Chaos

Polarizing crowds: Consensus and bipolarization in a persuasive arguments model

Cite as: Chaos **30**, 063141 (2020); https://doi.org/10.1063/5.0004504
Submitted: 12 February 2020 . Accepted: 05 June 2020 . Published Online: 19 June 2020

Federico Barrera Lemarchand, Viktoriya Semeshenko, Joaquín Navajas, and Pablo Balenzuela

Polarizing crowds: Consensus and bipolarization in a persuasive arguments model

Cite as: Chaos **30**, 063141 (2020); doi: 10.1063/5.0004504
Submitted: 12 February 2020 · Accepted: 5 June 2020 ·
Published Online: 19 June 2020

Federico Barrera Lemarchand,[1,2,3,a)] Viktoriya Semeshenko,[4] Joaquín Navajas,[1,3] and Pablo Balenzuela[2,5]

AFFILIATIONS

[1]Laboratorio de Neurociencia, Escuela de Negocios, Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, 1428 Buenos Aires, Argentina
[2]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina
[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290 (C1425FQB), Buenos Aires, Argentina
[4]Facultad de Ciencias Económicas, Universidad de Buenos Aires, Buenos Aires, Argentina and Instituto Interdisciplinario de Economía Política de Buenos Aires, CONICET–Universidad de Buenos Aires, Av. Córdoba 2122, C1120 AAQ Buenos Aires, Argentina
[5]Instituto de Física de Buenos Aires (IFIBA), CONICET, Av. Cantilo s/n, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina

Note: This article is part of the Focus Issue, Dynamics of Social Systems.
[a)]**Author to whom correspondence should be addressed:** fedex192@gmail.com

ABSTRACT

Understanding the opinion formation dynamics in social systems is of vast relevance in diverse aspects of society. In particular, it is relevant for political deliberation and other group decision-making processes. Although previous research has reported different approaches to model social dynamics, most of them focused on interaction mechanisms where individuals modify their opinions in line with the opinions of others, without invoking a latent mechanism of argumentation. In this paper, we present a model where changes of opinion are due to explicit exchanges of arguments, and we analyze the emerging collective states in terms of simple dynamic rules. We find that, when interactions are equiprobable and symmetrical, the model only shows consensus solutions. However, when either homophily, confirmation bias, or both are included, we observe the emergence and dominance of bipolarization, which appears due to the fact that individuals are not able to accept the contrary information from their opponents during exchanges of arguments. In all cases, the predominance of each stable state depends on the relation between the number of agents and the number of available arguments in the discussion. Overall, this paper describes the dynamics and shows the conditions wherein deliberative agents are expected to construct polarized societies.

In democratic societies, trivial and important decisions emerge from group discussions where people interchange opinions and try to persuade others using their own ideas. With two alternative choices, this situation eventually leads to consensus or coexistence of opposite opinions. In this paper, we explore the role of a social influence mechanism based on opinion change using exchanges of arguments, on the emergence of collective states in groups of agents, of different sizes. The results show the importance of accepting contrary evidence in order to reach consensus and the role of confirmation bias and homophily to avoid it. The framework developed in this paper can provide interesting and helpful insights for future experiments on opinion formation dynamics and explain the macroscopic collective states observed in societies.

I. INTRODUCTION

People engage in different types of debates and discussions on a daily basis. Very often, arguments are invoked and exchanged in these interactions, which can eventually lead to modifying initial opinions and decisions to be made. Political debates (e.g., the presidential debates), referendums (e.g., the "Brexit" referendum of 2016), discussion groups (e.g., medical panels, juries,

etc.), or even work meetings are all examples of discussions, at vastly different scales, where important decisions have to be made. These situations guided the development of several experimental and theoretical studies focused on fully understanding opinion dynamics. From a theoretical standpoint, statistical physics provides very powerful quantitative tools, useful for studying complex systems comprising many interacting agents (as can be seen in various comprehensive recent reviews).[1–4] In this frame of reference, agent-based modeling is a widely used technique employed to derive macroscopic states from simple microscopic interactions between *agents*. It has been efficiently applied to study various phenomena and environments, such as pedestrian behavior and traffic,[2] agriculture,[5] economy,[6] socio-ecological systems,[7] demography,[8] among many others.[2,9,10]

Past efforts in agent-based modeling of opinion dynamics typically involved the use of discrete opinions, as, for instance, *Voter* models,[11–16] Sznajd models,[17–23] or *majority rules* models, which use different topologies of interactions and rules for opinion dynamics.[24–28]

The use of continuous opinions gave place to a different stream of models, starting from the original DeGroot model.[29] Typically, the coexistence of different macroscopic opinions in this type of models was introduced by the concept of *bounded confidence*,[30] which involved interacting agents ignoring each other's opinions only when they were extremely different. These models were extensively studied in recent years.[31–36]

There were also some contributions that combine both approaches (continuous and discrete opinions). A model for opinions observed as discrete actions but represented internally as a continuous opinion was presented in Ref. 20. A similar approach was presented in Ref. 37 where discrete external opinions are emergent from underlying internal postures. The dynamics of this model mimic a process of information gathering: an agent's posture may change after interaction with another agent, and this in turn can induce an opinion change, if his posture crosses a specific threshold. Thus, the external opinion somehow determines a specific binary decision, and the internal posture accounts for how convinced an agent is or not about the manifested opinion. Even though other models based on accumulation of information were previously developed,[38] they do not involve the combination between discrete external opinions and continuous underlying postures mediated by thresholds.

Opinion dynamics yields a variety of naturally occurring macroscopic states in our society.[36] Given their political, economic, and social relevance, these states have been widely studied.[1,39–45] *Consensus* (all members of the group adhere to the same opinion) and *Bipolarization* (two distinct, polarized groups are formed) are relevant examples, commonly found in topics with binary statements (a pro-against issue, for example). Although bipolarization can be induced by introducing a mechanism of negative influence (purposeful distancing from dissimilar agents), insufficient and controversial evidence supporting their existence[46] has led to the proposal of other mechanisms able to reproduce this collective state.[47] Here, authors proposed a model to find bipolarization in the absence of negative influence by introducing an explicit exchange of arguments, along with homophily (increased probability of interaction with similar rather than dissimilar agents). They based their assumptions on the Persuasive Arguments Theory (PAT), which states that changes in opinion are derived from exchanges of arguments and depend on the number and strength (or persuasiveness) of these arguments.[48–53]

In this work, we present an agent-based threshold model where agent's states are represented by an external discrete opinion with continuous underlying internal postures,[37] but the interactions are given by an explicit exchange of arguments,[47] based on PAT.[48,49,51,53] We develop a set of rules for this exchange and explore the new parameter space, introducing also modifications aimed at modeling certain cognitive biases previously reported in the literature, which have been linked to the macroscopic states observed in society.[54]

This paper is organized as follows: in Sec. II, we describe the implementation of our model and the theoretical background it is grounded on. In Sec. III, we present our main results regarding the collective states produced by the model in the unbiased version (Subsection III A) and in the biased version with confirmation bias (Subsection III B) and homophily (Subsection III C). Finally, in Sec. IV, we outline the main conclusions and discuss possible future modifications for our model.

