

Extremism, segregation and oscillatory states emerge through collective opinion dynamics in a novel agent-based model

Beth M. Stokes, Samuel E. Jackson, Philip Garnett & Jingxi Luo

To cite this article: Beth M. Stokes, Samuel E. Jackson, Philip Garnett & Jingxi Luo (2022): Extremism, segregation and oscillatory states emerge through collective opinion dynamics in a novel agent-based model, *The Journal of Mathematical Sociology*, DOI: [10.1080/0022250X.2022.2124246](https://doi.org/10.1080/0022250X.2022.2124246)

To link to this article: <https://doi.org/10.1080/0022250X.2022.2124246>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 09 Oct 2022.



Submit your article to this journal [↗](#)



Article views: 639




View related articles [↗](#)



View Crossmark data [↗](#)

Extremism, segregation and oscillatory states emerge through collective opinion dynamics in a novel agent-based model

Beth M. Stokes^{a,b}, Samuel E. Jackson^c, Philip Garnett^d, and Jingxi Luo ^a

^aSchool of Mathematics, University of Birmingham, Edgbaston, Birmingham, UK; ^bDepartment of Mathematical Sciences, University of Bath, Bath, UK; ^cDepartment of Mathematical Sciences, Durham University, Durham, UK; ^dSchool for Business and Society, University of York, Heslington, York, UK

ABSTRACT

Using mathematics to model the evolution of opinions among interacting agents is a rich and growing field. We present a novel agent-based model that enhances the explanatory power of existing theoretical frameworks, corroborates experimental findings in social psychology, and reflects observed phenomena in contemporary society. Bespoke features of the model include: a measure of pairwise affinity between agents; a memory capacity of the population; and a generalized confidence bound called the interaction threshold, which can be dynamical and heterogeneous. Moreover, the model is applicable to opinion spaces of any dimensionality. Through analytical and numerical investigations, we study the opinion dynamics produced by the model and examine the effects of various model parameters. We prove that as long as every agent interacts with every other, the population will reach an opinion consensus regardless of the initial opinions or parameter values. When interactions are limited to be among agents with similar opinions, segregated opinion clusters can be formed. An opinion drift is also observed in certain settings, leading to collective extremisation of the whole population, which we quantify using a rigorous mathematical measure. We find that collective extremisation is likely if agents cut off connections whenever they move away from the neutral position, effectively isolating themselves from the population. When a population fails to reach a steady state, oscillations of a neutral majority are observed due to the influence exerted by a small number of extreme agents. By carefully interpreting these results, we posit explanations for the mechanisms underlying socio-psychological phenomena such as emergent cooperation and group polarization.

ARTICLE HISTORY



Received 3 August 2021
Revised 1 September 2022
Accepted 9 September 2022

KEYWORDS

Collective dynamics; belief formation; agent-based modeling; extremism

1. Introduction

Creating mathematical models to explain the dynamics of opinions is a research endeavor dating back to French (1956) and remains a frontier today (Castellano et al., 2009; Flache et al., 2017; Noorazar et al., 2020). By

CONTACT Jingxi Luo  j.luo.5@bham.ac.uk  School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

quantifying the interconnections among a social group, opinion dynamics models provide unique insights into the stimuli behind individuals evolving their views, and reveal mechanisms by which the group forms a consensus or fails to do so. This paper puts forward a new mathematical model of the emergence of extremism and segregation through opinion dynamics in a closed community. Interpreting the term ‘opinion’ broadly, we design the theory to be applicable to a variety of contexts including cultural evolution, language dynamics, economic games, animal societies, and so on.

One of the foundational models of opinion dynamics is due to DeGroot (1974), where a group of agents iteratively update their positions to weighted averages of other agents’ positions. Extending DeGroot’s model, Friedkin and Johnsen (1990) incorporates exogenous variables and other effects to simulate conflict and conformity behaviors. The DeGroot-Friedkin paradigm, which represents opinion updates as linear maps, remains influential today. In an insightful nonlinear generalization, Dandekar et al. (2013) shows that polarization can be a consequence of biased assimilation, a well-known psychological phenomenon where one is influenced most strongly by people with similar views (Lord et al., 1979). An important development of the theory introduces the effect of stubbornness: by allowing agents to have some attachment to their initial beliefs, it is found that the more stubborn agents hold more social power over time (Tian et al., 2021).

Bounded Confidence Models generalize the DeGroot-Friedkin paradigm by allowing each agent to interact only with agents whose opinions fall within some ‘confidence bound.’ Hegselmann and Krause (2002), for example, models the process of opinion fragmentation by updating to the average opinions of agents within the confidence bound. The model by Deffuant et al. (2000) has agents interacting in pairs and only adjusting their opinions if they fall within each other’s confidence bound, a process which leads to clustering. Other developments of Bounded Confidence Models have accounted for various factors that affect opinion dynamics, in order to align the models with social realities. Examples of such factors include group pressure and in-group favoritism (Alizadeh et al., 2015; Cheng & Yu, 2019); social feedback (Banisch & Olbrich, 2019); cultural complexity (Flache & Macy, 2011; Turner & Smaldino, 2018); repulsion (Huet et al., 2008; Stadtfeld et al., 2020); private opinions (Ye et al., 2019); and randomness in the confidence bounds (Kurahashi-Nakamura et al., 2016).

