Depolarization of opinions on social networks through random nudges

Ritam Pal,[*] Aanjaneya Kumar,[†] and M. S. Santhanam[‡]
Department of Physics, Indian Institute of Science Education and Research, Dr. Homi Bhabha Road, Pune 411008, India

(Received 23 February 2023; accepted 17 August 2023; published 15 September 2023)

Polarization of opinions has been empirically noted in many online social network platforms. Traditional models of opinion dynamics, based on statistical physics principles, do not account for the emergence of polarization and echo chambers in online network platforms. A recently introduced opinion dynamics model that incorporates the homophily factor—the tendency of agents to connect with those holding similar opinions as their own—captures polarization and echo chamber effects. In this work, we provide a nonintrusive framework for mildly nudging agents in an online community to form random connections. This is shown to lead to significant depolarization of opinions and decrease the echo chamber effects. Though a mild nudge effectively avoids polarization, overdoing this leads to another undesirable effect, namely, radicalization. Further, we obtain the optimal nudge probability to avoid the extremes of polarization and radicalization outcomes.

DOI: 10.1103/PhysRevE.108.034307

I. INTRODUCTION

The information revolution has lowered the entry barrier for nearly everyone to participate and contribute to shaping opinions and policies on various issues. This has been largely aided by the easy availability of social media infrastructure through mobile devices. Increasingly, the collective opinions expressed through various social media platforms are thought to be one barometer of the public mood on any contentious issue of the day [1]. This provides an interesting testing ground for the dynamics and statistical physics of interacting multiagent systems since the online nature of interactions provides fine-grained data for quantitative analysis and comparison with model results.

The study of opinion formation and its dynamics has attracted researchers for decades. The analysis of opinion dynamics from the statistical physics perspective can be traced back to the work of DeGroot [2], which provides a framework for reaching a consensus. Other discrete models, including the voter [3,4] model, Sznajd model [5,6], and their variants which have a strong basis in a framework of interacting spins, suggest that large participatory interactions among agents might also lead to the emergence of consensus. However, empirical results have shown that the distribution of opinions tends to show a bimodal distribution pattern corresponding to polarization, especially on controversial issues of the day [7–9]. Culture dissemination model [10], one of the first higher-dimensional modeling approaches to opinion dynamics, which also incorporates the human tendency to interact with similar persons, shows that despite there being local convergence, global polarization can be reached. Other discrete models [11–14] explain the effects of consensus, attitude changes in groups, and the spreading of minority opinions. In the presence of stubborn agents, these models can also capture the effect of polarization [15–17]. Different variants of the bounded confidence model [18,19] can also capture many empirically found trends in the distribution of opinions. These models can reproduce consensus, bimodal, or multi-modal opinion distributions depending on the confidence interval.

Another empirical feature that could not be accounted for by early models (at least by their original versions) was the phenomenon of echo chambers [20]. This refers to a scenario in which one agent's opinion is similar to the agents in their "social neighborhood," and one tends to reinforce the other. Lack of sufficient engagement with opposing opinions leads to positive reinforcement of one's own opinion within a close-knit social network. Empirical evidence for this effect has been reported from several social media platforms [21–24]. Few recent opinion dynamics models [25–28] have qualitatively captured the features of echo chambers, which have been shown to arise from personalized interactions among peers in an online setting, which might be accelerated through the platform's recommendation engine.

The model introduced by Baumann *et al.* accounted for several observed features from empirical data along with echo chambers in social media. The features that (a) most active users tend to be strongly opinionated and (b) locally connected agents have a convergence of opinions can be linked to the mechanism of reinforcement of opinion among agents and the tendency of agents to interact more with those with similar opinions (homophily [29,30]). Even if the model starts from an initial distribution of opinions without clear preferences, highly homophilic interactions induce the formation of echo chambers and polarized states.

Though having diverse opinions might be a desired outcome, extreme polarization leads to network segregation [31], which often bottlenecks the information flow in social networks. Also, echo chambers, often linked to polarization, are known to be responsible for sustaining misinformation for

[*] ritam.pal@students.iiserpune.ac.in
[†] kumar.aanjaneya@students.iiserpune.ac.in
[‡] santh@iiserpune.ac.in

a longer time on social networks [32,33]. These problems call for intervention mechanisms, which should be safe and noninvasive.

It might appear that in the case of controversial topics, the interaction and the debate will always lead to polarized states of opinion. But the underlying mechanism for polarization, the reinforcement of opinions through interaction between like-minded people, leaves us wondering if any intervention will help to reconcile disparate opinions.

In this work, we show that if agents are nudged slightly, then the cycle of reinforcement of opinions can be broken, and depolarization can be achieved. In social networks, the nudges are effected by exposing the agents to diverse opinions. We also show that overdoing this leads to radicalization [34,35], a state where all the agents have the same stance on an issue. We formulate an optimization problem that avoids polarization and radicalization and computes the right amount of nudge probability required to achieve this optimal scenario.