II. THE MODEL

We consider a population of N individuals whose opinion O emerges from an internal posture P, which is supported by a set of arguments. Individuals engage in pairwise interactions in which they exchange arguments, which eventually can lead to a change in their opinions. The state of the agents and the dynamics of interactions are detailed below.

A. States of the agents

The state of each agent is represented by two non-independent variables[37] and a set of arguments that define these variables, and it is sketched in panel (a) of Fig. 1. The discrete variable O accounts for the public opinion and takes three possible values: $O = 0, \pm 1$. The continuous variable P stands for the agent's posture and takes any value in the range $[-P_{\max}, P_{\max}]$, while the threshold P_t determines an agent's opinion as a function of P. If an agent's posture is in the range $[-P_{\max}, -P_t)$, the opinion is $O = -1$; if an agent's posture lies between $[-P_t, P_t]$, the opinion is $O = 0$; and within the range $[P_t, P_{\max})$, the opinion is $O = 1$. We set $P_{\max} = 3$ and $P_t = 1$, which ensures that the intervals of posture generating each opinion are equal in size. We say an agent is *oriented* if its opinion is $O = \pm 1$ and *moderate* if $O = 0$.

In agreement with the persuasive argument theory, agent's postures P are derived from the arguments they possess in their finite memory. Let M be the number of arguments each agent can recall and N_A be the number of existing arguments available to agents. We assume that half of the N_A arguments will be positive (i.e., in favor of the issue), the other half will be negative (i.e., against the issue). Also, each argument A will have a specific weight w, with possible integer values $1, \ldots, N_A/2$ and sign $V = \pm 1$. In this way, for every argument A_j of a given strength w_j and sign V_j, there exists an opposite argument of strength w_j and sign $-V_j$. The posture of an agent

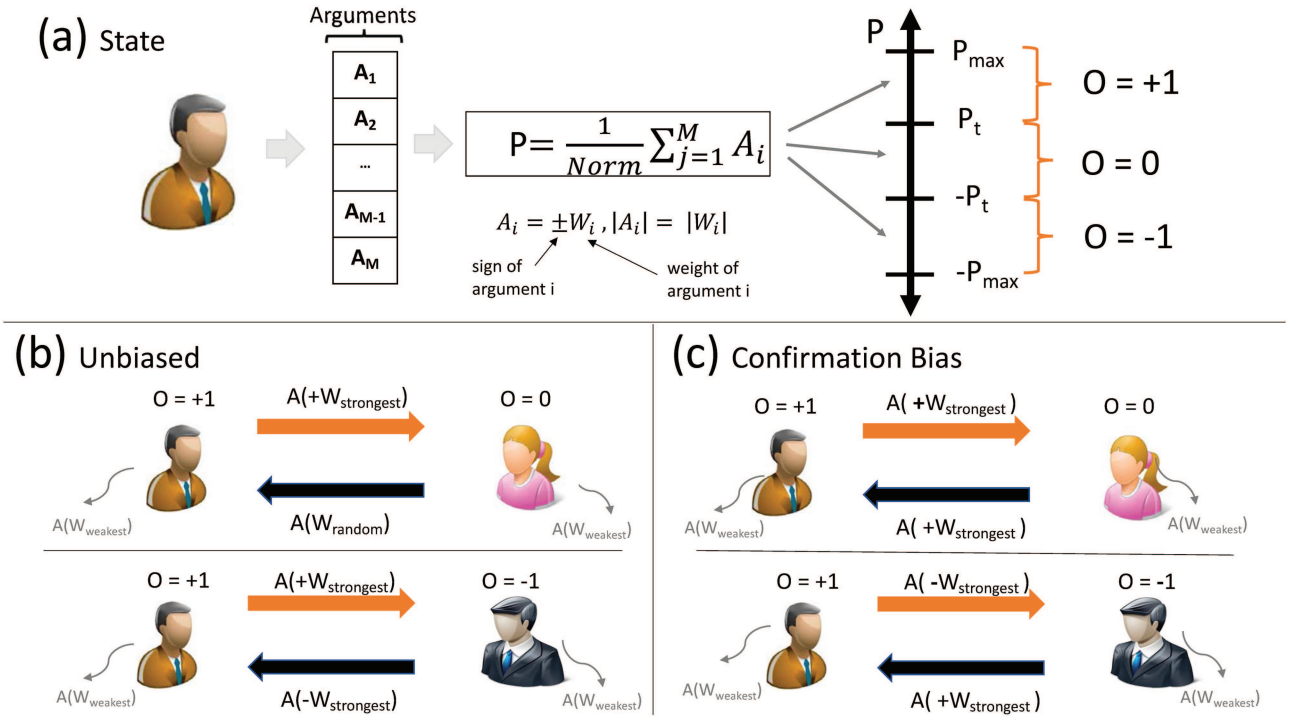

FIG. 1. State of the agents and interaction dynamics. (a) State of the agents. Each agent possesses a set of M arguments, which determine his posture P according to Eq. (1). If $-P_{max} \leq P < -P_t$ or $P_t < P \leq P_{max}$, the individual will hold an extreme opinion ($O = -1$ and $O = +1$, respectively). If $-P_t \leq P \leq P_t$, agents will hold a moderate opinion (O) about the discussed topic. (b) Unbiased interactions. In each bidirectional interaction, oriented agents will always share their most important argument (i.e., the strongest one), meanwhile, moderate agents will share a random one. (c) Biased interactions. In each interaction, oriented agents will incorporate the strongest argument from their opponent, aligned with their own orientation (i.e., a positive one if he has $O = +1$).

with M arguments in memory is defined as

$$P = \frac{1}{N_A/2}\sum_{j=1}^{M} A_j = \frac{1}{N_A/2}\sum_{j=1}^{M} V_j.w_j, \tag{1}$$

where the argument A_j has sign (V_j) and weight (w_j).

B. Dynamics

In this work, we consider only pairwise bidirectional interactions between agents. These interactions are mediated by the exchange of arguments, which lead to changes in agent's posture [according to Eq. (1)] and eventually may cause change in opinions (if posture's changes cross the thresholds P_t). In each interaction, each agent will share an argument with his opponent and receive another one from him. In this exchange, both will discard the weakest one.

In what follows, we consider two models of interactions: unbiased and mediated by confirmation bias, as detailed below and sketched in panels (b) and (c) of Fig. 1.

1. Unbiased interactions

The basic idea here is that agents with extreme opinions ($O = \pm 1$) will share the strongest argument aligned with their own opinion, meanwhile, a moderate agent ($O = 0$) will share a random one. The detailed rules read as follows:

1. An oriented agent j will always share an argument of the same sign V as its opinion O and will try to exchange the strongest argument (e.g., if $O_j = +1$, the given argument will be the strongest one with sign $+1$). If the other agent already has that argument, then j will exchange the second strongest and so forth. This rule is independent of the other agent's opinion.
2. A moderate agent will always share a randomly selected argument, regardless of its sign. As before, if the other agent already has the mentioned argument, another one will be chosen randomly regardless of the other agent's opinion.