The Voter Model by Holley and Liggett (1975) is distinct in character from Bounded Confidence Models; it considers a set of ‘voters’ who change their opinions at random to that of one of their neighbors, without accounting for the opinion they currently hold. Similarly, in the Neutral Model by Bentley et al. (2011), agents copy an existing opinion at random from the population or, with a low probability, invent a new opinion. To augment the Voter Model, the concept of ‘inertia’ has been developed, allowing voters to have conviction

in their previously held opinions (Stark et al., 2008a, 2008b). Inertia has subsequently been applied to the Noisy Voter Model which, when combined with supportive interactions, produce strong drifts of opinions (Artime et al., 2018; Kononovicius, 2021).

Elsewhere, models with adaptive networks have been shown to promote the formation of echo chambers (Benatti et al., 2020); Weighted Balance Theory encompassing multiple weighted attitudes has been validated against American National Election Survey data (Schweighofer et al., 2020); a statistical physics approach has successfully integrated data from the 2008 US presidential election (Galesic & Stein, 2019); various graph theoretical approaches have been developed to investigate opinion convergence (M. Cao et al., 2008; Hendrickx et al., 2014; Nedić & Liu, 2016; Ren & Beard, 2005); and models with memory-based connectivity have been shown to produce opinion clusters (Mariano et al., 2020).

In this paper, we develop a novel agent-based model that is nonlinear and deterministic; it incorporates and improves elements from the DeGroot-Friedkin, Bounded Confidence, Voter, and other modeling frameworks, creating a new theory with significantly enhanced explanatory power. The model unifies and enhances many of the aforementioned socio-psychological factors or phenomena, for instance: the ‘stubbornness is power’ effect, biased assimilation, a dynamic confidence bound, inertia-induced opinion drift, and memory-based connectivity. Specific features, which are either important inclusions or upgrades from existing models, are as follows.

Many Bounded Confidence Models allow each agent to hold only one opinion. We propose that each agent holds and communicates multiple opinions at a time. Equivalently, we say that each agent holds an opinion with multiple components, which will be represented as components of a multi-dimensional vector. Two agents will interact if and only if the Euclidean distance between their opinions falls within some prescribed bound. Instead of taking discrete values as in Voter Models, the opinion vectors will take continuous values, allowing a greater variety of simulation outcomes to emerge. In a move akin to the introduction of inertia to Voter Models, we will define the concept of ‘memory capacity,’ describing the number of past states of the population that each agent takes into account when deciding whether or not to interact with another. The resulting non-Markovian process of opinion updating bears a stronger resemblance to real-world decision-making than its Markovian counterparts, and is reducible to the Markovian process if the memory capacity is minimized.

Concepts similar to this memory capacity have been examined in a small number of studies from which we have taken inspiration (Anderson & Ye, 2019). Most notably, the network-based model by Mariano et al. (2020) includes a memory state variable that reflects an agent’s opinion history and a parameter that controls how quickly an agent ‘forgets’ the past. In a similar

fashion, the connectivity of agents in our model is also dependent upon the history of opinions, but the model's realism is now improved by allowing the graph of connectivity to be directed: i need not influence j if j influences i , which is a sensible feature of social interactions in the real world.

Another field of research from which we have taken inspiration is the modeling of collective animal motion. The generalizability of collective motion models to the field of opinion dynamics has previously been addressed (M. M. Cao et al., 2008; Vicsek & Zafeiris, 2012), drawing parallels between convergence properties of self-synchronizing animal systems and quorum-finding mechanisms in social groups. The model of spontaneous order in bird flocks by Cucker and Smale (2007), in particular, has strongly influenced this paper (see, Section 2).

This paper's scope and structure is as follows. We outline the principal ideas behind our model and present its mathematical formulation in Section 2, followed by a key theorem on consensus formation. A novel concept of pairwise 'affinity' will be introduced which describes how closely aligned two agents are in their recent history and is parameterized by the aforementioned memory capacity. In Section 3, we use numerical simulations to explore the phenomena of clustering, opinion drift, and extremisation, exploring real-world implications of the simulation results in the context of cooperative networks. We also examine the natural emergence of extreme views from the system when each agent's threshold for interaction evolves with their opinions. Section 3.4 proves that the model admits periodic solutions under certain conditions that we specify explicitly, which is a particularly intriguing feature. The oscillatory dynamics that arise when the system fails to converge are investigated in detail. Finally, we will draw conclusions and discuss future directions in Section 4.

2. The model and preliminary analysis

In the model, an 'opinion' has any number of components, which will be represented as coordinates in D -dimensional space. For example, an agent's preference for sweet or savory popcorn can be one dimension, while their conservative or liberal politics may be another. It is assumed that, in general, an agent evolves their entire opinion – all dimensions included – as a whole, rather than evolve the components independently. Thus, the opinion space is \mathbb{R}^D , with the origin representing the opinion that is neutral in every dimension. The Euclidean distance from the origin to an opinion is a measure of that opinion's 'extremeness.'

