Micro-foundation of opinion dynamics: Rich consequences of the weighted-median mechanism

Wenjun Mei *
Department of Mechanics and Engineering Science, Peking University, Beijing, China

Francesco Bullo
Center for Control, Dynamical Systems, and Computation, University of California, Santa Barbara, California 93106, USA

Ge Chen
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Julien M. Hendrickx
Institute of Information and Communication Technologies, Electronics, and Applied Mathematics, UCLouvain, Louvain-la-Neuve, Belgium

Florian Dörfler
Automatic Control Laboratory, ETH Zurich, Zurich, Switzerland

(Received 8 November 2020; revised 13 January 2022; accepted 11 April 2022; published 14 June 2022)

The key to obtaining a mechanistic and reliable understanding of complex public opinion formation processes is to identify the main mechanism governing interpersonal influence. Researchers have long been exploring simple yet predictive mathematical models of opinion dynamics. Although most models are based on the assumption that individuals update their opinions by averaging others' opinions, researchers might need to rethink this universally adopted micro-foundation. The deceivingly simple weighted-averaging mechanism features a non-negligible unrealistic implication, which brings unnecessary difficulties in seeking a proper balance between model complexity and predictive power. In this paper, we fundamentally resolve this problem by proposing the weighted-median mechanism as a new micro-foundation of opinion dynamics. Such an inconspicuous change from averaging to median leads to rich consequences. The weighted-median mechanism, derived from the cognitive dissonance theory in psychology, is well supported by online experiment data. It also broadens the applicability of opinion dynamics models to multiple-choice issues with ordered discrete options, e.g., political elections. Moreover, comparative studies show that the weighted-median mechanism predicts various real-world patterns of opinion evolution while some widely studied averaging-based models fail to, including how group structure affects the likelihood of reaching consensus and how extreme opinions are located in social networks.

DOI: 10.1103/PhysRevResearch.4.023213

I. INTRODUCTION

The key discourse in democratic society starts from exchanges of opinions in deliberative groups, or via social media, to eventually reaching consensus or disagreements. Mathematical models play a key role in obtaining mechanistic understandings of how empirically observed macroscopic opinion-formation phenomena emerge from certain microscopic social-influence mechanisms, as well as certain social network structures. Due to the complexity of social interactions, the key challenge in building predictive and mathematically tractable models is to identify the "salient features," i.e., the micro-foundations, that govern the interpersonal influence processes.

Most existing opinion dynamics models are based upon a common micro-foundation: the weighted-averaging mechanism, also known as the classic *DeGroot model* [1,2]. In the DeGroot model, individuals' opinions on a certain issue are denoted by real numbers, and are assumed to be updated by taking some weighted averaging opinions of others. The weights individuals assign to each other define a weighted and directed graph, referred to as the *influence network*. The DeGroot model is deceivingly elegant but leads to an overly simplified prediction that the individuals' opinions reach consensus whenever the influence network has a globally reachable and aperiodic strongly connected component [2,3] (see Sec. I of the Supplemental Material [4] for a brief review of graph theory). This bold conclusion under mild connectivity conditions leaves tricky puzzles for researchers, e.g., Axelrod's puzzle, "If people tend to become more alike in their beliefs, attitudes, and behavior when they interact, why do not all such differences eventually disappear?" [5],

*Corresponding author: mei@pku.edu.cn

and Abelson's puzzle, "Since universal ultimate agreement is an ubiquitous outcome ... what on earth must one assume in order to generate the bimodal outcome of community cleavage studies?" [6].

As efforts to resolve the above puzzles, numerous important extensions of the DeGroot model have been proposed by introducing additional assumptions and parameters; e.g., see some representative models [6–15] and survey papers [3,16,17]. Some of these extensions manage to generate the phenomena of persistent disagreement, opinion polarization, or opinion clustering, however, at the cost of losing model simplicity or mathematical tractability. These models would also be at risk of being overparametrized if they were to be further extended to capture a broader set of features of real-world opinion evolution instead of one specific phenomenon. Moreover, despite recent developments, the research of opinion dynamics, as remarked by Flache *et al.* [18], faces two main challenges: (1) an "urgent need for more theoretical work comparing, relating, and integrating alternative models"; (2) "a strong imbalance between a proliferation of theoretical studies and a dearth of empirical work."

In this paper, we propose an opinion dynamics model that is the simplest in form but, surprisingly, addresses all the puzzles and concerns above. Via identifying and fundamentally resolving an intrinsic unrealistic feature of the universally adopted weighted-averaging mechanism, we propose a new micro-foundation of opinion dynamics, i.e., the *weighted-median mechanism*. This new mechanism is inspired by the cognitive dissonance theory in psychology [19,20] and derived in the framework of network games [21]. As indicated by a complete set of studies, such an inconspicuous change from averaging to median leads to rich consequences.

First, in the weighted-median mechanism, Axelrod's puzzle [5] is no longer a puzzle. Although the weighted-median mechanism still implies that individuals tend to be attracted by others' opinions, it does not necessarily lead to eventual consensus.

Second, due to the nature of the median operation, the new mechanism is independent of numerical representation of opinions and broadens the applicability of opinion dynamics models to multiple-choice issues with ordered discrete options, e.g., political elections among parties over an ideological spectrum.

Third, the weighted-median mechanism, derived from first principles, is also empirically well supported. Analysis of an online experiment data set [22] indicates that median-based mechanisms enjoy significantly lower errors than averaging-based mechanisms in predicting individuals' opinion shifts under social influence.

Finally, via numerical studies, we directly compare in various aspects the predictions by the weighted-median mechanism and some widely studied extensions of the DeGroot model. The simulation results lead to the following meaningful observations:

(i) On the group level, only the weighted-median mechanism fully captures the empirically supported feature that consensus is less likely to be achieved in larger groups or groups with more clustered structures.

(ii) Regarding how extreme opinions are located in social networks, only predictions by the weighted-median mechanism are consistent with the patterns revealed by a real Twitter data set. That is, extreme opinions tend to reside in "peripheral areas" of the network and form densely connected small local clusters.

(iii) Without deliberately tuning model parameters, only the weighted-median mechanism generates various empirically observed steady public opinion distributions. Namely, the weighted-median mechanism offers perhaps the simplest answer to Abelson's puzzle [6].

To sum up, while it is implausible for one single model to explain every aspect of real-world opinion dynamics, the evidence above supports the weighted-median mechanism as a well-founded mechanism of social influence. Moreover, due to the simplicity of the weighted-median mechanism, it could serve well as a new foundation for further improvements via extensions in various directions.