Here, oriented agents with opposite opinions behave in the same way as if they were interacting with agents holding their same opinion. However, and contrary to this rule, a vast amount of empirical evidence suggests that people do behave in a biased way when exposed to denying arguments. More specifically, humans are prone to confirmation biases,[55–57] wherein individuals overtly seek information that favors preexisting beliefs, expectations, and hypotheses.[58]

This phenomenon involves discounting, or even ignoring, opposing evidence as well as focusing disproportionately on favorable information. Confirmation biases have been reported in a variety of relevant situations (e.g., crime investigations,[59,60] medical deliberations,[55,61] scientific inquiry and hypothesis testing,[56,62,63] diffusion of fake news, rumors, or conspiracy theories,[64] and even in the prevalence of social stereotypes and consequential discrimination[57]), and there have been recent efforts in modeling its impact on collective states.[64–66]

2. Confirmation bias interaction

In order to simulate the confirmation bias, we slightly modified the interaction rule between agents with opposing opinions: each time agents interact, each will get a favorable argument from the other one, ignoring the undesirable arguments. In other words, agent i with $O_i = +1$ will only accept the strongest argument of sign $+1$ from agent j with $O_j = -1$. In turn, agent j will receive the strongest argument of sign -1 from agent i. All other rules remain unchanged.

3. Homophily interaction

Based on an extensive literature suggesting that similarity breeds interpersonal attraction,[67] we also explored the role of homophily in the pairwise interactions between agents. We define the similarity[47] between agents i and j as $S_{i,j} = 1 - |P_i - P_j|/(2P_{\max})$, $\forall j$. This definition ensures that agents with the same posture will have $S_{i,j} = 1$, and agents with maximally opposing postures ($P_i = -P_j$, and $|P_i| = P_{\max}$) will have $S_{i,j} = 0$. Homophily suggests that the more similar agents are, the greater the chance of interaction they have. This interaction probability is ruled by

$$Q_{i,j} = \frac{(S_{i,j})^h}{\sum_{p=1,p\neq i}(S_{i,p})^h}, \tag{2}$$

where h is a free parameter depicting the strength of the homophily.

Summarizing, when two agents interact in this condition, first, agent i is selected with uniform probability, and then agent j with probability $Q_{i,j}$ according to Eq. (2).

III. RESULTS

We consider a population of N agents without any underlying connectivity topology. At $t = 0$, each agent is provided with different M arguments picked at random from the full list of N_A available arguments. At each time step $\Delta t = 1/N$, two agents are chosen at random and put to interact following the rules described above (one time step corresponds to N random interactions between agents). The system evolves until a stationary state is reached.

A. Unbiased model

We explored the final collective states reached by the system as a function of three free parameters of the model, namely, the number of agents (N), the number of existing arguments (N_A), and the memory size (M). We start with a set of $M = 6$ fixed arguments. Figure 2(a) shows the phase diagram with space regions where each type of solution predominates (i.e., is most probable). In order to define the regions, we estimated the probability of observing a given state s as N_s/N_{tot}, where N_s is the number of realizations that converged to that state and $N_{tot} = 1000$ is the total number of simulations ran in that point of the parameter space.

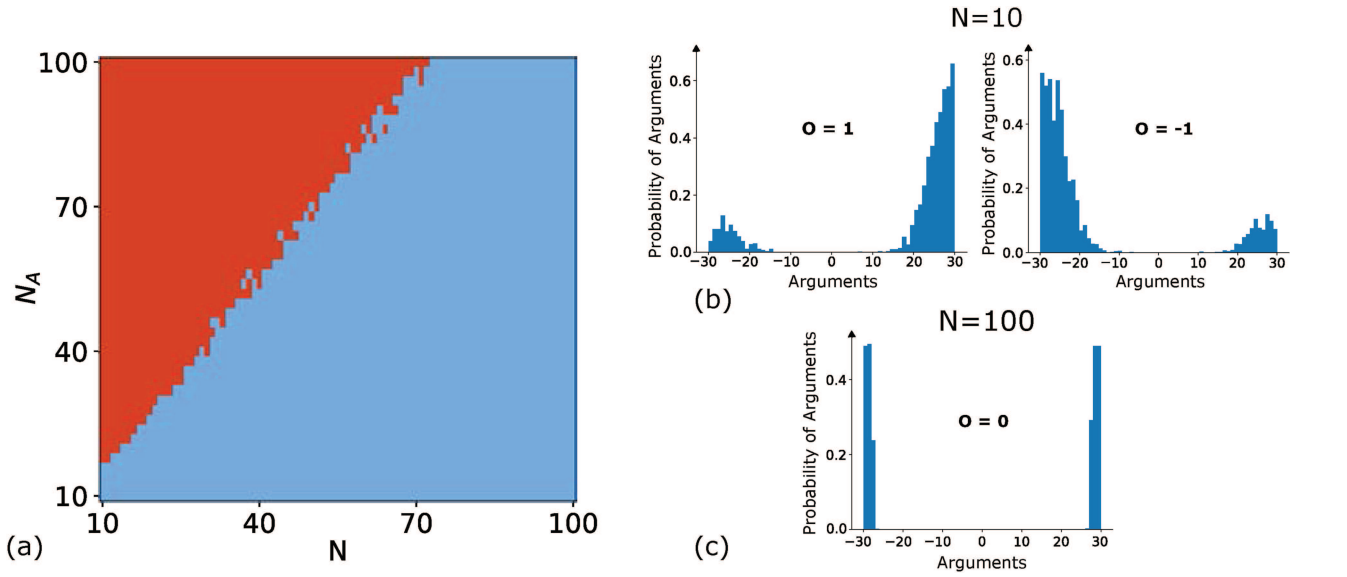

FIG. 2. Dominant solution as a function of N and N_A for the unbiased model. (a) Phase diagram of the space parameter where a given state is more likely than the other. The region where moderate consensus predominates (lower right) is depicted in light blue and the region where oriented consensus predominates (upper left) is shown in red. (b) Final distribution of arguments in oriented consensus for $N = 10$ and $N_A = 60$. (c) Final distribution of arguments in moderate consensus for $N = 100$ and $N_A = 60$.

We found two well separated regions: one with moderate consensus (light blue, $O = 0$) and a smaller region with oriented consensus (red, $O = \pm 1$).

On one hand, the oriented consensus is a dominant solution (although not the only one) when there are roughly more arguments than agents [the upper-left region in panel (a) of Fig. 2]. In this region, the probability that all arguments were represented in the initial state of the system reduces, and therefore, particular cases appear in which arguments of a given sign outweigh arguments of the opposite sign, generating final states of oriented consensus.