We consider $N \geq 2$ agents whose opinions are represented by D -dimensional real-valued vectors, $\mathbf{v}_1(t), \mathbf{v}_2(t), \dots, \mathbf{v}_N(t)$, where $t = 0, 1, 2, \dots$ are discrete times. To update their opinion at each time, every agent tries to

align with a select group of other agents. More precisely, every pair of agents, i and j , share an *affinity*, $a_{ij}(t)$, which we define in [Section 2.1](#); every agent has a *threshold*, $\rho_i(t)$, and agent i will try to align with agent j at time t if and only if $a_{ij}(t) > \rho_i(t)$. The model is therefore of the bounded confidence type, except pairwise influence is determined not only by the opinion difference between the pair, but by the more sophisticated measure of pairwise affinity which involves a collective *memory capacity* of the population. We now proceed to detail the mathematical model in [Section 2.1](#), before proving a result on the consensus formation in [Section 2.2](#).

2.1. Mathematical formulation

A vital element of the model is the pairwise affinity, $a_{ij}(t)$, between agents i and j , which we require to possess several properties. Firstly, the affinity must be symmetric ($a_{ij} = a_{ji}$). Secondly, it should always take positive values no larger than 1, with higher values indicating that i and j are more ‘alike.’ Thirdly, the affinity should depend not only on the opinion difference between i and j at the current time, but also on a recent history of opinion differences. This memory property represents an important generalization from existing bounded confidence models. For an affinity measure that satisfies all these requirements, we take

$$a_{ij}(t) = \frac{1}{\left(1 + \sum_{\tau=0}^t w(\tau; t, \mu) \|\mathbf{v}_j(\tau) - \mathbf{v}_i(\tau)\|^2\right)^{1/2}}, \quad (1)$$

where $w(\tau; t, \mu)$ is a weight function given by

$$w(\tau; t, \mu) = \begin{cases} 1, & \text{if } \tau > t - \mu, \\ 0, & \text{if } \tau \leq t - \mu, \end{cases} \quad (2)$$

and we have introduced the integer parameter $\mu \geq 1$, which we call the memory capacity, representing the number of steps (including the present step) that every agent takes into account when calculating affinities. If the current time $t < \mu$, then the sum will be over all time-steps. If $\mu \leq t$, then w assigns unit weight to times from $t - \mu + 1$ to the present while assigning zero weight to all prior times. We say that opinion differences prior to $t - \mu + 1$ ‘drop out’ of memory. The sum in the denominator of a_{ij} is a weighted sum of the square of opinion difference over the most recent μ time-steps, where $\|\cdot\|$ denotes the Euclidean norm and so $\|\mathbf{x} - \mathbf{y}\|$ denotes the Euclidean distance between \mathbf{x} and \mathbf{y} . Although we have chosen the Euclidean (l_2) norm as the distance measure, it is worth noting that other choices of norm would also be suitable. In particular, the convergence results in [Section 2.2](#) remain true for the l_1 and l_∞ norms (see details in [Lemma 2.1](#) and [Proposition 2.2](#)), meaning

that the same consensus behavior would be observed given an alternative norm. We choose the Euclidean norm as it provides the most moderate measure of distance out of the three candidates since, in general, $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \|\cdot\|_1$.

The choice of weight function is one of the simplest, yet it fully captures the affinity's memory property. The addition of 1 in the denominator of a_{ij} ensures that the affinity never exceeds 1, and $a_{ij} = 1$ (maximum affinity) if and only if: either $i = j$ (one has maximum affinity with oneself), or i and j have held exactly the same opinion in the most recent μ time-steps including current time. The fact that $a_{ij} > 0$ means no two agents will ever share exactly zero affinity, no matter how much their opinions differ. Moreover, if two agents i and j hold their opinions fixed, with $\mathbf{v}_i(\tau) \neq \mathbf{v}_j(\tau)$, then their affinity a_{ij} decreases over time, representing the tendency for people to become less connected if they keep disagreeing with each other.

With the affinity measure in place, and with every agent having some threshold, ρ_i (to be defined), we let the opinions $\mathbf{v}_i(t)$ in the population evolve as follows.

$$\text{For all } i : \mathbf{v}_i(t+1) = \mathbf{v}_i(t) + \frac{1}{Q_i(t; \rho_i)} \sum_{j=1}^N c_{ij}(t; \rho_i) a_{ij}(t) (\mathbf{v}_j(t) - \mathbf{v}_i(t)), \quad (3)$$

where

$$c_{ij}(t; \rho_i) = \begin{cases} 1, & \text{if } a_{ij}(t) > \rho_i, \\ 0, & \text{if } a_{ij}(t) \leq \rho_i, \end{cases} \quad (4)$$