In the next section, we discuss the basic model and motivate the random nudges in the subsequent section. In Sec. IV, we demonstrate our results and discuss their implications. We formulate an optimization problem in Sec. V, which emerges from a tradeoff between depolarization due to the proposed random nudges and the tendency to move toward a radicalized state. We conclude with a discussion of future directions.

II. BASIC MODEL AND METHODS

To analyze polarization and to introduce possible intervention methods for reducing polarization, we adapt a recently introduced model for opinion dynamics [25]. This model qualitatively captures a few aspects of opinion dynamics when agents' opinions evolve due to interactions in social media platforms. The model can reproduce the empirical features such as polarization and echo chambers and the fact that more active people on social media tend to have extreme opinions.

The model has N interacting agents, and it is assumed there are only two possible sides to an issue. This is typical of many, but not all, the issues—for example, to allow abortion or not. Opinion on a given issue is denoted by x_i, which can take any real value in the range $(-\infty, \infty)$. The sign of the x_i corresponds to the stance of the agent in the corresponding issue, and $|x_i|$ denotes the conviction of the agent in their respective stance. This implies that the larger the value of $|x_i|$, the more extreme the agent's opinion is. The model used to capture the evolution of opinion is activity driven [36–39], i.e., at each time step, only active agents can influence other agents. Based on empirical data [36,38], the distribution of agent's activity is chosen to be

$$F(a) = \frac{1-\gamma}{1-\varepsilon^{1-\gamma}} a^{-\gamma}, \tag{1}$$

where a is the activity, ε is the minimum activity (chosen in this work to be 10^{-2}), and γ controls how steep the function $F(a)$. It is chosen to be $\gamma = 2.1$. Agents' opinions evolve based on their interactions with other agents, and this information is encoded in the time-dependent adjacency matrix $A_{i,j}(t)$. Further, opinion evolution also depends on the strength of social interaction $K > 0$ and the controversialness of the issue $\alpha > 0$. The opinion dynamics is given by the following N coupled differential equation [25]:

$$\dot{x}_i = -x_i + K\left(\sum_{j=1}^{N} A_{ij}(t) \tanh{(\alpha x_j)}\right). \tag{2}$$

In this, $A_{i,j}(t)$ is the temporal adjacency matrix of interaction at time t. If at time t agent j influences agent i, then $A_{i,j}(t) = 1$, and $A_{i,j}(t) = 0$ otherwise. If agent i is active at time t, they will interact with m other agents, weighted by the probability $P_{i,j}$. Further, the probabilistic reciprocity factor $r \in [0, 1]$ determines the chance that an interaction is mutually influential, i.e., $A_{ij}(t) = A_{ji}(t) = 1$. The interaction probability is defined to be a function of the magnitude between two agents' opinions:

$$P_{ij} = \frac{|x_i - x_j|^{-\beta}}{\sum_k |x_i - x_k|^{-\beta}}, \tag{3}$$

where β is the homophily factor which quantifies the tendency for agents with similar opinions to interact with each other; $\beta = 0$ refers to the absence of interaction preference; and $\beta > 0$ implies that the agents with similar opinions are more likely to interact with one another. Evidently, Eq. (3) is modeled as a power-law decay of connection probabilities with only a small chance for agents with opposite opinions to interact. Since most of the interactions tend to occur between agents with similar opinions, this can lead to the formation of echo chambers.

The interaction dynamics in the model is enforced by the activity-driven temporal network that is fully encoded by the parameters $(\varepsilon, \gamma, m, \beta, r)$, together with the parameters that characterize the issue, (K, α). Asymptotically, this model features three distinct states in the distribution of opinions. If the social interaction K is sufficiently small, then the opinion of every agent decays to zero, and this state is known as the neutral consensus state. However, if social interaction K is large, but the homophily factor β is small, then, due to statistical fluctuations, all the opinions either become positive or negative. This state, where each agent has the same stance (the sign of x_i for all i is the same) with possibly different convictions, is called radicalization. It is important to note that radicalization is an absorbing state of this model. This is because when all agents have opinions with the same sign, the dynamics does not allow for a sign change of any agent's opinion. The most interesting case emerges when social interaction K and homophily factor β are large enough. In this case, a metastable polarized state emerges, which is characterized by a bimodal opinion distribution.

III. RANDOM NUDGES AND POLARIZATION

Echo chambers are increasingly becoming more apparent in online social media platforms. A generic tendency to interact with people who hold similar opinions as ours can lead to echo chambers, and this effect is, in turn, amplified by the recommendation engines on social media platforms. These algorithmically driven engines recommend similar connections or content in order to keep the users of those platforms engaged.