On the other hand, in the region where there are less arguments than agents, we observe that moderate consensus is comprised of agents holding strong arguments of both signs, as shown in panel (c) of Fig. 2. This is because the probability that all arguments are

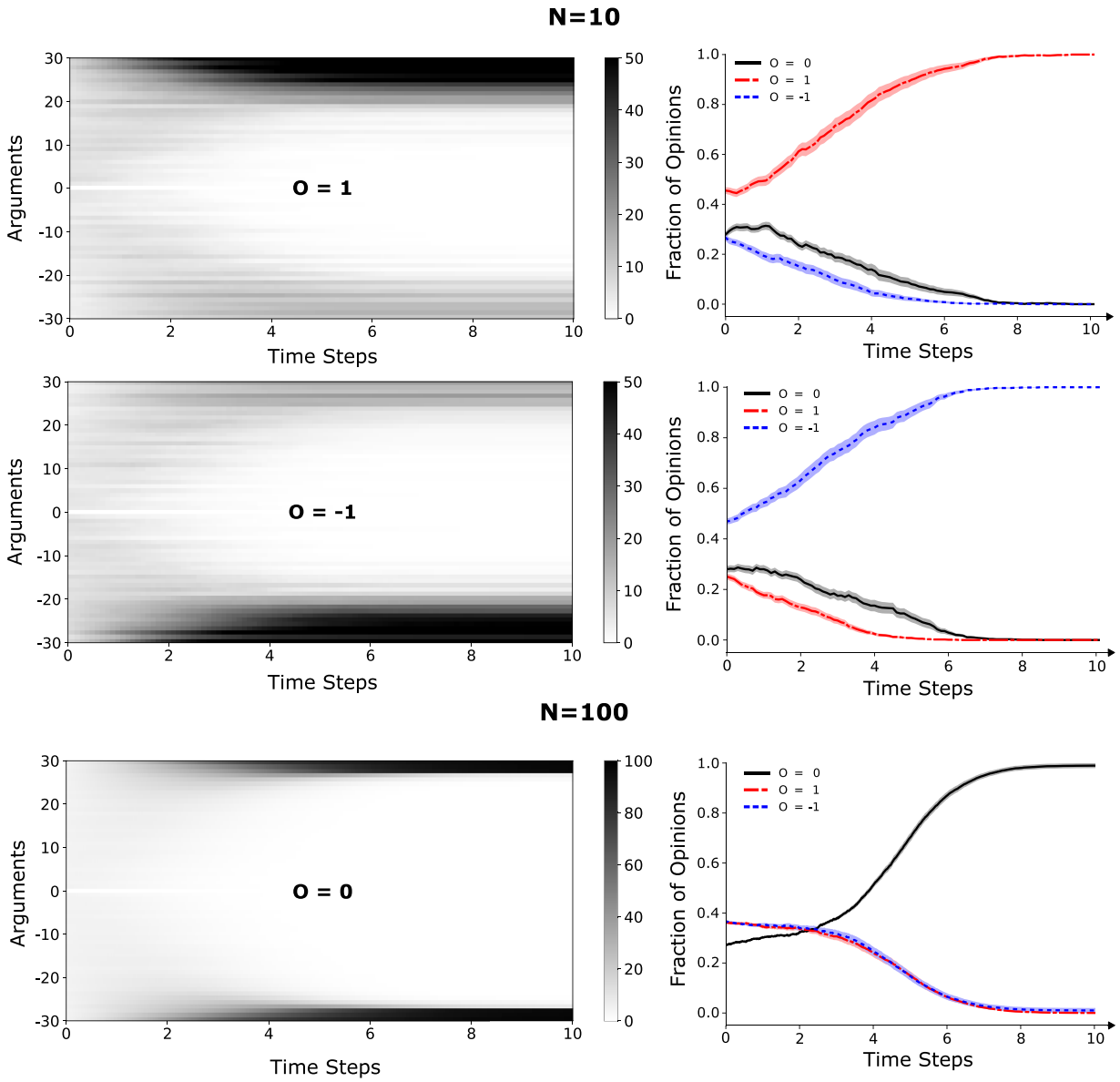

FIG. 3. Dynamics of the unbiased model. Left column: histograms exhibiting the evolution of arguments for different final states of the system. We saturated the scale for a better appreciation of the initial arguments. Right column: evolution of the fraction of agents of each opinion state. Each of the six panels was constructed based on 100 random realizations of initial conditions ($N_{tot} = 100$), with parameters $N_A = 60$ arguments and $M = 6$ memory size. The shaded regions denote the respective 95% confidence intervals. The first two rows correspond to the oriented consensus region (the red region in Fig. 2, $N = 10$ agents), and the third row corresponds to the moderate consensus region (the light blue region in Fig. 2, $N = 100$). One time step corresponds to N interactions.

present at the beginning of the simulations is very high, which leads to $M/2$ strongest arguments of each sign surviving and inevitably reaching a moderate consensus state. Both behaviors can be seen in the time evolution of the arguments and opinions, as plotted in Fig. 3. Specifically, the evolution of arguments in the moderate region ($N = 100$, third row) clearly shows how all arguments are initially present, and only the strongest ones of each sign remain in the final state. Furthermore, in the oriented region ($N = 10$, first two rows), not all arguments are initially present (as evidenced by the heterogeneity of initial states), and while some strong arguments can remain for the opposite opinion, the dominant one has stronger arguments in average. In the same figure, the evolution of the dynamics of opinions can be observed for the same regions.

Given the behavior observed in Fig. 2(a), where the dominance of each solution depends roughly on the ratio N_A/N, in Fig. 4, we plot the dominant collective state as a function of M and N_A/N. We observe here the same solutions as in the previous plot: oriented consensus (red) and moderate consensus (light blue). We can see that the oriented consensus states dominate the collective behavior for large values of N_A/N (which means there are many more arguments than agents), especially if the memory size is not too big. This can be explained in a similar way: the memory size is related to the number of arguments available in the initial conditions. When the number of agents is small with respect to the number of arguments, for reduced memory size, many arguments will be missing at the beginning of the simulation, which will bring the system to the oriented consensus. However, when the memory size increases, and/or the number of arguments decreases with respect to the number of agents (low values of N_A/N), it becomes more likely that all existing arguments are present, therefore leading to moderate consensus. Only for large values of N_A/N, and small values of memory, it becomes possible to find oriented consensus.

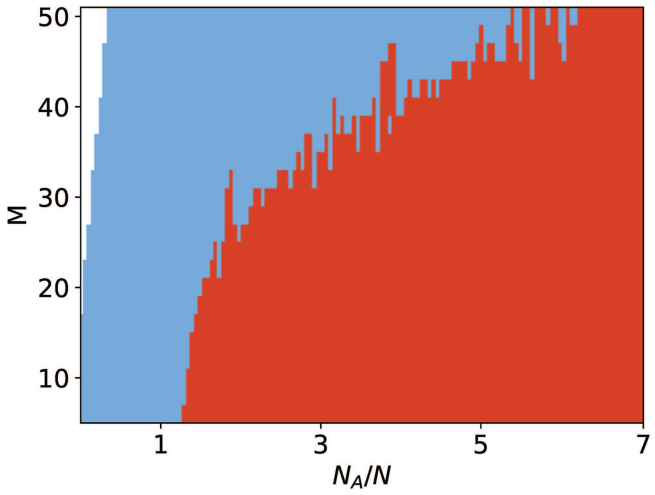

FIG. 4. Dominant solution as a function of M and N_A/N for the unbiased model. Light blue: Region where moderate consensus is more probable (upper left). Red: Region where oriented consensus predominates (lower right). The white region corresponds to the cases where $M \geq N_A$, where no arguments are forgotten (which is not physically relevant). We use 1000 random realizations of initial conditions for each pair of parameters ($N_{tot} = 1000$). We swept N and N_A from 10 to 100, the former in steps of 1 and the latter in steps of 2. One value of N_A/N comes from the average of all possible combinations of N and N_A that give that specific value.