II. FROM WEIGHTED AVERAGING TO WEIGHTED MEDIAN

A. An unrealistic feature of weighted averaging

Consider a group of n individuals indexed by $i = 1, 2, \ldots, n$. Denote by $x_i(t)$ the opinion of i on a certain issue at time t. The DeGroot model [1,2] is a discrete-time dynamics taking the following form:

$$x_i(t+1) = \text{Mean}_i(x(t); W) = \sum_{j=1}^{n} w_{ij} x_j(t), \tag{1}$$

where w_{ij} denotes the weight individual i assigns to individual j's opinion, i.e., individual j's influence on i. The *influence matrix* $W = (w_{ij})_{n\times n}$ induces a directed and weighted graph, referred to as the influence network and denoted by $G(W)$; see an example in Fig. 1(a). Namely, each individual is a node in $G(W)$, and each entry w_{ij} corresponds to a link from i to j with weight w_{ij}. By definition, $w_{ij} \geqslant 0$ for any i, j, and $w_{i1} + \cdots + w_{in} = 1$ for any i. As predicted by the DeGroot model, consensus is always achieved, i.e., $x_i(t) - x_j(t) \to 0$ as $t \to \infty$ for any i, j, whenever the influence network $G(W)$ has a globally reachable and aperiodic strongly connected component. This is a bold prediction under a mild connectivity condition, considering that persistent disagreement is at least as prevalent as consensus in human groups.

The intuition behind the DeGroot model's always-consensus prediction is that the weighted-averaging mechanism leads to a non-negligibly unrealistic implication, illustrated via the following simple example and visualized in Fig. 1(b): Suppose an individual i is influenced by individuals j and k via the weighted-averaging mechanism:

$$x_i(t+1) = x_i(t) + w_{ik}(x_k(t) - x_i(t)) + w_{ij}(x_j(t) - x_i(t)).$$

The equation above implies that whether i's opinion moves toward $x_k(t)$ or $x_j(t)$ is determined by whether $w_{ik}|x_k(t) - x_i(t)|$ is larger than $w_{ij}|x_j(t) - x_i(t)|$. That is, the "attractiveness" of opinion $x_j(t)$ to individual i is proportional to the opinion distance $|x_j(t) - x_i(t)|$. Such proportionality implies overly large "attractive forces" between distant opinions, which override the effects of any delicate network structures on forming local opinion clusters. As a result, the DeGroot model is driven to consensus under mild network connectivity conditions.

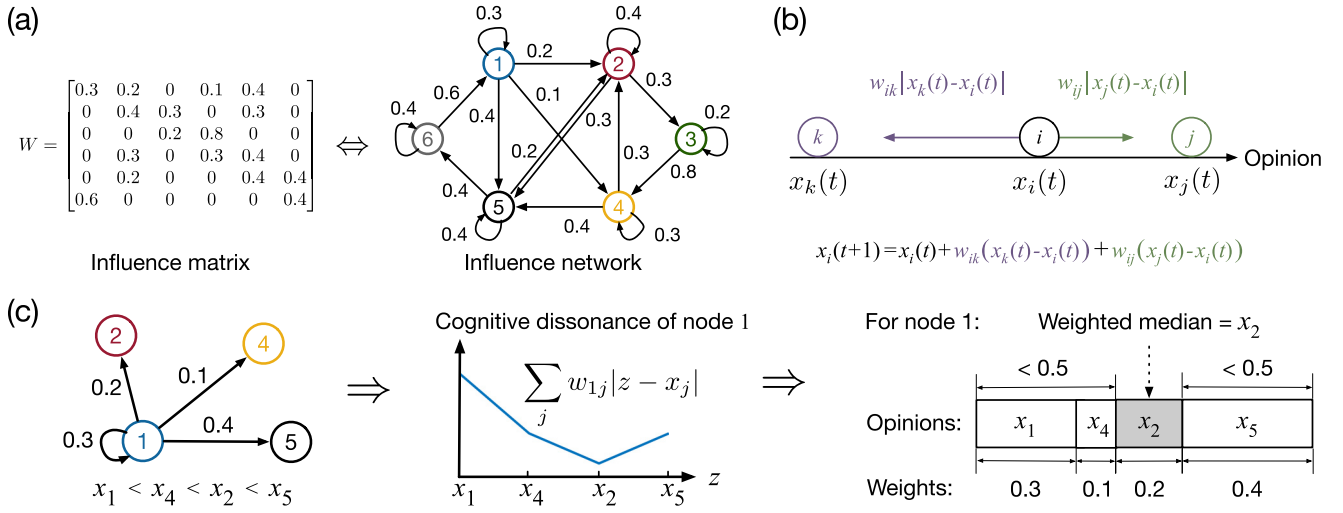

FIG. 1. Implications of the weighted-averaging and the weighted-median mechanisms. Panel (a) is an example of a 6×6 influence matrix and the corresponding influence network with 6 nodes. Panel (b) illustrates the underlying implication of the weighted-averaging opinion update. Panel (c) plots the cognitive dissonance function for node 1 in the influence network shown in panel (a), following the weighted-median mechanism. Node 1 updates its opinion by first sorting its social neighbors' opinions and picking the one such that the cumulative weights assigned to the opinions on its both sides are less than 0.5.

Moreover, the notion of opinion distance depends on the numerical representation of opinions, which could be arbitrary if the opinions are not numerical by nature.

Many extensions of the DeGroot models, as mentioned in the Introduction, can be considered as different efforts to remedy the above unrealistic feature by introducing additional model assumptions and parameters. For instance, Abelson [6] assumes that the interpersonal influences, i.e., the weights w_{ij}, decay with opinion distances. In a more recent paper [23], individuals with more extreme opinions are assumed to assign more weights to themselves. These modified averaging mechanisms, however, still lead to opinion consensus under mild network connectivity conditions. The Friedkin-Johnsen model [8] introduces individuals' persistent attachments to their initial conditions, which resist the attractions by others' opinions. However, this model almost surely generates persistent disagreement; i.e., consensus becomes almost impossible. Moreover, undesirably, the additional individual-level dynamics introduced to the model do not reflect any role of the influence network structure. In the biased-assimilation model [11], individuals process weighted averages of others' opinions in a highly nonlinear manner, by weighing confirming evidence more heavily than disconfirming evidence. The bounded-confidence models [9,10] assume that opinion attractiveness first increases proportionally with opinion distance and is then truncated to zero once the distance exceeds a preassumed threshold. These two models capture opinion polarization and opinion clustering, respectively, but are mathematically intractable due to their highly nonlinear dynamics.