These two features are modeled by the interaction probability, controlled by the homophily factor β. Large values

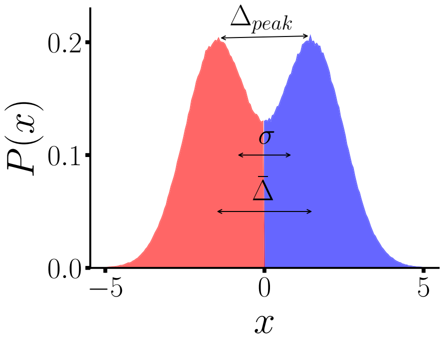

FIG. 1. A schematic to illustrate three measures of polarization. $\bar{\Delta}$ is the distance between mean positive and negative opinions. Δ_{peak} denotes the distance between two peaks in the opinion distribution, and $\boldsymbol{\sigma}$ denotes the standard deviation of the opinion distribution.

of β represent how closed the echo chambers are. To disrupt the formation of echo chambers, even while keeping the platforms as engaging as possible and without violating the users' privacy, we adopt the following intervention in the opinion dynamics model: With probability $p < 1$, the active agents will interact uniformly with any other agents, and with probability $(1 - p)$, the active agents will interact with others according to the homophily probability given in Eq. (3). We call p the random nudge probability. As p does not depend on the opinions of the agents, the intervention is noninvasive (the recommendation engine need not interpret the opinion of the agents). For small enough values of p, it is hoped that the platform is still engaging while maintaining enough diversity to ensure there is no echo chamber. With this intervention, we propose a modified interaction probability as

$$\widetilde{P}_{ij} = p \times \frac{1}{N-1} + (1-p) \times P_{ij}. \tag{4}$$

This is used in the rest of the results shown in this paper.

Quantifying Polarization. Before we delve into the details of the results, we discuss the three quantities employed to measure the degree of polarization based on the opinion distribution $P(x)$. They are defined as (a) Polarization measured through $\bar{\Delta}$, defined as the distance between the average of positive opinions and the average of negative opinions. (b) When opinion distribution exhibits a bimodal character, the distance between the two peaks, denoted by Δ_{peak}, can also be used as a measure of polarization [41]. (c) A gross measure of polarization could also be the standard deviation σ of the entire opinion distribution [27]. Figure 1 illustrates the schematics of all three measures of polarization. It must be noted that if polarization decreases due to the intervention proposed in Eq. (4), ideally, all three quantifiers must decrease.

We also define f_{ext} as the fraction of agents with conviction $|x| > x_{th}$, where x_{th} (chosen to be five) is a positive threshold. This quantifies the prevalence of extreme opinions among the agents, which at least should not increase when we nudge the agents.

IV. RESULTS

With the intervention strategy introduced in Sec. III, we find that with sufficiently small random nudge probability p, significant depolarization can be obtained, which is evident as the opinion distributions approach toward a unimodal distribution along with the decay of all three measures of polarization. To see the effects of nudge, we perform numerical simulations of the basic model in Eq. (2) using the interaction probability given in Eq. (3) and the intervention model in Eq. (4). The simulations are performed with $N = 5000$ agents for 1000 time steps with $dt = 0.01$. At initial time, x_i is uniformly chosen from a small interval, i.e., $x_i \in [-1, 1]$ for $i = 1, 2, \dots, N$. The model parameters are chosen to be $\alpha = 3$, $\beta = 3$, $K = 3$, $m = 10$, $\gamma = 2.1$, $\varepsilon = 0.01$, and $r = 0.5$ for all the simulations unless mentioned otherwise. The parameters chosen for the simulations lead to a polarized state in the original model without intervention.

In Fig. 2, we show the contrast between the trajectories of individual opinions and the opinion distribution with and without the application of a nudge. In the absence of nudge ($p = 0$), the simulation results in Fig. 2(a) show fewer trajectories with opinions $x_i \approx 0$. This leads to a bimodal distribution of opinions characteristic of a polarized state. In contrast, in Fig 2(b), a small nudge with a probability of $p = 0.01$ is applied, and we find significantly more trajectories with moderate opinions. This, effectively, is seen to lead to an absence of polarization, and is evident from the unimodal opinion distribution. The magnifications of the region around $x_i = 0$ and its distribution (shown in Fig. 2) reveal a clear distinction between these two scenarios.

To examine the effect of network nudge, we analyze the underlying time-averaged structures of the temporal interactions network. Without nudge, the interaction network has two distinct clusters; most of the connections are among positive opinionated agents or negative opinionated agents. There exist very few connections between these two groups other than for the agents with extreme opinions. This is expected since the agents with extreme opinions are also those who tend to be more active on social networks fora; hence, on average, they form more connections. This enables them to be relatively more connected to the agents with opposing opinions. These results are visually depicted in Fig. 3 as two snapshots of evolving network diagrams. If $p = 0$, no nudge is applied. In this case, as Fig 3(b) shows, a polarized network, made up of two distinct blue- and red-colored clusters, is formed. Blue color corresponds to nodes with $x > 0$, and red color to $x < 0$. The opinion distribution shown in Fig. 3(a) confirms the existence of polarization.