B. Biased model

Here, we study the role played by *confirmation bias*. We repeated our parameter exploration, focusing once again on the relationship between N and N_A, for fixed $M = 6$ [Fig. 5(a)]. We found that solutions with oriented consensus have now vanished from the parameter space and have been replaced by states of "bipolarization," where two oriented opinions coexist in the population. This solution has now become prevalent in the parameter space, being the most likely state in all regions previously dominated by oriented states and also reducing the size of the region of moderate consensus. As in the unbiased model, moderate consensus was mostly found for $N > N_A$, which is also explained by the presence of all existing arguments in the initial state of the system.

The reason underlying the emergence of bipolarization is related to the new interaction rule proposed in the model, to account for the confirmation bias. Now, when two agents with opposite opinions interact, they exchange arguments that reinforce their own positions. This effectively acts as a repulsion force, which drives oriented agents toward extreme postures. One should note that bipolarization is not necessarily balanced: there can be more agents of one opinion than the other. This can be observed in Fig. 5(b). For smaller N, it is possible to find more agents of one opinion than the other (lighter green to yellow color). For larger N, bipolarization tends to become balanced (darker green color).

The exploration of parameter M yields similar results, only this time bipolarization takes the role previously occupied by oriented consensus, which in turn completely vanishes for medium ranged values of M. It is omitted to avoid redundancy.

The results obtained for both biased and unbiased models led us to explore the transition between these two cases. To this purpose, we define a new parameter: the probability of confirmation bias, P_{CB}. Every time two oriented agents of opposing opinions interact, P_{CB} determines the probability of doing it with confirmation bias. It is important to note that the unbiased model is equivalent to $P_{CB} = 0$, and the biased model is equivalent to $P_{CB} = 1$. This leads to expect the final states of the system to be more similar to the unbiased model for values close to 0 and more similar to the biased model for values close to 1. However, the result was unexpected: there were no bipolarization states for any value $P_{CB} < 1$ [see Fig. 6(a)]. Furthermore, moderate and oriented consensus remained mostly unaltered.

To further study this phenomenon, we defined a new magnitude: the contrary information flux, φ, which measures the accumulated weights of opposite opinion arguments acquired by an oriented agent. This may occur either through interactions with a moderate agent (from whom he takes a randomly chosen argument) or through interactions with another oriented agent, but with an opposing opinion.

For each interaction between oriented agents with opposing opinions, we add the weights of the arguments that have been

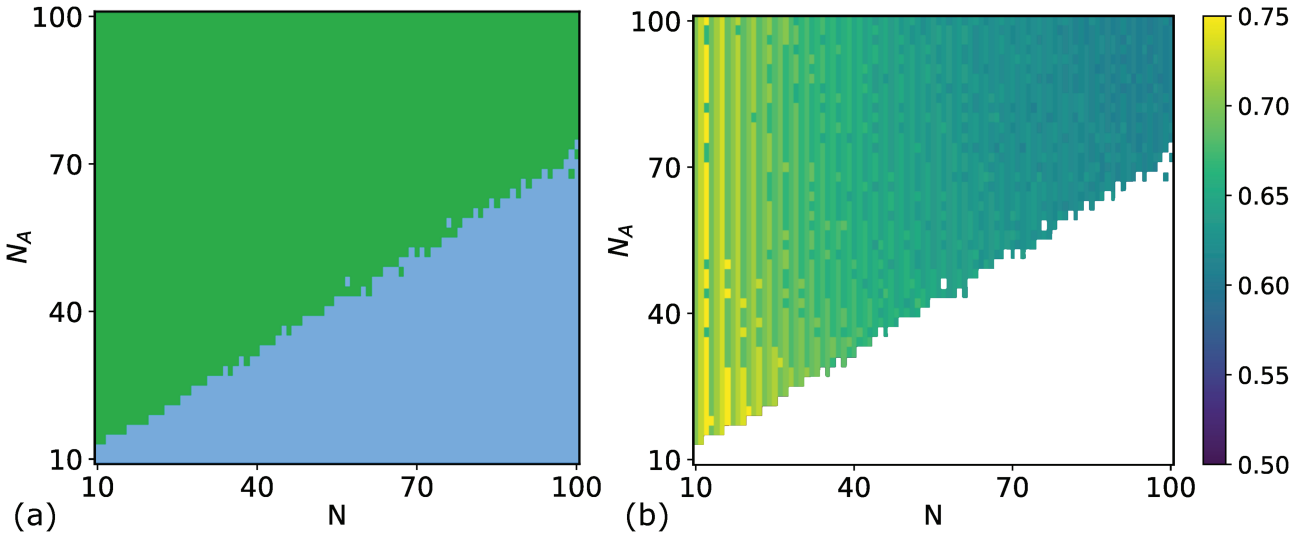

FIG. 5. Dominant solution as a function of N and N_A for the confirmation bias model. (a) Regions of the space parameter where a given state is more probable than the others. (Light blue) Region where moderate consensus is more probable (lower right). (Green) Region where bipolarization dominates (upper left). (b) Mean number of agents with a dominant oriented opinion, normalized by N. In both panels, each point is an average over 1000 realizations of the random initial conditions.

exchanged. This way, we obtain the contrary information flux for each interaction, φ_{int}. Summing over all interactions gives the total contrary information flux, φ [Eq. (3)]. The results of this information flux are plotted in Fig. 6(b),

$$\varphi = \sum_{int} \varphi_{int} = \sum_{int} (w_{i,int} + w_{j,int}). \tag{3}$$

Figure 6 shows that bipolarization can only be found for $P_{CB} = 1$, and this in turn is related to $\varphi = 0$. This demonstrates that bipolarization is reached due to the lack of contrary information acquired through the interaction between agents. For every other value of P_{CB}, $\varphi \neq 0$, which means that eventually, oriented agents get contrary arguments, thus becoming moderated.

C. Homophily

Previous research has shown that the combination of homophily and the exchange of arguments may produce states

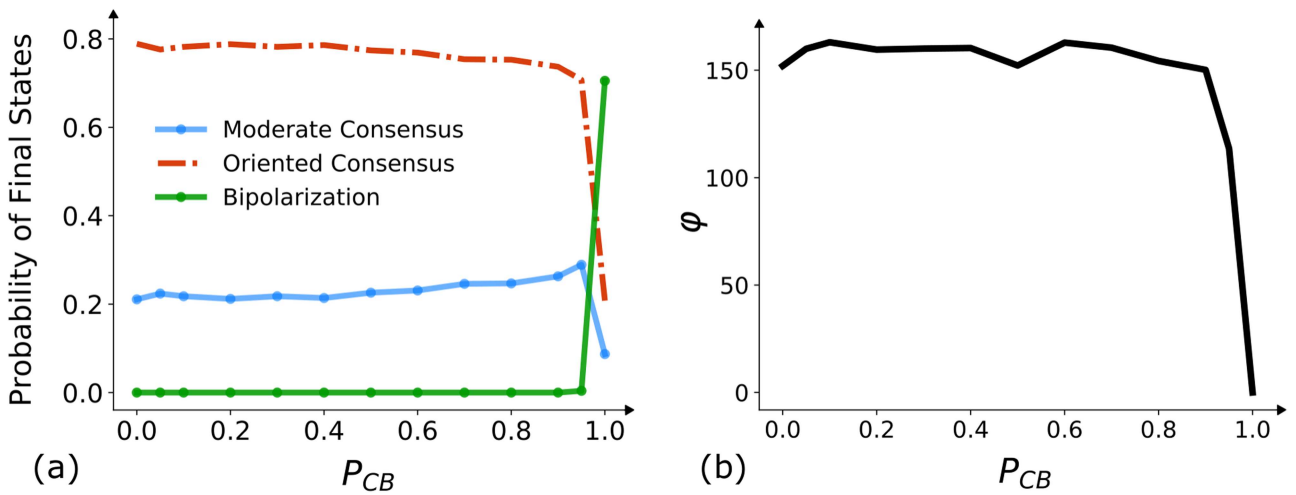

FIG. 6. Transition between unbiased and biased models. (a) Probability of each final state vs P_{CB}. (b) φ vs P_{CB}. Numerical values used: $N_{tot} = 1000$, $N = 10$, $N_A = 60$, and $M = 6$.