B. Model derivation and setup

1. Model derivation

In this paper, we resolve the inherent unrealistic features of the weighted-averaging mechanism in a more fundamental way. Instead of further extending the DeGroot model, we propose the weighted-median mechanism as a new micro-foundation of opinion dynamics, in which opinion attractiveness and opinion distance are not intrinsically coupled. The derivation of the weighted-median mechanism is inspired by network games and the *cognitive dissonance* theory in psychology: Individuals experience cognitive dissonance by disagreeing with others and tend to reduce the dissonance by adjusting their opinions [19,20]. Such dissonance can be mathematically formalized in different ways [21] and the arguably most parsimonious form is

$$u_i(x_i, x_{-i}) = \sum_{j:\, w_{ij}>0} w_{ij}|x_i - x_j|^{\alpha}, \text{ for any individual } i,$$

where x_{-i} denotes the opinions of all the other individuals except i, and $\alpha > 0$ is an important model parameter. In this context, individuals' opinion updates can be modeled as the following best-response dynamics: for any i,

$$x_i(t+1) \in \operatorname{argmin}_{z\in\mathbb{R}} \sum_{j:\, w_{ij}>0} w_{ij}|z - x_j(t)|^{\alpha}. \qquad (2)$$

Although it might be overly assertive to claim that such "dissonance functions" really exist and are being minimized in human minds, the above framework does help derive opinion-update mechanisms with clear sociological interpretations. For example, due to the convexity of x^{α} for $x \geqslant 0$, $u_i(x_i, x_{-i})$ with $\alpha > 1$ implies that moving toward a distant opinion reduced more dissonance than moving toward a nearby opinion by the same distance. Namely, distant opinions are more attractive. In particular, $\alpha = 2$ results in the DeGroot model [24]. On the other hand, $\alpha < 1$ implies that nearby opinions are more attractive. In this paper, we adopt the neutral hypothesis $\alpha = 1$, which does not imply any preassumption on how opinion attractiveness is coupled with opinion distance. If necessary, one could incorporate any such coupling by assuming opinion-dependent weights $w_{ij}(x)$, which is a formidable research direction but out of the scope of this paper. It turns out that Eq. (2) with $\alpha = 1$ derives

the weighted-median mechanism, illustrated in Fig. 1(c) and formalized below. The detailed derivation is given in Appendix B.

2. Model setup

The weighted-median model is formalized as a discrete-time stochastic process. Given the influence matrix $W = (w_{ij})_{n\times n}$ and the initial condition $x(0) \in \mathbb{R}^n$, at each time $t+1$, an individual i is randomly activated and updates their opinion via the following weighted-median mechanism:

$$x_i(t+1) = \mathrm{Med}_i(x(t); W), \tag{3}$$

where $\mathrm{Med}_i(x(t); W)$ denotes the *weighted median* of the n-tuple $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ associated with the weights $(w_{i1}, w_{i2}, \ldots, w_{in})$, i.e., the ith row of the matrix W. The value of $\mathrm{Med}_i(x(t); W)$ is in turn given as follows: $\mathrm{Med}_i(x(t); W) = x^* \in \mathbb{R}$ if x^* satisfies

$$\sum_{j: x_j < x^*} w_{ij} \leqslant \frac{1}{2}, \quad \text{and} \quad \sum_{j: x_j > x^*} w_{ij} \leqslant \frac{1}{2}.$$

For generic weights W, $\mathrm{Med}_i(x(t); W)$ is unique. Otherwise, let $\mathrm{Med}_i(x(t); W)$ be the weighted median closest to $x_i(t)$, which again guarantees its uniqueness; see Appendix A for a detailed discussion.

3. Model applicability

The weighted-median operator is well defined as long as opinions are ordered. This prominent feature broadens the applicability of opinion dynamics models to multiple-choice issues with discrete and ordered options, which have not been extensively studied before by quantitative models. Debates and decisions about ordered multiple-choice issues are prevalent in reality. For example, in modern societies, many political issues are evaluated along one-dimensional ideology spectra and political solutions often do not lend themselves to a continuum of viable choices. At a fundamental level, the weighted-median mechanism is independent of numerical representations of opinions. Such representations may be nonunique and artificial for any issue where the opinions are not intrinsically quantitative. Obviously, a nonlinear opinion rescaling leads to major changes in the evolution of the averaging-based opinion dynamics. It is notable that the human mind often perceives and manipulates quantities in a nonlinear fashion, e.g., the perception of probability according to prospect theory [25].

The weighted-averaging mechanism dictates that each individual opinion changes as a linear superposition of the attractions by their neighbors' opinions. In the weighted-median mechanism, the attraction of an opinion to an individual is not independent but depends on what other opinions they are exposed to, how these opinions are ordered, and how the weights are assigned. In this sense, the weighted-averaging mechanism is more relevant in the persuasive events in dyadic or "gossip-like" situations. In order to apply the weighted-median mechanism to the scenario of dyadic conversations, one might need to assume that individuals adjust their opinions by simultaneously taking into account opinions they have been exposed to in some previous conversations.

III. EMPIRICAL VALIDATION

The weighted-median mechanism, derived from psychological theory and first principles, is also supported by empirical evidence. Analysis of an online experiment data set [22] indicates that median-based mechanisms enjoy significantly lower errors than averaging-based mechanisms in predicting individuals' opinion shifts under social influence. In each such experiment, 6 anonymous individuals answer 30 questions sequentially within tightly limited time. The questions are guessing the number of dots in a certain color in a given image; see Fig. 2(a) for one example. For each question, the 6 participants answer for 3 rounds. After each round, they see all the 6 participants' answers anonymously as feedback and possibly alter their own answers. The data set records the participants' answers in each round of the 30 questions. Such experiment design has several desired features. First, the questions being asked can be considered as judgmental issues, since there is no systematic way to solve them in limited time but subjective guessing. Second, since the participants see each other's answers anonymously, the underlying influence network is conceivably all-to-all with uniform weights. Namely, the experiment design rules out any other factor, e.g., prejudice or communication pattern, but focuses on the core comparison between median and average.

We randomly sampled 18 experiments from the data set, in which 71 participants answer all the 30 questions at each round. For each question, we predict the participants' third-round answers based on their second-round answers using the following hypotheses H1–H6 in pairs: Each participant i's answer $x_i(t+1)$ at the $(t+1)$th round is given by

H1: $x_i(t+1) = \mathrm{Median}(x(t))$,

H2: $x_i(t+1) = \mathrm{Average}(x(t))$,

H3: $x_i(t+1) = \gamma_i(t)x_i(t) + [1-\gamma_i(t)]\mathrm{Median}(x(t))$,

H4: $x_i(t+1) = \beta_i(t)x_i(t) + [1-\beta_i(t)]\mathrm{Average}(x(t))$,

H5: $x_i(t+1) = \tilde{\gamma}_i(t)x_i(1) + [1-\tilde{\gamma}_i(t)]\mathrm{Median}(x(t))$,

H6: $x_i(t+1) = \tilde{\beta}_i(t)x_i(1) + [1-\tilde{\beta}_i(t)]\mathrm{Average}(x(t))$,

where "Median" and "Average" means arithmetic median and average of all the six participants' answers, respectively. If there are two arithmetic medians, then $\mathrm{Median}(x(t))$ denotes the one closest to $x_i(t)$. Hypothesis H3 (H4 resp.) can be interpreted as the median (averaging resp.) mechanism with "inertia," while hypothesis H5 (H6 resp.) can be interpreted as the median (averaging resp.) mechanism with "prejudice." For hypotheses H3–H6, the parameters $\gamma_i(t)$, $\beta_i(t)$, $\tilde{\gamma}_i(t)$, and $\tilde{\beta}_i(t)$ are estimated by least-squares linear regression based on the participants' answers in the first 20 questions as the training set. Then these estimated parameters are used to predict their answers in the remaining 10 questions.