$$Q_i(t; \rho_i) = \sum_{k=1}^N c_{ik}(t; \rho_i). \quad (5)$$

We say that i 'listens to' j (or j influences i) at time t if $c_{ij}(t; \rho_i) = 1$, and Equation (4) expresses the fact that i listens to j if and only if their affinity exceeds i 's threshold. Thus, $Q_i(t; \rho_i)$ is simply the number of agents that i listens to (including i , since every agent is self-influencing with $a_{ii} = 1$). According to Equation (3), the amount by which agent i adjusts their opinion at each time is a weighted average of relative opinions from i to all agents that i listens to, with weights determined by affinities. By construction, every agent's self-confidence, $1 - \frac{1}{Q_i} \sum_{j \neq i} c_{ij} a_{ij}$, and all the other weights, $\frac{1}{Q_i} c_{ij} a_{ij}$ for $j \neq i$, add up to 1, meaning that the system's transition matrix is right-stochastic. Note that c_{ij} may not be symmetric: $a_{ij} > \rho_i$ does not imply $a_{ji} > \rho_j$, since ρ_i and ρ_j may be different (even though $a_{ji} = a_{ij}$). In other words, the fact that i listens to j does not necessarily mean j listens to i , since they may have different thresholds. Note also that if $\rho_i = 0$ for all i , then in the infinite-memory limit

($\mu \rightarrow \infty$), the model becomes analogous to Cucker and Smale (2007) which investigates the synchronization of bird flocks.

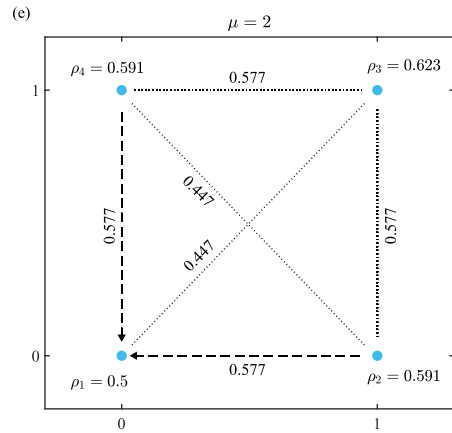
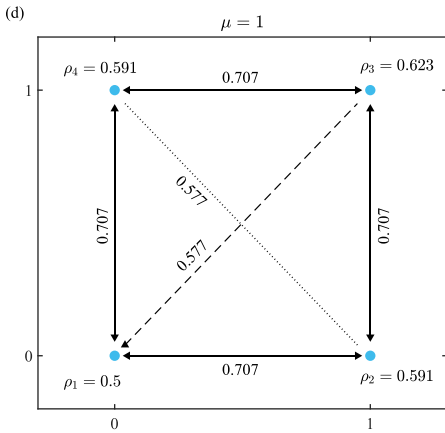
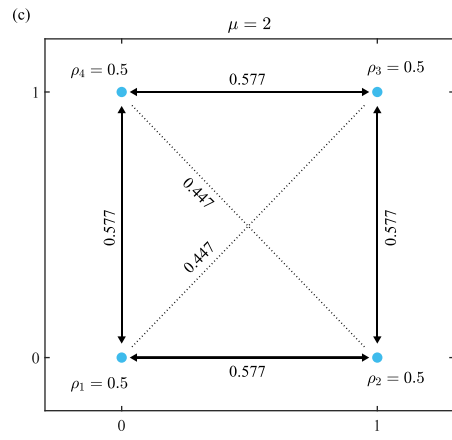
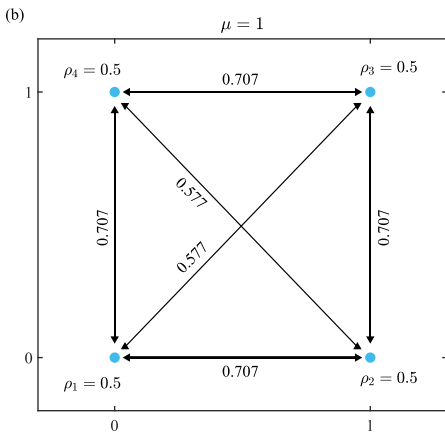
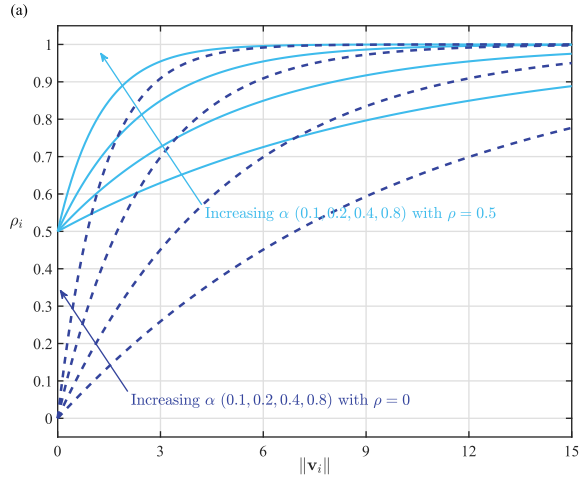
In this paper we consider two ways to assign thresholds to the population:

- (1) Every agent is equally susceptible to change at all times: $\rho_i(t) = \rho$ for all i and all t , where $0 \leq \rho < 1$ is some prescribed constant which we call the *universal threshold*. This is the simplest way to assign thresholds.
- (2) Every agent's threshold evolves over time, in such a way that the more extreme their opinion, the higher their threshold and hence the less susceptible they are to change:

$$\rho_i(t) = \rho + (1 - \rho) \left(1 - e^{-\alpha \|\mathbf{v}_i(t)\|} \right). \quad (6)$$