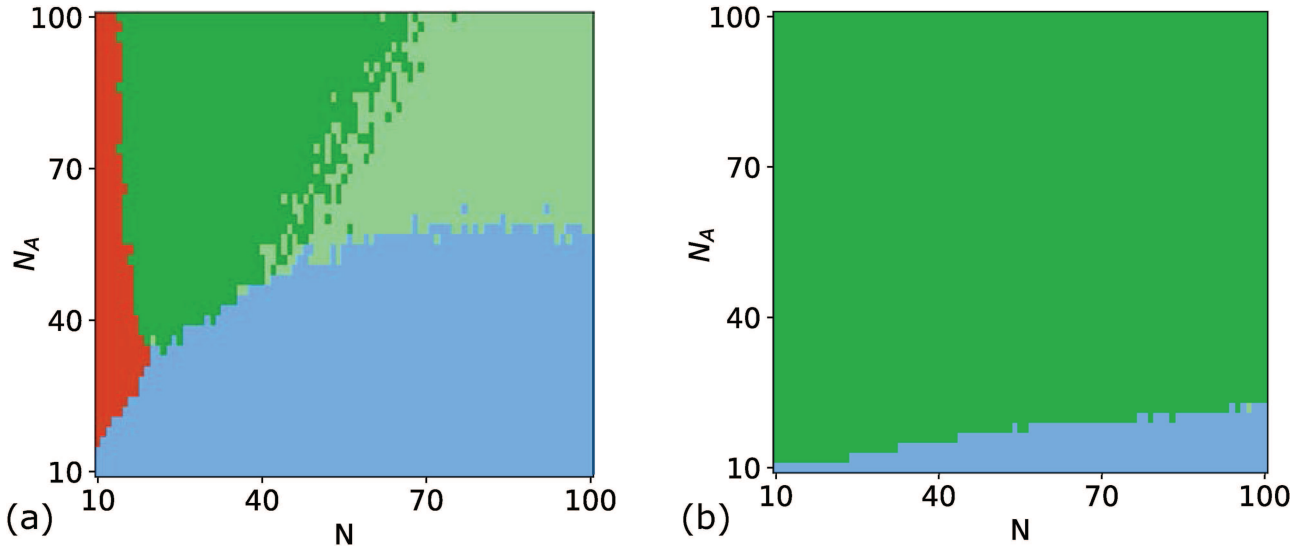

FIG. 7. Effects of homophily in biased and unbiased models. Regions of the space parameter where a given state is more probable than the other, for the unbiased model (a) and the biased model (b). In both cases, $h = 4$, and 1000 random realizations of initial conditions were used. (Light blue) Region where moderate consensus is more probable than the others (lower region). (Red) Region where oriented consensus predominates (left region). (Green) Region where bipolarization dominates. (Light green) Region where metastable states of bipolarization are the most probable states.

of bipolarization.[47] To explore the role played by homophily in polarization, we enriched our model by including its dynamics. Figure 7 presents the results for both the unbiased model with homophily (a) and the biased model with homophily (b), according to Eq. (2), for $h = 4$. As before, dominance of moderate consensus is depicted in light blue, oriented consensus in red, and bipolarization in green, and we added a light green region that corresponds to preponderance of metastable states of bipolarization. The latter could actually lead to final states of moderate consensus, oriented consensus, or bipolarization. These metastable states are barely present for low values of homophily; however, for higher values ($h > 3$), they become predominant in a region of space. They can only be found when confirmation bias is not present. The number of time steps required for them to lose stability and reach a final stable state is much larger than the mean number required by the other states (more than two standard deviations larger).

One important conclusion drawn from Fig. 7 is that bipolarization can only arise without confirmation bias in the presence of homophily. For different values of h, it is observed that the bipolarization region gets larger with h (the plots for other values of h are omitted to avoid redundancy). Oriented consensus is nearly replaced by bipolarization. Furthermore, when both confirmation bias and homophily are present [Fig. 7(b)], bipolarization dominates nearly the entire phase diagram, safe for a small region of moderate consensus. This suggests that homophily and confirmation bias reinforce each other, producing even more bipolarization than before.

The reason behind the bipolarization states produced by homophily is that agents who have extreme opposite values of posture (i.e., some with $P = -3$ and others with $P = 3$) cannot interact [see Eq. (2)]. If the probability of interaction between agents with similar posture is much greater than that between dissimilar ones, then interactions tend to produce groups of oriented agents on the opposite sides of the spectrum. After a few successive in-group interactions, all agents end up with all the strongest arguments of their own side and thus with the most extreme values of posture. The joint combination of homophily and confirmation bias makes these effects stronger.

With all this in mind, it can be concluded that both homophily and confirmation bias produce bipolarization states in our model. We can see that a strong homophily plays the same role as confirmation bias: to obstruct the flux of contrary information between groups with opposite opinions. When confirmation bias is present, there is no flux due to the lack of interactions between agents with opposite opinions; meanwhile, when homophily dominates, it is due to a reinforcement process in the interactions, as was shown above.

IV. CONCLUSIONS

In this paper, we present a model where the change of opinion is mediated by the exchange of arguments, and we analyze the collective states in terms of simple dynamic rules. A possible interpretation of our rules of interaction is that two agents engage in a discussion and listen to the arguments the other has to offer. In an unbiased discussion, an agent of a given oriented opinion will tend to convince any other one of its stance, which leads the other agent to incorporate an argument of that opinion (the strongest one, in this case). However, if the discussion is biased, when two agents of opposing opinions interact, they will only listen to what favors the opinions they already have and ignore opposing arguments. Any

"concession" will be exacerbated by the other agent, and thus, agents might end up even more convinced of their previous stance after discussing with someone of opposite stance. This phenomenon, known as the "boomerang effect," where a person attempts to persuade others and unintentionally reinforces their prior beliefs, has been previously observed in empirical research.[68] In contrast, oriented agents can become more moderate when interacting with moderate agents, but they could also become more oriented (and this remains unchanged in both variants of the model). Again, this effect, where opinions become more extreme after interacting with other individuals, has been extensively observed in psychological research.[69]

Our unbiased model was successful in reproducing consensus states, either oriented consensus ($O = 1$ or $O = -1$ for all agents) or moderate consensus ($O = 0$). The region in the parameter space where the latter predominates is larger and mostly confined to the region where the number of arguments is smaller than the number of agents ($N_A < N$), which means that there is a high probability that all arguments are present at the beginning of the simulations. While changes in memory size (M) are comparably less important, the increase in the number of arguments leads to an increase in oriented consensus states. The parameter P_t, which is the threshold between opinions, does not play a relevant role as long as the three opinions are mapped to posture intervals of the same length.[37] However, a deeper analysis involving a distribution of thresholds instead of a fixed value could play a relevant role, but we leave this open for future research.

