2 spans from 1 when the data are exactly explained by the model ($\Sigma_{\text{res}} = 0$) to 0 in the case of a baseline model which always predicts $\bar{x}$. Negative values of R^2 are associated with models under the baseline.

The adjusted $\overline{R}^2$ refines this metric by counteracting the tendency of R^2 to increase when additional independent variables are introduced to the model, even if these variables insignificantly contribute to the explanation of the variance. The adjusted measure is obtained considering a normalization based on the number of independent variables of the model k and the number of observations n:

$$\overline{R^2} = 1 - (1 - R^2)\frac{n - 1}{n - k - 1}$$

This adjustment addresses overfitting, as an excessive number of extraneous variables leads to a decrease in $\overline{R}^2$. The resulting values still range from 0 to 1 and retain the same interpretation.

We perform this analysis in the three scenarios: $N = 1000$, $M = 500$, $N = 2000$, $M = 1000$, and $N = 5000$, $M = 1000$. Additionally, for the popularity distribution, we include a comparison with the R obtained from simulations of a standard urn model. The summarized results are presented in Table I.

The results reveal high values for the adjusted R^2 for the distributions of ratings and similarity. As for the popularity distribution, the values are slightly smaller but still within an acceptable range. Furthermore, it is noteworthy that our model outperforms the standard urn model in all cases, as evidenced by the higher $\overline{R}^2$.

[1] The partisan divide on political values grows even wider, https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/.

[2] U. Chitra and C. Musco, Analyzing the impact of filter bubbles on social network polarization, in *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20 (Association for Computing Machinery, New York, 2020), pp. 115–123.

[3] M. Maes and L. Bischofberger, Will the personalization of online social networks foster opinion polarization? Available at SSRN (2015), doi: 10.2139/ssrn.2553436.

[4] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, Exploring the filter bubble: The effect of using recommender systems on content diversity, in *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14 (Association for Computing Machinery, New York, 2014), pp. 677–686.

[5] L. V. Bryant, The YouTube algorithm and the alt-right filter bubble, Open Info. Sci. **4**, 85 (2020).

[6] D. O'Callaghan, D. Greene, M. Conway, J. Carthy, and P. Cunningham, The extreme right filter bubble, arXiv:1308.6149.

[7] D. Muise, H. Hosseinmardi, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, Quantifying partisan news diets in Web and TV audiences, Sci. Adv. **8**, eabn0083 (2022).

[8] G. Linden, B. Smith, and J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Comput. **7**, 76 (2003).

[9] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, Recommender systems, Phys. Rep. **519**, 1 (2012).

[10] B. Smith and G. Linden, Two decades of recommender systems at amazon.com, IEEE Internet Comput. **21**, 12 (2017).

[11] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You* (Penguin UK, New York, 2011).

[12] T. R. Dillahunt, C. A. Brooks, and S. Gulati, Detecting and visualizing filter bubbles in Google and Bing, in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15 (Association for Computing Machinery, New York, 2015), pp. 1851–1856.

[13] S. Nagulendra and J. Vassileva, Understanding and controlling the filter bubble through interactive visualization: A user study, in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14 (Association for Computing Machinery, New York, 2014), pp. 107–115.

[14] P. Gravino, B. Monechi, and V. Loreto, Towards novelty-driven recommender systems, C. R. Phys. **20**, 371 (2019).

[15] B. Kirdemir and N. Agarwal, Exploring bias and information bubbles in YouTube's video recommendation networks, in *Complex Networks & Their Applications X. COMPLEX NETWORKS 2021*, Studies in Computational Intelligence, edited by R. M. Benito, C. Cherifi, H, Cherifi, E. Moro, L. M. Rocha, and M. Sales-Pardo, Vol 1073 (Springer, Cham, 2022).

[16] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, The echo chamber effect on social media, Proc. Natl. Acad. Sci. USA **118**, e2023301118 (2021).

[17] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, Quantifying echo chamber effects in information spreading over political communication networks, EPJ Data Sci. **8**, 35 (2019).

[18] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? Psych. Sci. **26**, 1531 (2015).

[19] J. L. Herlocker, J. A. Konstan, and J. Riedl, Explaining collaborative filtering recommendations, in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Association for Computing Machinery, New York, 2000), pp. 241–250.

[20] X. Su and T. M. Khoshgoftaar, A survey of collaborative filtering techniques, Adv. Artif. Intel. **2009**, 1 (2009).

[21] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. **81**, 591 (2009).

[22] P. Sen and B. K. Chakrabarti, *Sociophysics: An Introduction* (Oxford University Press, Oxford, 2014).