This assumption is grounded in empirical observations (Kozitsin, 2020; Lord et al., 1979; Tian et al., 2021). In Equation (6), ρ_i is a strictly increasing function of the extremeness $\|\mathbf{v}_i(t)\|$ of agent i 's opinion, and $\alpha > 0$ is a constant *reinforcement rate* determining how sharply one's threshold increases as one's opinion becomes more extreme (see, Figure 1a). The larger α is, the more sharply one's threshold increases. Note that $\rho_i \rightarrow 1$ as $\|\mathbf{v}_i(t)\| \rightarrow \infty$, and $\rho_i = \rho$ if $\|\mathbf{v}_i(t)\| = 0$. We therefore interpret ρ as a *baseline threshold*: the threshold that one has when one's opinion is entirely neutral. Note also that in the limit $\alpha \rightarrow 0$, we recover the uniformly constant $\rho_i(t) = \rho$.

Recall that the pairwise affinity $a_{ij}(t)$ decreases over time unless i and j adjust their opinions to align with each other. The implication of this fact at the population level is that, unless $\rho_i(t) = 0$ for all i and all t , then as agents fail to 'come together' in their opinions, the network of interpersonal influences will become less connected over time. Equivalently, given two systems with identical opinion histories and different memory capacities, the system with the larger memory capacity has a less connected network of influences. An illustration of this phenomenon is presented in Figure 1. Assume that agents 1–4 have held their two-dimensional opinions fixed (e.g., due to external influences) for at least μ steps, at $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$, respectively. Then, the Euclidean distance between any pair's opinions is fixed at either 1 or $\sqrt{2}$. Calculating the pairwise affinities by Equation (1), we find $a_{ij} = 1/\sqrt{2}$ or $1/\sqrt{3}$ if $\mu = 1$, and $a_{ij} = 1/\sqrt{3}$ or $1/\sqrt{5}$ if $\mu = 2$. Thus, in case $\rho_i = 0.5$ for all i : if $\mu = 1$ then everyone listens to everyone else, since $1/\sqrt{2} > 1/\sqrt{3} > 0.5$ (Figure 1b); if $\mu = 2$ then the most distant pairs of agents, $\{1, 3\}$ and $\{2, 4\}$, do not communicate, since $1/\sqrt{3} > 0.5 > 1/\sqrt{5}$ (Figure 1c). On the other hand, in case individual thresholds evolve from baseline $\rho = 0.5$ according to Equation (6) with reinforcement $\alpha = 0.2$: if $\mu = 1$ then $\{2, 4\}$ do not communicate while $\{1, 3\}$ communicate uni-directionally, the symmetry being broken due to the heterogeneous thresholds (Figure 1d); if $\mu = 2$ then the only



$i \longleftrightarrow j$: i and j listen to each other
 $i \dashleftarrow j$: i listens to j but not vice versa
 $i \cdots \cdots j$: i and j do not communicate

communications are agent 1 listening to agents 2 and 4, since no other affinity exceeds the relevant threshold (Figure 1e). This simple example shows that the connectivity of the system depends sensitively on multiple factors: the opinions, thresholds and memory capacity.

In the language of complexity theory, the system is ‘simple’ if the universal threshold ρ is close to 0 or 1, being always highly connected in the former case and always barely connected in the latter; and the complexity is maximized if ρ takes intermediate values since the connectivity can fluctuate greatly over time, as the example above demonstrates. In the simplest case, $\rho_i(t) = 0$ (meaning everyone listens to everyone else all the time), we establish analytically in Section 2.2 that any population is guaranteed to form a consensus over time, meaning $\mathbf{v}_i(t)$ converge to some common value for all i . In any other case ($\rho_i(t) = \rho > 0$ or Equation (6)), the system is not analytically tractable, so we will investigate the opinion dynamics using numerical methods in Section 3.

2.2. Sufficient conditions for convergence and for consensus

We say that the system converges to a steady state if and only if, for all i , there exists some constant \mathbf{v}_{*i} such that $\mathbf{v}_i(t) \rightarrow \mathbf{v}_{*i}$ as $t \rightarrow \infty$. We say that the system converges to consensus if and only if there exists some common constant \mathbf{v}_* such that, for all i , $\mathbf{v}_i(t) \rightarrow \mathbf{v}_*$ as $t \rightarrow \infty$. Whenever the system converges to a steady state but not to consensus, we say that the system converges to segregation. In this section, we consider the model (1)–(5) with some universal threshold $\rho_i(t) = \rho$, in which case we show that the system always converges to a steady state, and establish the following sufficient condition for consensus: $\rho < \rho_*$, where ρ_* is a critical value we will determine explicitly.