The moderate consensus state deserves a careful analysis. This is because in this model, *a priori*, agents could have a moderate opinion for two reasons: (1) because they possess weak arguments or (2) because they possess strong arguments from opposite sides so that both types of arguments cancel out [according to Eq. (1)] and their posture becomes close to zero. Given the dynamics of our model, it can only produce argument polarization, and therefore, we can only observe the second scenario where moderate agents are the result of agents holding strong arguments from both sides of the discussion.

One might argue that, assuming arguments can be objectively measured in strength, discussion without biases will invariably end in consensus (either moderate or oriented): if arguments are freely shared and accepted, eventually, the strongest arguments will be preserved, while the rest will be forgotten, and everyone will have those same arguments. However, the presence of a confirmation bias restricts the free flow of information: someone of a certain opinion will incorporate favorable arguments but will underestimate or ignore contrary information. This could in turn produce bipolarization, which led us to consider a variation of our model, with the inclusion of a confirmation bias.

When introducing the confirmation bias, bipolarization states appear. In contrast to the previous case, bipolarization replaces oriented consensus, and its region of preponderance becomes larger than that of the moderate consensus state. Nevertheless, after introducing a probability of confirmation bias (P_{CB}), we were unable to find a way to generate a smooth transition between our biased and unbiased models. Instead, we found that bipolarization was only present when $P_{CB} = 1$, which could be explained by the *contrary information flux* (φ). When $\varphi = 0$, no contrary information can be transmitted, and bipolarization becomes stable. In any other case, bipolarization ends up leading to consensus.

The introduction of homophily induced the appearance of bipolarization in the absence of confirmation bias. For high values of the homophily parameter (h), bipolarization nearly dominated all the regions where oriented consensus used to be the most probable state and some parts of the region where moderate consensus was preponderant. It is also possible to find a region where metastable states of bipolarization are dominant. Furthermore, when confirmation bias was also present, bipolarization nearly dominated all of the parameter space (with a reduced region of moderate consensus found for small values of N_A). Thus, superposition of both mechanisms produces even more bipolarization states than before.

Given the ubiquity of confirmation bias and homophily in social interactions, the argument exchange model developed in this paper could be very helpful to inform future experiments related to opinion formation dynamics and could potentially explain the macroscopic collective states observed in societies. The versatility of the interaction rules allows for simple modifications, for example, the introduction of networks of interaction with different topologies. These could lead to new insights on how people interact, modify their opinions, and make important decisions on different scales, from reduced work meetings to presidential elections.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Secretary, University of Buenos Aires (UBACyT), Argentina, under Grant No. 20020130100582BA and by the National Agency for Science and Technology Promotion (ANPCyT), Argentina, under Grant No. PICT-201-0215.

DATA AVAILABILITY

The data that support the findings of this study are openly available in GitHub, Ref. 70.

REFERENCES

[1] A. Baronchelli, "The emergence of consensus: A primer," R. Soc. Open Sci. **5**, 172189 (2018).
[2] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," Rev. Mod. Phys. **81**, 591–646 (2009).
[3] D. Helbing, *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes* (Springer, Berlin, 2010).
[4] W. Weidlich, "Physics and social science—The approach of synergetics," Phys. Rep. **204**, 1–163 (1991).
[5] T. Berger, "Agent-based spatial models applied to agriculture: A simulation tool for technology diffusion, resource use changes and policy analysis," Agric. Econ. **25**, 245–260 (2001).
[6] J. D. Farmer and D. Foley, "The economy needs agent-based modelling," Nature **460**, 685 (2009).
[7] T. Filatova, P. H. Verburg, D. C. Parker, and C. A. Stannard, "Spatial agent-based models for socio-ecological systems: Challenges and prospects," Environ. Model. Softw. **45**, 1–7 (2013).
[8] F. C. Billari and A. Prskawetz, *Agent-Based Computational Demography: Using Simulation to Improve Our Understanding of Demographic Behaviour* (Springer Science & Business Media, 2012).
[9] R. L. Goldstone and M. A. Janssen, "Computational models of collective behavior," Trends Cogn. Sci. **9**, 424–430 (2005).