[23] G. L. Ciampaglia, A. Nematzadeh, F. Menczer, and A. Flammini, How algorithmic popularity bias hinders or promotes quality, Sci. Rep. **8**, 15951 (2018).

[24] A. Sirbu, D. Pedreschi, F. Giannotti, and J. Kertesz, Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model, PLoS One **14**, e0213246 (2019).

[25] F. Freitas, A. R. Vieira, and C. Anteneodo, Imperfect bifurcations in opinion dynamics under external fields, J. Stat. Mech. (2020) 024002.

[26] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Modeling echo chambers and polarization dynamics in social networks, Phys. Rev. Lett. **124**, 048301 (2020).

[27] J. W. Baron, Consensus, polarization, and coexistence in a continuous opinion dynamics model with quenched disorder, Phys. Rev. E **104**, 044309 (2021).

[28] G. De Marzo, A. Zaccaria, and C. Castellano, Emergence of polarization in a voter model with personalized information, Phys. Rev. Res. **2**, 043117 (2020).

[29] G. Iannelli, G. De Marzo, and C. Castellano, Filter bubble effect in the multistate voter model, Chaos **32**, 043103 (2022).

[30] N. Perra and L. E. C. Rocha, Modelling opinion dynamics in the age of algorithmic personalisation, Sci. Rep. **9**, 7261 (2019).

[31] A. F. Peralta, M. Neri, J. Kertész, and G. Iñiguez, Effect of algorithmic bias and network structure on coexistence, consensus, and polarization of opinions, Phys. Rev. E **104**, 044312 (2021).

[32] A. F. Peralta, J. Kertész, and G. Iñiguez, Opinion formation on social networks with algorithmic bias: Dynamics and bias imbalance, J. Phys.: Complexity **2**, 045009 (2021).

[33] F. P. Santos, Y. Lelkes, and S. A. Levin, Link recommendation algorithms and dynamics of polarization in online social networks, Proc. Natl. Acad. Sci. USA **118**, e2102141118 (2021).

[34] F. Cinus, M. Minici, C. Monti, and F. Bonchi, The effect of people recommenders on echo chambers and polarization, in *Proceedings of the International AAAI Conference on Web and Social Media*, edited by Y.-R. Lin, M. Cha, and D. Quercia, Vol. 16 (AAAI Press, Palo Alto, California, 2022), pp. 90–101.

[35] C. M. Valensise, M. Cinelli, and W. Quattrociocchi, The dynamics of online polarization, arXiv:2205.15958.

[36] A. Ferrara, L. Espín-Noboa, F. Karimi, and C. Wagner, Link recommendations: Their impact on network structure and minorities, in *14th ACM Web Science Conference 2022* (Association for Computing Machinery, New York, 2022), pp. 228–238.

[37] L. Espín-Noboa, C. Wagner, M. Strohmaier, and F. Karimi, Inequality and inequity in network-based ranking and recommendation algorithms, Sci. Rep. **12**, 1 (2022).

[38] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, Methods and metrics for cold-start recommendations, in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Association for Computing Machinery, New York, 2002), pp. 253–260.

[39] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, Grouplens: An open architecture for collaborative filtering of netnews, in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (Association for Computing Machinery, New York, 1994), pp. 175–186.

[40] G. Pólya, Sur quelques points de la théorie des probabilités, Ann. Inst. Henri Poincaré **1**, 117 (1930).

[41] J. G. Mauldon, A generalization of the beta-distribution, Ann. Math. Stat. **30**, 509 (1959).

[42] I. Konstas, V. Stathopoulos, and J. M. Jose, On social networks and collaborative recommendation, in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Association for Computing Machinery, New York, 2009), pp. 195–202.

[43] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz, The dynamics of correlated novelties, Sci. Rep. **4**, 5890 (2014).

[44] D. Kowald, M. Schedl, and E. Lex, The unfairness of popularity bias in music recommendation: A reproducibility study, in *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42* (Springer, Lisbon, Portugal, 2020), pp. 35–42.

[45] G. Vigliensoni and I. Fujinaga, The music listening histories dataset, in *Proceedings of the 18th International Society for Music Information Retrieval Conference* (Suzhou, People's Republic of China, 2017), pp. 96–102.

[46] A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä, Cumulative effects of triadic closure and homophily in social networks, Sci. Adv. **6**, eaax7310 (2020).

[47] M. Hohmann, K. Devriendt, and M. Coscia, Quantifying ideological polarization on a network using generalized Euclidean distance, Sci. Adv. **9**, eabq2044 (2023).