Using the above method for $t = 2$, we obtain $71 \times 30 = 2130$ predictions of the participants' 3rd-round answers by each of H1 and H2, and $71 \times 10 = 710$ predictions by each of H3–H6. Figure 2(b) shows the scatter plots between the observed answers and the predictions by H1 and H2. We compute the error rate for each prediction by H1–H6 as follows:

$$\text{error rate} = \frac{|\,\text{predicted value} - \text{observed value}\,|}{\text{observed value}}.$$

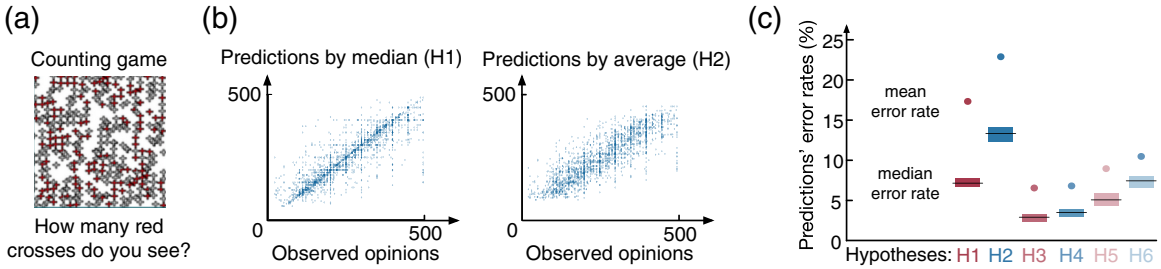

FIG. 2. Empirical analysis of the experiment data set [22]. Panel (a) shows an example of the counting game. Panel (b) shows the scatter plots between the participants' observed 3rd-round answers and the predictions by median (hypothesis H1) and average (hypothesis H2), respectively. Panel (c) is a visualized presentation of some indicative statistics of hypotheses H1–H6's prediction errors. The black bars indicate the medians of prediction error rates for each hypothesis, while the vertical ranges of the colored rectangles are the associated 95% confidence intervals, computed by the *binomial distribution method* [26]. The colored dots correspond to the means of the prediction error rates for each hypothesis.

Some indicative statistics of the prediction error rates for H1–H6 are visualized in Fig. 2(c) and are presented in detail in Fig. 1 of the Supplemental Material [4], according to which the median error rate of the predictions by median (H1) is 46.36% lower than that of the predictions by average (H2). In addition, for each pair of hypotheses, the median-based mechanism bears a lower median (and also mean) prediction error rate than the average-based counterpart. Notably, hypotheses H3 and H4 achieve remarkably low prediction errors by introducing individual inertia as additional parameters. Despite being useful for fitting the models, these parameters do not reflect intrinsic attributes of the individuals, nor are they stable over time. Hence, we refrain from such extensions and focus on the core issue, namely mean vs median. In addition, we also predict the participants' opinion shifts from the first round to the second round of each question. The results yield quantitatively similar conclusions; see Fig. 1 of the Supplemental Material [4].

IV. COMPARATIVE NUMERICAL STUDIES

Figure 3 shows a typical evolution of the weighted-median model on a lattice graph, from which some immediate observations can be obtained. First, unlike the DeGroot model, individuals in the weighted-median model do not always reach consensus but usually form into different opinion clusters.

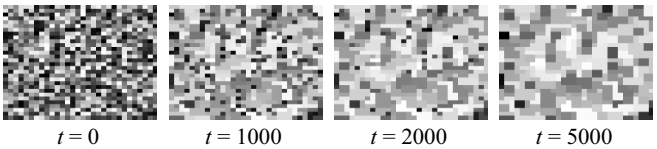

FIG. 3. One simulation of the weighted-median model on a 30×30 lattice graph. Each block is an individual and is bilaterally connected with all their adjacent blocks (not including the diagonally adjacent blocks). Each individual has a self-loop and uniformly assigns weights to all their neighbors including themself. Individuals' initial opinions are independently randomly generated according to the uniform distribution on $[-1, 1]$. The gray scale of each block is proportional to the absolute value of the individual's final opinion, i.e., their "degree of extremeness." After 5000 time steps, the evolution reaches an equilibrium.

Second, most of the extreme opinion holders (i.e., the dark gray blocks), initially scattered in the lattice, gradually convert to more moderate opinions. Namely, the typical effect of social influence on moderating the opinions of individuals in groups is still present but not overly strong as in the DeGroot model.

Further insights revealed by the weighted-median model are presented in the rest of this section. Particularly, we compare the behavior of the weighted-median model with some widely studied extensions of the DeGroot model, including the Friedkin-Johnsen model [8], the biased-assimilation model [11], and the networked bounded-confidence model [27], all with randomized parameters. Their mathematical forms and simulation setups are provided in Sec. III of the Supplemental Material [4].

A. Consensus probability and group structure

Since the weighted-median mechanism resolves the overly large attractions between distant opinions, effects of network structures on determining group consensus or disagreement naturally emerge. We investigate how group size and the clustering coefficient of the underlying influence network affect a group's probability of reaching consensus. We simulate different models on Watts-Strogatz small-world networks [28], whose structure is determined by three model parameters: the network size n, the average degree d, and the *rewiring probability* β. Specifically, the smaller β, the more clustered the network is. For the results shown in Figs. 4(a)–4(c), we fix the rewiring probability as $\beta = 1$ and estimate how the probability of reaching consensus changes with the network size n, under various fixed values of the average degree d. For the simulation results shown in Figs. 4(d)–4(f), we fix the network size as $n = 30$ and estimate how the probability of reaching consensus changes with the rewiring probability β, under various fixed values of the average degree d. For each model and network setup, the consensus probability is estimated over 5000 independent simulations.