In Lorenz (2005), it was shown that any system $\mathbf{V}(t+1) = \mathbf{M}(t)\mathbf{V}(t)$ where $\mathbf{V} \in \mathbb{R}^{N \times D}$ converges to a steady state if three conditions are met: all agents have positive self-confidence ($m_{ii} > 0$); confidence is mutual ($m_{ij} > 0 \Leftrightarrow m_{ji} > 0$); and there exists some $\delta > 0$ such that the time-sequence \mathcal{M}_t , defined by $\mathcal{M}_t = \min_{i,j} \{m_{ij}(t) > 0\}$, satisfies $\mathcal{M}_t > \delta$. By expressing (3) in

Figure 1. An example of the interplay between agents’ thresholds, pairwise affinities, and pairwise influences. Panel (a) shows the evolving threshold ρ_i of some agent i as a function of $\|\mathbf{v}_i\|$, as per Equation (6), with reinforcement rate $a = 0.1, 0.2, 0.4, 0.8$. Panels (b–e) are snapshots of the connections among four agents with two-dimensional opinions $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$ which we assume, for this illustration, have been held fixed for at least μ time-steps. The arrows denote the three distinct types of pairwise influence. The numbers that annotate the arrows are the pairwise affinities calculated by Equation (1) to 3dp precision. Panels (b,c) have universal threshold $\rho_i = 0.5$ for all i ; panels (d,e) have evolving thresholds with baseline $\rho = 0.5$ and reinforcement rate $a = 0.2$, calculated by Equation (6) to 3dp precision. The memory capacity is $\mu = 1$ in (b,d) and $\mu = 2$ in (c,e).

matrix form (putting the opinion of agent i in row i of \mathbf{V}), we find the diagonal elements

$$m_{ii} = 1 - \frac{1}{Q_i} \sum_{j \neq i} c_{ij} a_{ij} \geq 1 - \frac{Q_i - 1}{Q_i} \geq \frac{1}{N} > 0, \quad (7)$$

implying that the current model meets the “positive self-confidence” condition. For the “mutual confidence” condition, we look at the off-diagonal elements

$$m_{ij} = \frac{1}{Q_i} c_{ij} a_{ij} \geq 0, \quad i \neq j, \quad (8)$$

and note that equality holds if and only if $c_{ij} = 0$. Since $\rho_i = \rho$ for all i by assumption and $a_{ij} = a_{ji}$ by definition, we deduce from (4) that $c_{ij} = c_{ji}$, and therefore $m_{ij} = 0$ if and only if $m_{ji} = 0$. Lorenz’s second condition is thus met.

To show that Lorenz’s third and final condition is met, it suffices to find a positive lower bound for all positive off-diagonal m_{ij} for all time. To that end, note that

$$\text{for all } i \neq j \text{ such that } m_{ij} > 0 : \quad m_{ij} \geq \frac{1}{N} a_{ij}. \quad (9)$$

We therefore seek some constant $\rho_* > 0$ such that $a_{ij} \geq \rho_*$ for all i, j and all time, and we do so through the following lemma.

Lemma 2.1. Consider the system $\mathbf{V}(t+1) = \mathbf{M}(t)\mathbf{V}(t)$, where $t \in \{0, 1, 2, \dots\}$, $\mathbf{V} \in \mathbb{R}^{N \times D}$ and $\mathbf{M} \in \mathbb{R}^{N \times N}$. Let $R_t = \max_i \{\|\mathbf{r}_i(\mathbf{V}(t))\|\}$, where $\mathbf{r}_i(\mathbf{V})$ denotes the i^{th} row of \mathbf{V} and $\|\cdot\|$ the Euclidean norm. If $\|\mathbf{M}(t)\|_\infty \leq 1$ for all t , then the sequence $R_{t \geq 0}$ is non-increasing.

Proof. It is an established fact of linear algebra that R_t equals the $(2, \infty)$ -norm of $\mathbf{V}(t)$ (Lewis, 2010):

$$\max_i \{\|\mathbf{r}_i(\mathbf{V}(t))\|\} = \sup \left\{ \frac{\|\mathbf{V}(t)\mathbf{x}\|_\infty}{\|\mathbf{x}\|} : \mathbf{x} \neq \mathbf{0} \right\} = \|\mathbf{V}(t)\|_{2,\infty} \quad (10)$$

Combining the submultiplicity of induced norms:

$$\|\mathbf{M}\mathbf{V}\|_{2,\infty} \leq \|\mathbf{M}\|_\infty \|\mathbf{V}\|_{2,\infty} \text{ for any } \mathbf{M} \text{ and } \mathbf{V} \text{ where } \mathbf{M}\mathbf{V} \text{ is defined,} \quad (11)$$

with the assumption that $\|\mathbf{M}(t)\|_\infty \leq 1$, yields

$$\|\mathbf{V}(t+1)\|_{2,\infty} = \|\mathbf{M}(t)\mathbf{V}(t)\|_{2,\infty} \leq \|\mathbf{V}(t)\|_{2,\infty} \quad (12)$$