However, when a nudge is applied, even for the case when the nudge probability is as small as $p = 0.01$, we find the network to be well mixed (large blue and red clusters have disappeared) [Fig. 3(e)], and this leads to a significantly depolarized state indicated by the approximate unimodality of the opinion distribution as shown in Fig. 3(d). The term echo chamber describes a situation where the beliefs or opinions of people are reinforced by interactions among a closed group of people who hold similar opinions. In recent years, this has been widely discussed in the context of online communities [21–24]. However, some studies appear to suggest that the effects of echo chambers are over estimated [42]. To infer the presence of echo chamber-type effects, we calculate the average opinion of the nearest neighbors (NN) of each agent

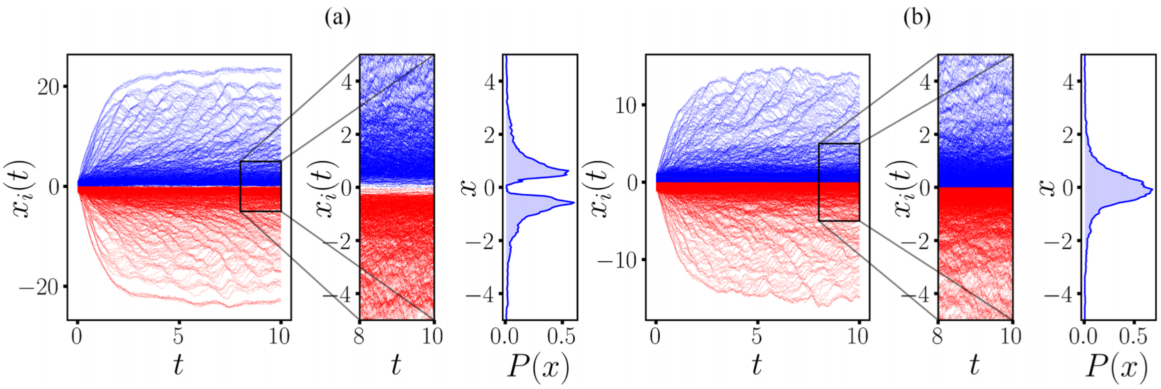

FIG. 2. Emergent polarized (and depolarized) states in the presence (and absence) of the nudge factor. The simulations are performed with 10 000 agents, and parameters are set to promote polarization. (a) The agents are not nudged. Hence the polarized state emerges. A magnification of the region around $x = 0$ reveals the absence of trajectories there, and the corresponding distribution shows a bimodal distribution with a near-zero density close to $x \approx 0$. (b) Network nudge is introduced with probability $p = 0.01$, and we find a significant depolarization. Opinion trajectories tend to crowd around $x = 0$, and the opinion distribution approaches an approximate unimodal and almost-symmetric distribution about $x = 0$.

[24,25]. This is denoted by

$$\langle x_{NN}\rangle = k_i^{-1}\sum_j a_{ij}x_j, \quad \text{and} \quad k_i = \sum_j a_{ij}, \tag{5}$$

where a_{ij} is the temporally aggregated (over the last 100 time steps) adjacency matrix. When a nudge is not applied ($p = 0$), a colored heatmap of x and $\langle x_{NN}\rangle$ in Fig. 3(c) reveals two disjoint hot spots corresponding to the two distinct echo chambers. A strong bimodality is observed in the marginal distributions. Now, when we apply a nudge with probability $p = 0.01$, we can observe only one hot spot indicating the existence of only one closed group [Fig. 3(f)]. All the agents are inside this closed group, and the echo chamber effect is largely diluted or nonexistent. We did not find perfect unimodality in the marginal distribution of x, which can be attributed to the fact that different realizations can lead to either of these three distributions: (a) slight bimodal distribution with significant reduction in all three polarization parameters, (b) unimodal distribution with a slight skew toward positive opinions, and (c) similar distribution with a skew toward negative opinions. As the heat maps and the marginal distributions are created from data averaged over 200 realizations,