As indicated by panels (a) and (d) of Fig. 4, in the weighted-median model, consensus is less likely to be achieved in larger or more clustered networks. This feature is consistent with previous empirical studies [29,30] and even everyday experience. On the other hand, predictions

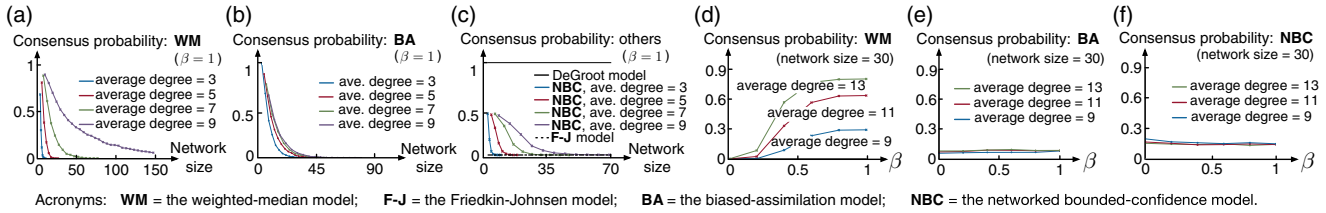

FIG. 4. Different models' predictions on how consensus probability depends on network size and clustering coefficient. These models are simulated on Watts-Strogatz small-world networks [28]. In panels (a)–(c), we fix the average individual degree d and the rewiring probability β, and plot how the consensus probability changes with the network size n. Panel (a) presents the predictions by the weighted-median model when $\beta = 1$. For other values of β, the results are qualitatively similar; e.g., see Fig. 2 of the Supplemental Material [4] for the results when $\beta = 0.3$. In panels (d)–(f), we fix n and d, and plot how the consensus probability changes with β. Panel (d) presents the predictions by the weighted-median model when $n = 30$. For other values of n, the results are qualitatively similar; e.g., see Fig. 2 of the Supplemental Material [4] for $n = 20$. Since the DeGroot and the Friedkin-Johnsen models lead to trivial predictions of either almost-sure consensus or almost-sure disagreement, their curves are not plotted in panels (d)–(f).

by other models are shown in panels (b), (c), (e), and (f) of Fig. 4: The Friedkin-Johnsen model almost surely leads to disagreement; the biased assimilation model and the networked bounded-confidence model capture the decreasing of consensus probability with network size, but do not show clear patterns regarding the relation between consensus probability and clustering coefficient.

B. Locations of extreme opinions in social networks

From Fig. 3, one could already see that extreme opinions in the lattice graph behave differently than moderate opinions. To further investigate how extreme opinions are located in social networks, we simulate different models 100 times independently on randomly generated scale-free networks [31] with 5000 nodes. The initial opinions are uniformly randomly generated from $[-1, 1]$ and opinions are classified into 4 categories; see Fig. 5(a). We estimate the in-degree centrality distributions for individuals holding different categories of opinions at the steady states of each simulation.

As Fig. 5(b) indicates, only in the weighted-median model, the in-degree distribution curves for different categories of opinions are clearly separated, and, moreover, the curve for extreme opinions decays the fastest as in-degree increases. That is, only the weighted-median model shows that extreme opinions tend to reside in peripheral areas of social networks. This feature is consonant with previous empirical, conceptual, and case studies [32–37], which explain opinion radicalization via social-influence processes and identify social marginalization as a key cause. Such a connection has barely been captured by quantitative opinion dynamics models and the weight-median mechanism provides perhaps the simplest explanation for it. To avoid the risk of bias due to the higher probability of being absolutely stubborn (self-weight $>1/2$) in the weighted-median model when the in-degree is small, we perform a second experiment on graphs without self-weights,

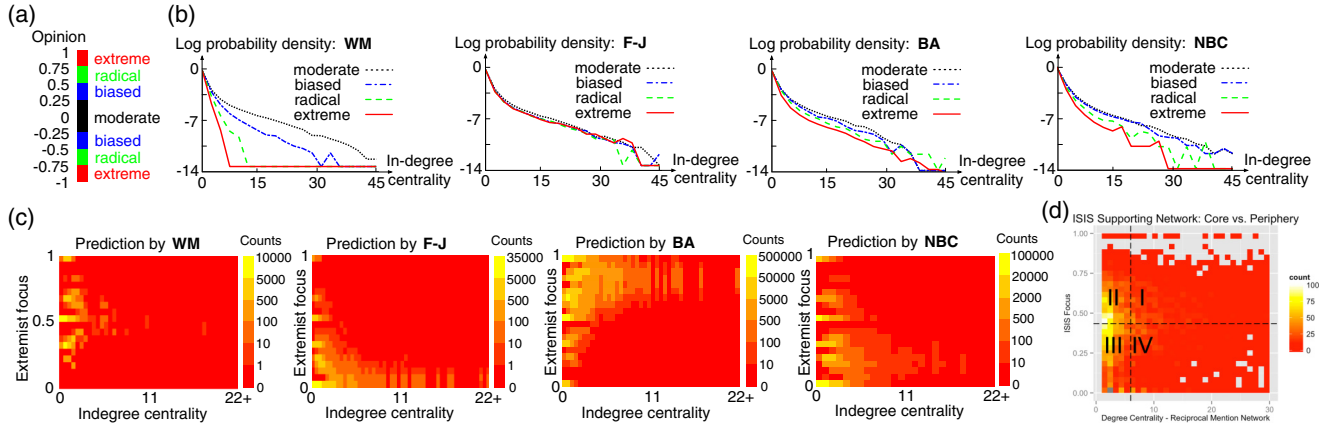

FIG. 5. Distributions of extreme opinions predicted by different models. Panel (a) is the categorization of opinions. Panel (b) shows different models' predictions on the in-degree centrality distributions for individuals holding different categories of opinions at the steady states. Panel (c) shows different models' predictions on the two-dimensional distributions, i.e., the in-degree and the extremist focus, for the extreme opinion holders at steady states. In each heat map, the last column "22+" records the number of extreme individuals with in-degrees larger than or equal to 22. Panel (d) is Fig. 5 in [38], licensed under Creative Commons CC0 public domain dedication (CC0 1.0). This figure plots the empirical distribution of randomly sampled Twitter users over the in-degree and the ISIS focus (the ratio of one's pro-ISIS social neighbors).