as required. \square

Note that Lemma 2.1 is a general result applicable to any system whose transition matrix has absolute row sums no larger than 1. Note also that Lemma 2.1 still holds if the l_1 or l_∞ norm were used to define the pairwise affinity and hence R_t . This is due to the following facts which are analogous to Equation (10; Lewis, 2010): $\max_i \{ \| \mathbf{r}_i(\mathbf{V}(t)) \|_1 \} = \| \mathbf{V}(t) \|_{\infty, \infty}$, $\max_i \{ \| \mathbf{r}_i(\mathbf{V}(t)) \|_\infty \} = \| \mathbf{V}(t) \|_{1, \infty}$. To apply Lemma 2.1 to the current model (1)–(5), we simply let $\mathbf{r}_i(\mathbf{V}(t)) = \mathbf{v}_i(t)$ and \mathbf{M} be the matrix with elements given by (7)–(8). Then, $R_t = \max_i \| \mathbf{v}_i(t) \|$ is the maximum Euclidean magnitude of all opinions at time t . Note that Lemma 2.1 implies the set of opinions is always ‘shrinking’ in the sense that R_t is non-increasing, *regardless of* $\rho_i(t)$. A useful interpretation of this result is that the opinions ‘shrink’ due to the agents interacting under attractive forces only, with no repulsive forces involved. Now, let

$$\rho_* = \frac{1}{(1 + 4\mu R_0^2)^{1/2}}, \quad (13)$$

then for all i, j, t , we have $\| \mathbf{v}_j(t) - \mathbf{v}_i(t) \| \leq 2R_0$ and hence

$$a_{ij}(t) \geq \rho_*, \quad (14)$$

which implies that Lorenz’s final condition for convergence is met. We are now ready to state the main result of the section.

Proposition 2.2. Consider a population of agents $i = 1, 2, \dots, N$, evolving their opinions $\mathbf{v}_i(t) \in \mathbb{R}^D$ according to the model (1)–(5), with some universal threshold $\rho_i(t) = \rho$ for all i, t .

- (1) Given any initial condition, the opinions converges to some steady state: $\lim_{t \rightarrow \infty} \mathbf{v}_i(t) = \mathbf{v}_{*i}$ for all i .
- (2) Given any initial condition and any $\rho < \rho_*$, where ρ_* is given by (13) with $R_0 = \max_i \{ \|\mathbf{v}_i(0)\| \}$, the opinions converge to a consensus: $\lim_{t \rightarrow \infty} \mathbf{v}_i(t) = \mathbf{v}_*$ for some common \mathbf{v}_* . Moreover,

$$\mathbf{v}_* = \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i(0), \quad (15)$$

is the initial mean opinion of the population.

Proof. Part (1) is already proven, by showing that the system meets all of Lorenz’s convergence criteria. For part (2), it follows immediately from (14) and $\rho < \rho_*$ that $a_{ij}(t) > \rho$ for all i, j, t , which implies $c_{ij}(t) = 1$ for all i, j, t . That is, every agent listens to every other for all time. The system therefore simplifies to

$$\mathbf{V}(t+1) = \mathbf{V}(t) + \frac{1}{N}(\mathbf{A}(t) - \mathbf{B}(t))\mathbf{V}(t), \quad (16)$$

where $\mathbf{r}_i(\mathbf{V}) = \mathbf{v}_i$, \mathbf{A} is the $N \times N$ matrix with elements a_{ij} , and \mathbf{B} is the $N \times N$ diagonal matrix with elements

$$b_{ii} = \sum_{j=1}^N a_{ij}, \quad b_{ij} = 0 \text{ if } i \neq j. \quad (17)$$

Define the *initial mean matrix*, $\bar{\mathbf{V}}_0$, with rows

$$\mathbf{r}_1(\bar{\mathbf{V}}_0) = \mathbf{r}_2(\bar{\mathbf{V}}_0) = \cdots = \mathbf{r}_N(\bar{\mathbf{V}}_0) = \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i(\mathbf{V}(0)). \quad (18)$$

It takes a straightforward calculation to show $\mathbf{A}\bar{\mathbf{V}}_0 = \mathbf{B}\bar{\mathbf{V}}_0$, which means that $\mathbf{V}(t) = \bar{\mathbf{V}}_0$ is a steady state of the system. Since convergence is already established, and since limits in $\mathbb{R}^{N \times D}$ are unique, it follows that $\mathbf{V}(t) = \bar{\mathbf{V}}_0$ is precisely the state to which the system converges. \square

Recall that Lemma 2.1 holds if the system were defined using the l_1 or l_∞ norm instead of the Euclidean norm; thus, the same can be said for Proposition 2.2. To apply Proposition 2.2, consider initial opinions in $\mathbb{R}^{N \times D}$ where each column is N values sampled from an independent standard normal distribution (as is done in Section 3). In this case, it is reasonable to assume that all initial opinions fall within a sphere of radius $R_0 = \sqrt{9D}$. Proposition 2.2 then implies that consensus is guaranteed whenever

$$\rho < \frac{1}{\sqrt{1 + 36\mu D}} \approx \frac{1}{6\sqrt{\mu D}}. \quad (19)$$

To conclude the section, we note that other than Lorenz (2005), different convergence criteria for opinion dynamics systems exist in the literature, for example, in Blondel et al. (2005). Hendrickx et al. (2014) proved general results concerning the existence of models that guarantee average consensus, using a graph-theoretic approach. Here, Proposition 2.2 can be stated in graph theoretical terms because a graph that represents the agents as nodes and pairwise influences as edges is indeed connected and undirected if $\rho < \rho_*$. Overall, the model with $\rho < \rho_*$ provides a mechanism for how an interacting social group can find common ground from initial disagreements, through a process of collective assimilation.