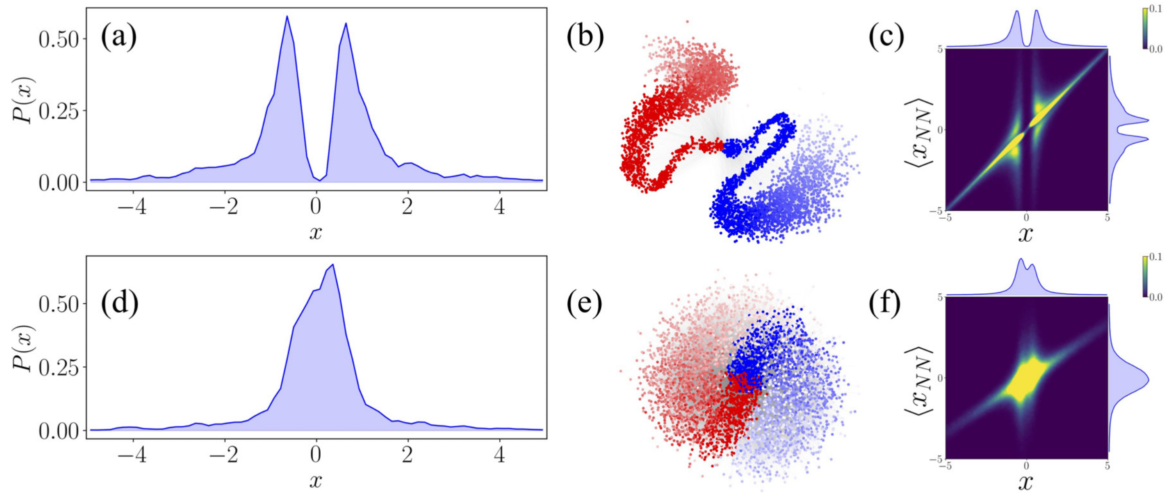

FIG. 3. Effect of the nudge on the opinion distribution, the structure of social interactions networks, and the signature of echo chambers. The networks are averaged over the last 100 time steps of simulation and are drawn using the `draw` function in `networkx` [40]. Nodes with blue color correspond to agents with positive opinions, and red corresponds to agents with negative opinions. The saturation of the color is mapped to the conviction of the agents; high saturation corresponds to a high level of conviction, and vice versa. The opinion of an agent x and the mean opinion of its nearest neighbors $\langle x_{NN}\rangle$ is averaged over 200 realizations to generate the heatmap to indicate the presence of echo chambers [see Eq. (5)]. The marginal distributions are shown in the corresponding axes. (a) For $p = 0$, i.e., without a nudge, the distribution is polarized, and the network has two distinct clusters (b), one formed by the agents with positive opinions and the other by the agents with negative opinions. (c) The presence of two distinct lobes in the heatmap indicates the echo chamber effect. (d) For $p = 0.01$, we observe an opinion distribution with a single peak, and the social interactions network is now well mixed (e). A depolarization state is reached. (f) A single lobe in the heatmap confirms the weakening of the echo chamber effect.

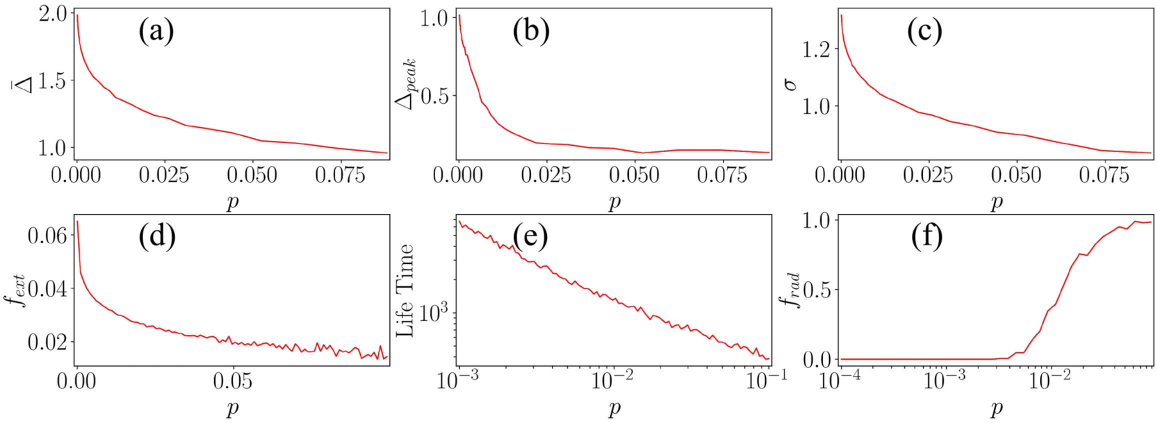

FIG. 4. Three measures of polarization, (a) $\bar{\Delta}$, (b) Δ_{peak}, (c) σ, and the fraction f_{ext} of agents with extreme opinions (d), as a function of nudge strength p. All four parameters are averaged over the last 100 time steps. The simulations were repeated 200 times, and only nonradicalized realizations were considered for ensemble averaging. The average lifetime until the whole population moves toward radicalization as a function of p is shown in panel (e). Panel (f) shows the fraction of simulations that lead to radicalization for different nudge strengths p.

all the above factors contribute to the slight bimodality in the marginal distribution of x. Nevertheless, the marginal distribution corresponds to a significant reduction in polarization and echo chambers.

V. OPTIMIZING THE NUDGE: POLARIZATION VERSUS RADICALIZATION

To obtain a global picture of how depolarization sets in as a function of nudge probability p, we plot the three measures of polarization as a function of p. All three measures, $\bar{\Delta}$, Δ_{peak}, and σ, have been computed from the simulation results. The results shown represent an average over the last 100 time steps of the simulations and averaged over 200 realizations. In Fig. 4, we observe that all three measures of polarization decrease as the strength of the nudge p increases. In particular, $\bar{\Delta}$ and σ are found to decrease as a stretched exponential function $\exp(-p^{\gamma})$, and the stretching factor γ is determined through regression to be approximately 0.3. A recent work studying the depolarization of echo chambers [41] considered adding an effective noise term dependent on a random sample of opinions to Eq. (2). While this approach succeeds in making the opinion distribution unimodal, it increases the width of the distribution significantly, which as a consequence, corresponds to an increase in extreme opinions. In contrast, the framework of nudging the mechanism of forming social connections in online interactions works well in decreasing width of the opinion distribution [Fig. 4(c)] as well as extreme opinions [Fig 4(d)] and also suggests direct algorithmic interventions for recommender systems.