[10]T. A. Kohler and G. G. Gumerman, *Dynamics in Human and Primate Societies: Agent-Based Modeling of Social and Spatial Processes* (Oxford University Press, 2000).
[11]P. Clifford and A. Sudbury, "A model for spatial conflict," Biometrika **60**, 581–588 (1973).
[12]R. A. Holley and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the voter model," Ann. Probab. **3**, 643–663 (1975).
[13]T. M. Liggett, *Interacting Particle Systems* (Springer Science & Business Media, 2012), Vol. 276.
[14]J. T. Cox and D. Griffeath, "Diffusive clustering in the two dimensional voter model," Ann. Probab. **14**, 347–370 (1986).
[15]P. Krapivsky, "Kinetics of a monomer-monomer model of heterogeneous catalysis," J. Phys. A: Math. Gen. **25**, 5831 (1992).
[16]L. Frachebourg and P. L. Krapivsky, "Exact results for kinetics of catalytic reactions," Phys. Rev. E **53**, R3009 (1996).
[17]K. Sznajd-Weron, "Mean-field results for the two-component model," Phys. Rev. E **71**, 046110 (2005).
[18]A. Chmiel, J. Sienkiewicz, and K. Sznajd-Weron, "Tricriticality in the *q*-neighbor Ising model on a partially duplex clique," Phys. Rev. E **96**, 062137 (2017).
[19]J. Schneider, "The influence of contrarians and opportunists on the stability of a democracy in the Sznajd model," Int. J. Modern Phys. C **15**, 659–674 (2004).
[20]A. Martins, "Continuous opinions and discrete actions in opinion dynamics problems," Int. J. Modern Phys. C **19**, 617–624 (2008).
[21]N. Crokidakis, V. H. Blanco, and C. Anteneodo, "Impact of contrarians and intransigents in a kinetic model of opinion dynamics," Phys. Rev. E **89**, 013310 (2014).
[22]V. Schwammle, M. C. González, A. A. Moreira, J. S. Andrade, and H. J. Herrmann, "Different topologies for a herding model of opinion," Phys. Rev. E **75**, 066108 (2007).
[23]G. Travieso and L. da Fontoura Costa, "Spread of opinions and proportional voting," Phys. Rev. E **74**, 036112 (2006).
[24]S. Galam, "Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach," J. Math. Psychol. **30**, 426–434 (1986).
[25]S. Galam, "Minority opinion spreading in random geometry," Eur. Phys. J. B **25**, 403–406 (2002).
[26]C. J. Tessone, R. Toral, P. Amengual, H. S. Wio, and M. San Miguel, "Neighborhood models of minority opinion spreading," Eur. Phys. J. B **39**, 535–544 (2004).
[27]S. Thomas *et al.*, *Micromotives and Macrobehavior* (W. W. Norton and Company, New York, 1978).
[28]M. Granovetter and R. Soong, "Threshold models of interpersonal effects in consumer demand," J. Econ. Behav. Organ. **7**, 83–99 (1986).
[29]M. DeGroot, "Reaching a consensus," J. Am. Stat. Assoc. **69**, 118–121 (1974).
[30]G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," Adv. Complex Syst. **3**, 87–98 (2000).
[31]R. Hegselmann, U. Krause *et al.*, "Opinion dynamics and bounded confidence models, analysis, and simulation," J. Artif. Soc. Soc. Simul. **5**, 1–33 (2002).
[32]T. Kurahashi-Nakamura, M. Mäs, and J. Lorenz, "Robust clustering in generalized bounded confidence models," J. Artif. Soc. Soc. Simul. **19**, 7 (2016).
[33]J. Lorenz, "Continuous opinion dynamics under bounded confidence: A survey," Int. J. Modern Phys. C **18**, 1819–1838 (2007).
[34]R. Hegselmann and U. Krause, "Opinion dynamics driven by various ways of averaging," Comput. Econ. **25**, 381–405 (2005).
[35]D. Urbig and J. Lorenz, "Communication regimes in opinion dynamics: Changing the number of communicating agents," arXiv:0708.3334 (2007).
[36]J. Lorenz, M. Neumann, and T. Schröder, "Individual attitude change and societal dynamics: Computational experiments with psychological theories," PsyArXiv (2020).
[37]P. Balenzuela, J. P. Pinasco, and V. Semeshenko, "The undecided have the key: Interaction-driven opinion dynamics in a three state model," PLoS One **10**, e0139572 (2015).
[38]J. K. Shin and J. Lorenz, "Tipping diffusivity in information accumulation systems: More links, less consensus," J. Stat. Mech.: Theor. Exp. **2010**, P06005.
[39]E. Burnstein and A. Vinokur, "Persuasive argumentation and social comparison as determinants of attitude polarization," J. Exp. Soc. Psychol. **13**, 315–332 (1977).
[40]C. R. Sunstein, "The law of group polarization," J. Political Philos. **10**, 175–195 (2002).
[41]W. Jager and F. Amblard, "Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change," Comput. Math. Organ. Theor. **10**, 295–303 (2005).
[42]V. Sampedro and F. S. Pérez, "The 2008 Spanish general elections: "Antagonistic bipolarization" geared by presidential debates, partisanship, and media interests," Int. J. Press/Politics **13**, 336–344 (2008).
[43]P. Dandekar, A. Goel, and D. T. Lee, "Biased assimilation, homophily, and the dynamics of polarization," Proc. Natl. Acad. Sci. U.S.A. **110**, 5791–5796 (2013).
[44]M. L. Jönsson, U. Hahn, and E. J. Olsson, "The kind of group you want to belong to: Effects of group structure on group accuracy," Cognition **142**, 191–204 (2015).
[45]A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz, "Models of social influence: Towards the next frontiers," J. Artif. Soc. Soc. Simul. **20**, 2 (2017).
[46]Z. Krizan and R. S. Baron, "Group polarization and choice-dilemmas: How important is self-categorization?," Eur. J. Soc. Psychol. **37**, 191–201 (2007).
[47]M. Mas and A. Flache, "Differentiation without distancing. Explaining bi-polarization of opinions without negative influence," PLoS One **8**, e74516 (2013).
[48]E. Burnstein and A. Vinokur, "Testing two classes of theories about group induced shifts in individual choice," J. Exp. Soc. Psychol. **9**, 123–137 (1973).
[49]E. Burnstein and A. Vinokur, "What a person thinks upon learning he has chosen differently from others: Nice evidence for the persuasive-arguments explanation of choice shifts," J. Exp. Soc. Psychol. **11**, 412–426 (1975).
[50]P. R. Laughlin and P. C. Earley, "Social combination models, persuasive arguments theory, social comparison theory, and choice shift," J. Pers. Soc. Psychol. **42**, 273 (1982).
[51]V. B. Hinsz and J. H. Davis, "Persuasive arguments theory, group polarization, and choice shifts," Pers. Soc. Psychol. Bull. **10**, 260–268 (1984).
[52]T. W. McGuire, S. Kiesler, and J. Siegel, "Group and computer-mediated discussion effects in risk decision making," J. Pers. Soc. Psychol. **52**, 917 (1987).
[53]H. Mercier and D. Sperber, "Why do humans reason? Arguments for an argumentative theory," Behav. Brain Sci. **34**, 57–74 (2011).
[54]A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," Science **185**, 1124–1131 (1974).
[55]M. C. Parmley, "The effects of the confirmation bias on diagnostic decision making," doctoral dissertation (Drexel University, 2006).
[56]P. E. Lehner, L. Adelman, B. A. Cheikes, and M. J. Brown, "Confirmation bias in complex analyses," IEEE Trans. Syst. Man Cybern. A Syst. Humans **38**, 584–592 (2008).
[57]J. A. Nelson, "The power of stereotyping and confirmation bias to overwhelm accurate assessment: The case of economics, gender, and risk aversion," J. Econ. Method. **21**, 211–231 (2014).
[58]A. Koriat, S. Lichtenstein, and B. Fischhoff, "Reasons for confidence," J. Exp. Psychol. Human Learn. **6**, 107 (1980).
[59]B. O'Brien, "Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations," Psychol. Public Policy Law **15**, 315 (2009).
[60]C. Hill, A. Memon, and P. McGeorge, "The role of confirmation bias in suspect interviews: A systematic evaluation," Leg. Criminol. Psychol. **13**, 357–371 (2008).
[61]F. Tschan, N. K. Semmer, A. Gurtner, L. Bizzari, M. Spychiger, M. Breuer, and S. U. Marsch, "Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making," Small Group Res. **40**, 271–300 (2009).
[62]P. E. Lehner, L. Adelman, R. J. DiStasio, M. C. Erie, J. S. Mittel, and S. L. Olson, "Confirmation bias in the analysis of remote sensing data," IEEE Trans. Syst. Man Cybern. A Syst. Humans **39**, 218–226 (2009).
[63]C. R. Mynatt, M. E. Doherty, and R. D. Tweney, "Confirmation bias in a simulated research environment: An experimental study of scientific inference," Q. J. Exp. Psychol. **29**, 85–95 (1977).
[64]M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," Proc. Natl. Acad. Sci. U.S.A. **113**, 554–559 (2016).

[65]M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Modeling confirmation bias and polarization," Sci. Rep. **7**, 40391 (2017).
[66]M. Starnini, M. Frasca, and A. Baronchelli, "Emergence of metapopulations and echo chambers in mobile agents," Sci. Rep. **6**, 31834 (2016).
[67]D. Byrne, "Interpersonal attraction and attitude similarity," J. Abnorm. Soc. Psychol. **62**, 713 (1961).
[68]A. R. Cohen, "A dissonance analysis of the boomerang effect," J. Pers. **30**(1), 75–88 (1962).
[69]S. Moscovici and M. Zavalloni, "The group as a polarizer of attitudes," J. Pers. Soc. Psychol. **12**, 125 (1969).
[70]F. Barrera Lemarchand (2020). "Polarizing crowds: Consensus and bipolarization in a persuasive arguments model," GitHub. https://github.com/fedex192/Pol-crowds-c-b-in-PAM.