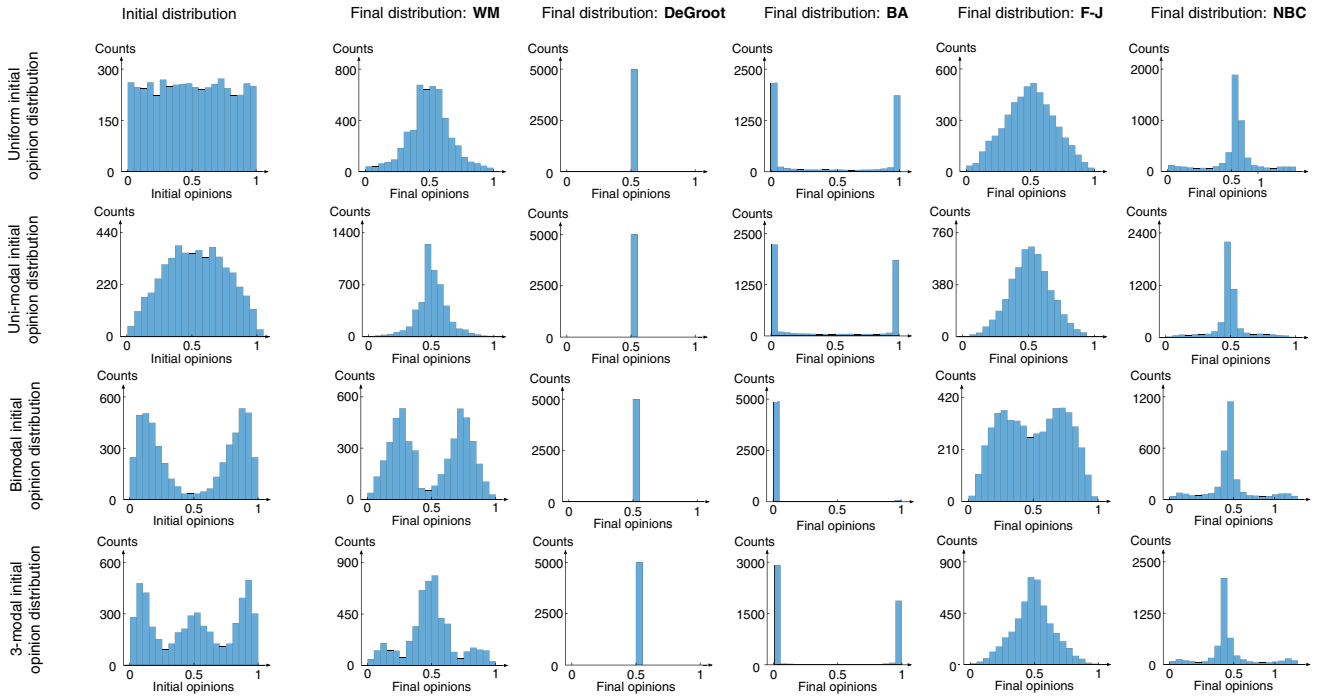

FIG. 6. Distributions of the initial opinions and the final opinions predicted by different models. All simulations are run on the same scale-free network with 5000 nodes and starting with the same randomly generated initial conditions.

and obtained similar results; see Fig. 4 of the Supplemental Material [4]. Simulations for closeness and between centrality or for different categorizations of opinions also lead to similar results and are presented in Figs. 3 and 5 of the Supplemental Material [4].

To obtain a deeper understanding of how extreme opinions are located in social networks, we further investigate the distribution of extreme opinions over two dimensions: the in-degree centrality and the *extreme focus*, i.e., the ratio of an individual's out-neighbors holding extreme opinions. For each model in comparison, we further simulate them on a scale-free network with 2000 nodes for 1000 times independently. To avoid the trivial cases that some individuals might stick to extreme opinions just because they have self-loops with weights larger than $1/2$, the simulations are conducted on networks without self-loops. For extreme opinion holders at the steady states, we compute their in-degree centrality and *extreme foci*, and then plot the corresponding two-dimensional distributions; see the heat maps in Fig. 5(c).

The heat map generated by the weighted-median model exhibits a clearly distinct pattern to those generated by the other models: In the weighted-median model, extreme opinion holders tend to have low in-degrees and their extreme foci concentrate around the value 0.5, which implies that they form into local clusters in peripheral areas of the networks. This observation indicates a mechanistic explanation for opinion radicalization among socially marginalized individuals: In social networks, some local clusters are formed by individuals with low centrality, which usually implies few social contacts. Inside those local clusters, if extreme opinions constitute the "mainstream," i.e., the weighted-median opinions, individuals will adhere to extreme opinions by yielding to social influence, due to the overwhelming social pressure and lack of diverse information sources. Remarkably, the heat map generated by the weighted-median model impressively resembles a real data set of the network among randomly sampled Twitter users, in which some users have their accounts suspended for posting pro-ISIS terrorism contents and are considered as extreme-opinion holders; see Fig. 5(d).

C. Steady public opinion distributions

Empirical evidence suggests public opinions do not only achieve persistent disagreement, but also form into certain steady distributions [15,39]. In fact, it has long been an open problem what mathematical models naturally lead to the emergence of various empirically observed steady public opinion distributions [40]. By simulating different models on a randomly generated scale-free network with 5000 nodes, we compare their predictions on the final steady opinion distributions, starting from various initial opinion distributions. Figure 6 shows a set of typical simulation results.

Among all the models in comparison, only the weighted-median model, without deliberately tuning any model parameters, naturally generates various types of empirically observed steady distributions of public opinions. Comparisons conducted on a small-world network [28] indicate similar conclusions and are provided in Fig. 6 of the Supplemental Material [4]. Namely, the weighted-median model provides perhaps the simplest explanation of the famous Abelson diversity puzzle [6] quoted in the Introduction, i.e., what models generate bimodal steady opinion distributions.

V. CONCLUSIONS AND FURTHER DISCUSSION

To sum up, with minimal assumptions, the weighted-median mechanism resolves the unrealistic proportionality between opinion attractiveness and opinion distance implied by the widely adopted weighted-averaging mechanism. Despite its simplicity in form, the weighted-median mechanism leads to higher accuracy in quantitatively predicting individuals' opinion shifts in an online experiment and captures various interesting real-world phenomena. While it is implausible for one single model to explain every aspect of real-world opinion evolution, all the aforementioned features support the weighted-median mechanism as a well-founded and expressive micro-foundation of opinion dynamics, especially for multiple-choice issues with discrete and ordered options.

A major limitation of the weighted-median mechanism is that no new opinion is created during the opinion updates. Therefore, it does not capture the behavior that individuals compromise at intermediate opinions. This limitation could be resolved by assuming that individuals move toward instead of directly taking their weighted-median opinions. In addition, one could make the model more realistic by considering state-dependent weights, e.g., by assuming that individuals with more extreme opinions become more stubborn, as in [23]. Moreover, many nontrivial extensions introduced to the classic DeGroot model can also be incorporated into the weighted-median mechanism, e.g., the presence of antagonistic relations [41], individual prejudice [8], the logical constraints among issues [14], and the issue alignments [42,43]. In addition, rigorous analysis of the dynamic behavior, e.g., convergence and graph-theoretic conditions for consensus/disagreement, of the weighted-median model and its variations would also be of important theoretical value.

ACKNOWLEDGMENTS

We acknowledge financial support from the National Natural Science Foundation of China under Grants No. 72131001, No. 72192804, and No. 12071465, the U.S. Army Research Laboratory and the U.S. Army Research Office under Grants No. W911NF-15-1-0577 and No. W911NF-16-1-0005, the RevealFlight Concerted Research Action (ARC) of the Federation Wallonie-Bruxelles, the Incentive Grant for Scientific Research (MIS) "Learning from Pairwise Data" of the F.R.S.-FNRS, as well as ETH Zurich funds. We also thank Dr. Nicolò Pagan, Dr. Noah E. Friedkin, Dr. Bary Pradelski, Dr. Bernhard Clemm Von Hohenberg, Dr. Aming Li, Dr. Ulrik Brandes, Dr. Christoph Stadtfeld, and Dr. Saverio Bolognani for their participation in discussions.