In the original model, the authors found the polarized state to be metastable and showed that with an increased value of β, the lifetime of the state has a faster than exponential growth. Our intervention adds more randomness to the system and increases statistical fluctuations. Hence, for large p, we observe a drastic decrease in the average lifetime of the polarized and depolarized states. An approximate straight line in the log-log plot indicates the lifetime of polarized or depolarized states decreases as a power law as nudge strength p is increased [see Fig. 4(e)]. Figure 4(f) also captures the same effect as we see that radicalization is either nonexistent or a rarity for $p < 10^{-2}$, but it increases quickly and becomes the norm for $p > 10^{-2}$.

In many situations, radicalization is as much undesirable as polarization. Hence, to solve the issue of radicalization at a high value of nudge probability, rather than nudging all the people in the population, at each time step of the simulation, we randomly selected a fraction f of the population and nudged them. We define a simple linear utility function $U(\bar{\Delta}, f_{\text{rad}}) = \tilde{\bar{\Delta}} + f_{\text{rad}}$, where $\tilde{\bar{\Delta}}$ is $\bar{\Delta}$, linearly scaled to be between zero and one, and f_{rad} is the fraction of radicalized simulations. The structure of the utility function is the same for the other two measures of polarization. Figure 5 depicts the heat map of the utility functions corresponding to the three utility functions. The optimal population fraction and nudge probability is numerically found to follow the curve $p \cdot f^{A} = B$, where A and B are constants.

VI. ROBUSTNESS OF THE FRAMEWORK

To ensure the robustness of our intervention framework, we applied network nudge to another recent model of opinion dynamics, namely the social compass model [43,44], which, together with homophily, exhibits the effect of echo chambers. The original model describes the dynamics of opinions on two interdependent topics in polar coordinates. We reinterpret the polar angle in the original model as the opinion on a single topic to adapt the model to our framework. The dynamics of this modified model is governed by the following N coupled differential equation:

$$\dot{x}_i(t) = |x_i| \sin\left(x_i^0 - x_i\right) + K\left(\sum_{j=1}^{N} A_{ij}(t) \sin\left(x_i - x_j\right)\right). \quad (6)$$

In contrast to the original model [43], the variable x_i is chosen to be the opinions of the people on a single topic, and the temporal adjacency matrix is formed according to homophily probability 3. x_i^0 is the initial opinion of agent i, and all the other variables and parameters have the same meaning as in

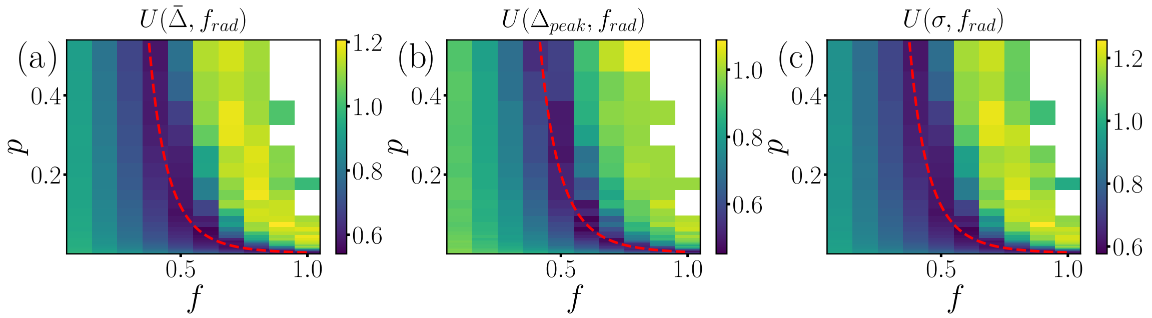

FIG. 5. The heat map of the utility as a function of nudge strength and nudged population fraction. Panels (a), (b), and (c) correspond to the corresponding utility of $\bar{\Delta}$, Δ_{peak}, and σ, respectively. The red dashed curve, which is found to follow the curve $p \cdot f^A = B$, $(A, B = \text{constants})$, denotes the optimal values of population fraction and nudge strength.

the previous model 2. In Fig. 6, we show that when the social interaction and the homophily factor are high enough ($K = 4$, $\beta = 4$), many echo chambers are formed, which is clear from the trajectories of the opinion as well as from the multiple communities seen in the aggregated network [Figs. 6(a) and 6(b)]. But when we introduce a slight nudge with $p = 0.002$, the effect of echo chambers is reduced drastically. The opinion trajectories seem to converge to a moderate value, and the interaction network is well connected without any obvious segregated communities Figs. 6(c) and 6(d).