APPENDIX A: UNIQUENESS OF WEIGHTED MEDIAN

The notion of weighted median is formalized as follows:

Definition 1 (weighted median). Given any n-tuple of real numbers $x = (x_1, \ldots, x_n)$ and the associated n-tuple of nonnegative weights $w = (w_1, \ldots, w_n)$, where $\sum_{i=1}^{n} w_i = 1$, the *weighted median* of x, associated with the weights w, is denoted by $\mathrm{Med}(x; w)$ and defined as the real number $x^* \in \{x_1, \ldots, x_n\}$ such that

$$\sum_{i:\, x_i < x^*} w_i \leqslant 1/2, \quad \text{and} \quad \sum_{i:\, x_i > x^*} w_i \leqslant 1/2.$$

By carefully examining this definition, one could observe that, associated with certain specific weights w, there might exist multiple weighted medians of x satisfying the definitions above. Here we point out the following facts:

(i) The weighted median of x associated with w is unique if and only if there exists $x^* \in \{x_1, \ldots, x_n\}$ such that

$$\sum_{i:\, x_i < x^*} w_i < \frac{1}{2}, \quad \sum_{i:\, x_i = x^*} w_i > 0, \quad \text{and} \quad \sum_{i:\, x_i > x^*} w_i < 1/2.$$

In this case, x^* is the unique weighted median.

(ii) The weighted medians of x associated with w are NOT unique if and only if there exists $z \in \{x_1, \ldots, x_n\}$ such that $\sum_{i:\, x_i < z} w_i = \sum_{i:\, x_i \geqslant z} w_i = 1/2$. Among all these weighted medians of x, the smallest one, denoted by $\underline{x}^*$, satisfies

$$\sum_{i:\, x_i < \underline{x}^*} w_i < \frac{1}{2}, \quad \sum_{i:\, x_i = \underline{x}^*} w_i > 0, \quad \text{and} \quad \sum_{i:\, x_i > \underline{x}^*} w_i = \frac{1}{2},$$

while the largest weighted median, denoted by $\overline{x}^*$, satisfies

$$\sum_{i:\, x_i < \overline{x}^*} w_i = \frac{1}{2}, \quad \sum_{i:\, x_i = \overline{x}^*} w_i > 0, \quad \text{and} \quad \sum_{i:\, x_i > \overline{x}^*} < \frac{1}{2}.$$

Moreover, if there exists any $\hat{x} \in \{x_1, \ldots, x_n\}$ such that $\underline{x}^* < \hat{x} < \overline{x}^*$, then $\hat{x}$ is also a weighted median and it must hold that $\sum_{i:\, x_i = \hat{x}} w_i = 0$.

For generic weights, e.g., if $w_1, \ldots, w_n$ are independently randomly generated from some continuous probability distributions, the case in fact 2 almost never occurs since, with probability 1, no subset $\theta \in \{1, \ldots, n\}$ would satisfy exactly $\sum_{i \in \theta} w_i = 1/2$. Therefore, given generic weights w, the weighted median of x is unique.

Regarding the weighted-median opinion dynamics defined in Sec. II B, in order to avoid unnecessary mathematical complexity, we would like to make each individual's opinion update well defined and deterministic. Therefore, we slightly change the definition of weighted-median opinion when it is not unique according to Definition 1. Specifically, for any individual $i \in \{1, \ldots, n\}$, if at some point of time their weighted-median opinion is not unique, then let $\mathrm{Med}_i(x(t); W)$ be the weighted median that is the closest to $x_i(t)$. This setup guarantees the uniqueness of $\mathrm{Med}_i(x; W)$ since only one of the following three cases can occur when the weighted-median opinions are not unique:

(i) $x_i \leqslant \underline{x}^*$, where $\underline{x}^*$ is the smallest weighted median of x associated with the weights $(w_1, \ldots, w_n)$. In this case, $\mathrm{Med}_i(x; W) = \underline{x}^*$ is unique.

(ii) $x_i \geqslant \overline{x}^*$, where $\overline{x}^*$ is the largest weighted median of x associated with the weights $(w_1, \ldots, w_n)$. In this case, $\mathrm{Med}_i(x; W) = \overline{x}^*$ is unique.

(iii) $\underline{x}^* < x_i < \overline{x}^*$. According to fact 2 for the weighted median in last paragraph, this must imply that $\sum_{j:\, x_j = x_i} w_{ij} = 0$ and x_i is also a weighted median of x associated with the weights $(w_1, \ldots, w_n)$. Therefore, in this case, $\mathrm{Med}_i(x; W) = x_i$ is also unique.

By the nature of weighted median, for any given initial condition $x(0) = (x_{0,1}, \ldots, x_{0,n})^\top$, the solution $x(t)$ to the weighted-median opinion dynamics satisfies $x_i(t) \in \{x_{0,1}, \ldots, x_{0,n}\}$ for any $i \in \{1, \ldots, n\}$ and any $t \geqslant 0$. Moreover, along the weighted-median opinion dynamics, for each node i and at each time t,

$$x_i(t+1) > x_i(t) \quad \text{if and only if} \quad \sum_{j: x_j(t) > x_i(t)} w_{ij} > 1/2,$$

and

$$x_i(t+1) < x_i(t) \quad \text{if and only if} \quad \sum_{j: x_j(t) < x_i(t)} w_{ij} > 1/2.$$

APPENDIX B: DERIVATION OF THE WEIGHTED-MEDIAN MECHANISM FROM THE ABSOLUTE-VALUE COGNITIVE DISSONANCE FUNCTION

Consider an influence network $G(W)$ with n individuals. Given the opinion vector x, each individual i's cognitive dissonance generated by disagreeing with others can be modeled as

$$C_i(x_i, x_{-i}) = \sum_{j=1}^{n} w_{ij} |x_i - x_j|^\alpha,$$

and individual i's opinion update can be modeled as the best response to minimize the cognitive dissonance $C_i(x_i, x_{-i})$. That is, the updated opinion of individual i, denoted by x_i^+, satisfies

$$x_i^+ = \text{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{n} w_{ij} |z - x_j|^\alpha. \tag{B1}$$

We use equality here in the sense that the right-hand side of the equation above is unique for generic weights w_{ij}. The following proposition states the relation between the system given by Eq. (B1) and the weighted-median mechanism, when we set the parameter $\alpha = 1$.

Proposition 1 (weighted-median update as best-response dynamics). Given the row-stochastic influence matrix $W = (w_{ij})_{n \times n}$ and the vector $x = (x_1, \ldots, x_n)^\top$, the following statements holds: For any $i \in \{1, \ldots, n\}$,

(i) If there exists $x^* \in \{x_1, \ldots, x_n\}$ such that

$$\sum_{j: x_j < x^*} w_{ij} < \frac{1}{2}, \quad \text{and} \quad \sum_{j: x_j > x^*} w_{ij} < \frac{1}{2},$$

then

$$\text{Med}_i(x; W) = x^* = \text{argmin}_z \sum_{j=1}^{n} w_{ij} |z - x_j|.$$

(ii) If there does not exist such x^*, then the set

$$M_i(x; W) = \left\{ y \in \{x_1, \ldots, x_n\} \,|\, \sum_{j: x_j \leqslant y} w_{ij} \leqslant \frac{1}{2}, \right.$$
$$\left. \times \sum_{j: x_j > y} w_{ij} \leqslant \frac{1}{2} \right\}$$

is nonempty and

$$\begin{aligned} \text{Med}_i(x; W) &= \text{argmin}_{y \in M_i(x;W)} |y - x_i| \\ &\in [\inf M_i(x; W), \ \sup M_i(x; W)] \\ &= \text{argmin}_z \sum_{j=1}^{n} w_{ij} |z - x_j|. \end{aligned}$$

This proposition is a straightforward consequence of Definition 1 in this document and Lemma 3.1 in the paper by Sabo *et al.* [44].

[1] J. R. P. French, Jr., A formal theory of social power, Psychol. Rev. **63**, 181 (1956).

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

[3] A. V. Proskurnikov and R. Tempo, A tutorial on modeling and analysis of dynamic social networks. Part I, Annu. Rev. Control **43**, 65 (2017).

[4] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.4.023213 for a brief review of graph theory, the detailed simulation setups, as well as some supplementary empirical and simulation results.

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

[6] R. P. Abelson, Mathematical models of the distribution of attitudes under controversy, in *Contributions to Mathematical Psychology*, Vol. 14, edited by N. Frederiksen and H. Gulliksen (Holt, Rinehart, & Winston, 1964), pp. 142–160.

[7] D. Acemoglu, G. Como, F. Fagnani, and A. Ozdaglar, Opinion fluctuations and disagreement in social networks, Math. Oper. Res. **38**, 1 (2013).

[8] N. E. Friedkin and E. C. Johnsen, Social influence and opinions, J. Math. Soc. **15**, 193 (1990).

[9] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Mixing beliefs among interacting agents, Adv. Complex Syst. **3**, 87 (2000).

[10] R. Hegselmann and U. Krause, Opinion dynamics and bounded confidence models, analysis, and simulations, J. Artif. Soc. Soc. Simul. **5**(3), 2 (2002).

[11] P. Dandekar, A. Goel, and D. T. Lee, Biased assimilation, homophily, and the dynamics of polarization, Proc. Natl. Acad. Sci. USA **110**, 5791 (2013).

[12] G. Shi, A. Proutiere, M. Johansson, J. S. Baras, and K. H. Johansson, The evolution of beliefs over signed social networks, Oper. Res. **64**, 585 (2016).

[13] T. Kurahashi-Nakamura, M. Mäs, and J. Lorenz, Robust clustering in generalized bounded confidence models, J. Artif. Soc. Soc. Simul. **19**(4), 7 (2016).

[14] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov, Network science on belief system dynamics under logic constraints, Science **354**, 321 (2016).

[15] J. Lorenz, Modeling the evolution of ideological landscapes through opinion dynamics, in *Advances in Social Simulation 2015* (Springer, 2017), pp. 255–266.

[16] D. Acemoglu and A. Ozdaglar, Opinion dynamics and learning in social networks, Dynamic Games Appl. **1**, 3 (2011).

[17] A. V. Proskurnikov and R. Tempo, A tutorial on modeling and analysis of dynamic social networks. Part II, Annu. Rev. Control **45**, 166 (2018).
[18] 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**(4), 2 (2017).
[19] L. Festinger, *A Theory of Cognitive Dissonance* (Stanford University Press, 1957).
[20] D. C. Matz and W. Wood, Cognitive dissonance in groups: The consequences of disagreement, J. Pers. Soc. Psychol. **88**, 22 (2005).
[21] P. Groeber, J. Lorenz, and F. Schweitzer, Dissonance minimization as a microfoundation of social influence in models of opinion formation, J. Math. Soc. **38**, 147 (2014).
[22] C. Vande Kerckhove, S. Martin, P. Gend, P. J. Rentfrow, J. M. Hendrickx, and V. D. Blondel, Modelling influence and opinion evolution in online collective behaviour, PLoS ONE **11**, 1 (2016).
[23] V. Amelkin, F. Bullo, and A. K. Singh, Polar opinion dynamics in social networks, IEEE Trans. Autom. Control **62**, 5650 (2017).
[24] D. Bindel, J. Kleinberg, and S. Oren, How bad is forming your own opinion?, Games Econ. Behav. **92**, 248 (2015).
[25] D. Kahneman and A. Tversky, Prospect theory: An analysis of decision under risk, Econometrica **47**, 263 (1979).
[26] M. Bland, *An Introduction to Medical Statistics* (Oxford University Press, 2015).
[27] R. Parasnis, M. Franceschetti, and B. Touri, On graphs with bounded and unbounded convergence times in social Hegselmann-Krause dynamics, in *IEEE Conference on Decision and Control* (IEEE, 2019), pp. 6431–6436.
[28] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).
[29] A. P. Hare, A study of interaction and consensus in different sized groups, Am. Sociol. Rev. **17**, 261 (1952).
[30] Y. Yoo and M. Alavi, Media and group cohesion: Relative influences on social presence, task participation, and group consensus, MIS Q. **25**, 371 (2001).
[31] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[32] J. Woelfel, J. Woelfel, J. Gillham, and T. McPhail, Political radicalization as a communication process, Commun. Res. **1**, 243 (1974).
[33] C. McCauley and S. Moskalenko, Mechanisms of political radicalization: Pathways toward terrorism, Terror. Political Violence **20**, 415 (2008).
[34] J. R. Halverson and A. K. Way, The curious case of Colleen LaRose: Social margins, new media, and online radicalization, Media War Confl. **5**, 139 (2012).
[35] E. C. Hug, The role of isolation in radicalization: How important is it?, Master's thesis, Naval Postgraduate School, Monterey, CA, 2013.
[36] E. Tsintsadze-Maass and R. W. Maass, Groupthink and terrorist radicalization, Terror. Political Violence **26**, 735 (2014).
[37] S. Lyons-Padilla, M. J. Gelfand, H. Mirahmadi, M. Farooq, and M. V. Egmond, Belonging nowhere: Marginalization and radicalization risk among Muslim immigrants, Behav. Sci. Policy **1**(2), 1 (2015).
[38] M. C. Benigni, K. Joseph, and K. M. Carley, Online extremism and the communities that sustain it: Detecting the ISIS supporting community on twitter, PloS ONE **12**, e0181405 (2017).
[39] K. Janda, J. M. Berry, J. Goldman, D. Schildkraut, and P. Manna, *The Challenge of Democracy: American Government in Global Politics* (Cengage Learning US, 2019).
[40] N. E. Friedkin, The problem of social control and coordination of complex systems in sociology: A look at the community cleavage problem, IEEE Control Syst. **35**, 40 (2015).
[41] G. Shi, C. Altafini, and J. S. Baras, Dynamics over signed networks, SIAM Rev. **61**, 229 (2019).
[42] 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).
[43] 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).
[44] K. Sabo and R. Scitovski, The best least absolute deviations line—properties and two efficient methods for its derivation, ANZIAM J. **50**, 185 (2008).