VII. DISCUSSION

The widespread use of the internet, and consequently, social media platforms, have drastically altered the way humans consume, interact with, and exchange information. Polarization and the formation of echo chambers have been shown to negatively impact constructive discussions and debates—two fundamental pillars of a healthy democracy. Building on the recent advances in the modeling of opinion dynamics in social networks, in this work, we study the possibility of depolarizing a population using a stochastic nudge.

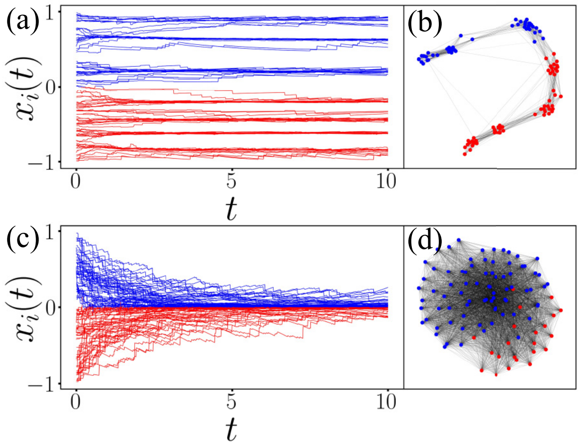

FIG. 6. The effect of nudge in the opinion dynamics model, governed by Eq. (6). Panels (a) and (c) show the trajectories of opinions in the absence and presence of network nudge, respectively. Panels (b) and (d) show the corresponding interaction network structure. Clearly, we see the presence of echo chambers in the absence of a network nudge, and the effect decreases when a slight nudge is applied.

Our results suggest that a small number of randomized interactions, which are otherwise dominated by homophily driven mechanisms, can lead to a significant reduction in polarization. This reduction was quantitatively captured by three different measures of polarization. While we show that minimal nudges can burst echo chambers and lead to socially desirable distributions of opinions, increasing the strength of this nudge can result in radicalization. Given this sensitivity on the nudge strength, we show that a possible resolution is obtained if, instead of nudging each agent, only a fraction f of the agents are nudged. We highlight that this interplay of the nudge strength p and the fraction f of nudged individuals leads to an interesting optimization problem. This optimization can help inform the fraction of individuals to be nudged for a fixed nudge strength for optimal depolarization.

We believe that the strongest case for the application of such randomized nudges can be made to recommendation systems. While ubiquitous, recommender algorithms are optimized for increasing engagement [45], which we now know can come at the cost of creating echo chambers [46], increase in the representation of extreme ideologies [47], and even the tampering of users' preferences [48]. In such settings, the randomized nudges can be potentially operationalized as the poisoning of a viewer's watch history with a limited amount of random content, uncorrelated with the viewer's preferences [49]. While there are several ethical and legal considerations that must be accounted for before implementing any such interventions, it certainly opens up several interesting avenues for future research to build on. Noninvasive interventions may be important to reduce the detrimental effects of polarization. However, an important first step is to build reliable tools to quantify polarization from data [50], which in itself constitutes an intriguing direction for future research.

ACKNOWLEDGMENTS

R.P. and A.K. gratefully acknowledge the Prime Minister's Research Fellowship of the Government of India for financial support. M.S.S. acknowledges the support of MATRICS Grant MTR/2019/001111 from SERB, Government of India. The authors acknowledge the National Supercomputing Mission for the use of PARAM Brahma at IISER Pune.

[1] S. C. McGregor, Social media as public opinion: How journalists use social media to represent public opinion, Journalism **20**, 1070 (2019).

[2] M. H. DeGroot, Reaching a consensus, J. Am. Stat. Assoc. **69**, 118 (1974).

[3] R. A. Holley and T. M. Liggett, Ergodic theorems for weakly interacting infinite systems and the voter model, Ann. Probab. **3**, 643 (1975).

[4] S. Redner, Reality-inspired voter models: A mini-review, C. R. Phys. **20**, 275 (2019).

[5] K. Sznajd-Weron and J. Sznajd, Opinion evolution in closed community, Int. J. Mod. Phys. C **11**, 1157 (2000).

[6] K. Sznajd-Weron, J. Sznajd, and T. Weron, A review on the sznajd model-20 years after, Physica A **565**, 125537 (2021).

[7] C. G. Lord, L. Ross, and M. R. Lepper, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. J. Pers. Soc. Psychol. **37**, 2098 (1979).

[8] P. DiMaggio, J. Evans, and B. Bryson, Have american's social attitudes become more polarized? Am. J. Sociol. **102**, 690 (1996).

[9] D. Baldassarri and A. Gelman, Partisans without constraint: Political polarization and trends in american public opinion, Am. J. Sociol. **114**, 408 (2008).

[10] R. Axelrod, The dissemination of culture: A model with local convergence and global polarization, J. Conflict Resol. **41**, 203 (1997).

[11] S. Galam, Y. Gefen, and Y. Shapir, Sociophysics: A new approach of sociological collective behaviour. i. mean-behaviour description of a strike, J. Math. Sociol. **9**, 1 (1982).

[12] S. Galam and S. Moscovici, Towards a theory of collective phenomena: Consensus and attitude changes in groups, Eur. J. Soc. Psychol. **21**, 49 (1991).

[13] S. Galam, Minority opinion spreading in random geometry, Eur. Phys. J. B **25**, 403 (2002).

[14] S. Galam, Sociophysics: A Physicist's Modeling of Psycho-political Phenomena, *Understanding Complex Systems* (Springer, New York, 2012).

[15] S. Galam and F. Jacobs, The role of inflexible minorities in the breaking of democratic opinion dynamics, Physica A **381**, 366 (2007).

[16] S. Galam, Stubbornness as an unfortunate key to win a public debate: An illustration from sociophysics, Mind & Society **15**, 117 (2016).

[17] S. Galam, Collective beliefs versus individual inflexibility: The unavoidable biases of a public debate, Physica A **390**, 3036 (2011).

[18] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Mixing beliefs among interacting agents, Adv. Complex Syst. **03**, 87 (2000).

[19] R. Hegselmann and U. Krause, Opinion dynamics and bounded confidence: Models, analysis, and simulation, J. Art. Soc. Soc. Simul. (JASSS) **5** (2002).

[20] R. K. Garrett, Echo chambers online?: Politically motivated selective exposure among internet news users, J. Comput.-Mediat. Comm. **14**, 265 (2009).

[21] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, Echo chambers: Emotional contagion and group polarization on facebook, Sci. Rep. **6**, 37825 (2016).

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

[23] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship, in *Proceedings of the 2018 World Wide Web Conference* (ACM Digital Library, New York, 2018) pp. 913–922.

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

[25] 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).

[26] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Emergence of Polarized Ideological Opinions in Multidimensional Topic Spaces, Phys. Rev. X **11**, 011012 (2021).

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

[28] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, Social influence and unfollowing accelerate the emergence of echo chambers, J. Comput. Soc. Sci. **4**, 381 (2021).

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. **27**, 415 (2001).

[30] A. Bessi, F. Petroni, M. D. Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, Homophily and polarization in the age of misinformation, Eur. Phys. J.: Spec. Top. **225**, 2047 (2016).

[31] V. V. Vasconcelos, S. M. Constantino, A. Dannenberg, M. Lumkowsky, E. Weber, and S. Levin, Segregation and clustering of preferences erode socially beneficial coordination, Proc. Natl. Acad. Sci. **118**, e2102153118 (2021).

[32] P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion, PLoS One **13**, e0203958 (2018).

[33] 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. **113**, 554 (2016).

[34] D. G. Myers and H. Lamm, The group polarization phenomenon. Psychological Bulletin **83**, 602 (1976).

[35] D. J. Isenberg, Group polarization: A critical review and meta-analysis. J. Pers. Soc. Psychol. **50**, 1141 (1986).

[36] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Activity driven modeling of time varying networks, Sci. Rep. **2**, 469 (2012).

[37] M. Starnini and R. Pastor-Satorras, Topological properties of a time-integrated activity-driven network, Phys. Rev. E **87**, 062807 (2013).

[38] A. Moinet, M. Starnini, and R. Pastor-Satorras, Burstiness and Aging in Social Temporal Networks, Phys. Rev. Lett. **114**, 108701 (2015).

[39] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Controlling Contagion Processes in Activity Driven Networks, Phys. Rev. Lett. **112**, 118702 (2014).

[40] A. Hagberg, P. Swart, and D. S Chult, Exploring network structure, dynamics, and function using NetworkX, Tech. Rep. No. LA-UR-08-05495, Los Alamos National Lab. (LANL), Los Alamos, NM, United States (2008).

[41] C. B. Currin, S. V. Vera, and A. Khaledi-Nasab, Depolarization of echo chambers by random dynamical nudge, Sci. Rep. **12**, 9234 (2022).

[42] E. Dubois and G. Blank, The echo chamber is overstated: The moderating effect of political interest and diverse media, Inf. Commun. Soc. **21**, 729 (2018).

[43] J. Ojer, M. Starnini, and R. Pastor-Satorras, Modeling Explosive Opinion Depolarization in Interdependent Topics, Phys. Rev. Lett. **130**, 207401 (2023).

[44] J. Ojer, M. Starnini, and R. Pastor-Satorras, Vanishing threshold in depolarization of correlated opinions on social networks, arXiv:2306.01329.

[45] S. Milano, M. Taddeo, and L. Floridi, Recommender systems and their ethical challenges, AI Soc. **35**, 957 (2020).

[46] E. Noordeh, R. Levin, R. Jiang, and H. Shadmany, Echo chambers in collaborative filtering based recommendation systems, arXiv:2011.03890.

[47] J. Whittaker, S. Looney, A. Reed, and F. Votta, Recommender systems and the amplification of extremist content, Internet Policy **10**, 1 (2021).

[48] C. Evans and A. Kasirzadeh, User tampering in reinforcement learning recommender systems, arXiv:2109.04083.

[49] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak, Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations, arXiv:2203.10666.

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