3. Numerical simulations: results and discussions

In this section, we investigate how the opinion dynamics are affected by the model parameters, focusing mainly on the threshold, ρ_i . Two cases are considered: $\rho_i = \rho$ constant for all time (in Section 3.2), and ρ_i evolving with individual opinions according to Equation (6; in Section 3.3). We also examine the effects of different dimensions (D) of the opinion vector, and the memory capacity (μ).

3.1. Methodology

Numerical simulations were run in MATLAB (code available at https://github.com/bmstokes/belief_dynamics/releases/tag/v.1.0.0 under the Mozilla Public License 2.0). Every simulation is for $N = 100$ agents. The D components of every initial opinion are drawn randomly from D independent standard normal distributions. We adopt this simplistic initialization method on the basis of its universality: in the absence of any specific context, it is reasonable to consider a normally distributed initial population of opinions, which can then be standardized to enable comparison across the dimensions. We do, however, acknowledge that some real-world scenarios may not be well represented by this initial sampling; we will present some examples in Section 4 and discuss how they can be investigated using the model in future work. The simulation results presented here serve to demonstrate the power of the model: rich and varied phenomena emerge from simple initial conditions and hold strong explanatory power, as we will demonstrate in the following sections. Similarly rich phenomena are bound to emerge from more complex initial states, which any user of the model is always free to specify. For the present study, under each value of D , we generated 1000 distinct initial states; and for each set of other parameter values (some combination of μ, ρ, α , see, Table 1), we ran 1000 simulations using that common set of initial states, allowing us to control for the parameters μ, ρ and α .

The system is in a steady state from time t_0 onwards if $\mathbf{v}_i(t+1) - \mathbf{v}_i(t) = 0$ for all i for all $t \geq t_0$. In the special case of zero universal threshold ($\rho_i = \rho = 0$), Proposition 2.2 has established that the only possible steady state is consensus, and that the system converges to it from any initial state in the sense of $\mathbf{v}_i(t+1) - \mathbf{v}_i(t) = 0$ for all i as $t \rightarrow \infty$. For other choices of ρ_i , systems may converge to other (non-consensus) types of steady state, and some systems may not converge to any steady state. For the practical purpose of numerical simulations, where it is impossible to let $t \rightarrow \infty$, we use the following procedure to determine whether a system has reached a (pseudo-) steady state, allowing us to terminate the simulation at some finite time.

- (1) Two agents are in the same *cluster* if the Euclidean distance between their opinions is less than 10^{-6} . The *clustering* of the population refers

Table 1. Parameters used in numerical simulations.

Parameter	Values used
D (Dimensionality of every opinion)	1, 2, 3, 5
μ (Memory capacity of population)	2, 10
ρ (Universal threshold in Section 3.2; baseline threshold in Section 3.3)	0, 0.01, 0.02, ..., 0.99
α (Reinforcement rate; only in Section 3.3)	0.1, 0.2, 0.4, 0.8

to the partition of the N agents into their clusters. For example, if agents labeled by even numbers are in one cluster (all pairwise distances less than 10^{-6}) while all odd-numbered agents are in a different cluster, at some time t , then we say that the clustering at this time is $\{(1, 3, 5, \dots), (2, 4, 6, \dots)\}$.

- (2) If there exists some time $t_c \geq 0$ such that, for $t = t_c + 1, t_c + 2, \dots, t_c + 100$:
- (a) The clustering of the population remains the same as the clustering at time t_c ; and
 - (b) No agent ‘accelerates’ by more than 10^{-6} in any dimension at any time, i.e., $\max_{i,j}\{v_{ij}(t+1) - v_{ij}(t)\} \leq 10^{-6}$; and
 - (c) no agent’s opinion at $t = t_c + 100$ is further than 10^{-6} away in any dimension from their opinion at $t = t_c$, i.e., $\max_{i,j}\{v_{ij}(t_c + 100) - v_{ij}(t_c)\} \leq 10^{-6}$;

then we say that the system has reached a (pseudo-)steady state at time t_c , and stop the simulation. We call t_c the *convergence time* of this system.

In short and roughly speaking, we stop the simulation at time t_c if all agents have barely moved for 100 time-steps, and our definition of ‘barely’ is a very strict condition. The ‘pseudo-steady’ states therefore serve as very good proxies for the real (analytical) steady states of the system, so we will refer to them simply as steady states. If, by the criteria above, the system fails to converge to any steady state within 5000 time-steps, then we declare that the system in that particular configuration (of initial state and parameters) fails to converge.

Through our simulations, we find that a system with universal threshold ($\rho_i = \rho$ taking values as per Table 1) always converges to some steady state, regardless of the other parameters and initial state (see, Section 3.2). On the other hand, a system with individually evolving heterogeneous thresholds, where the reinforcement rate takes values as per Table 1, sometimes fails to converge in interesting ways (see, Section 3.3).

3.2. Universal threshold: consensus versus segregation

In this part of the investigation, we assume that all agents have the same threshold, which remains constant for all time. That is, $\rho_i(t) = \rho = \text{constant}$ for all i .

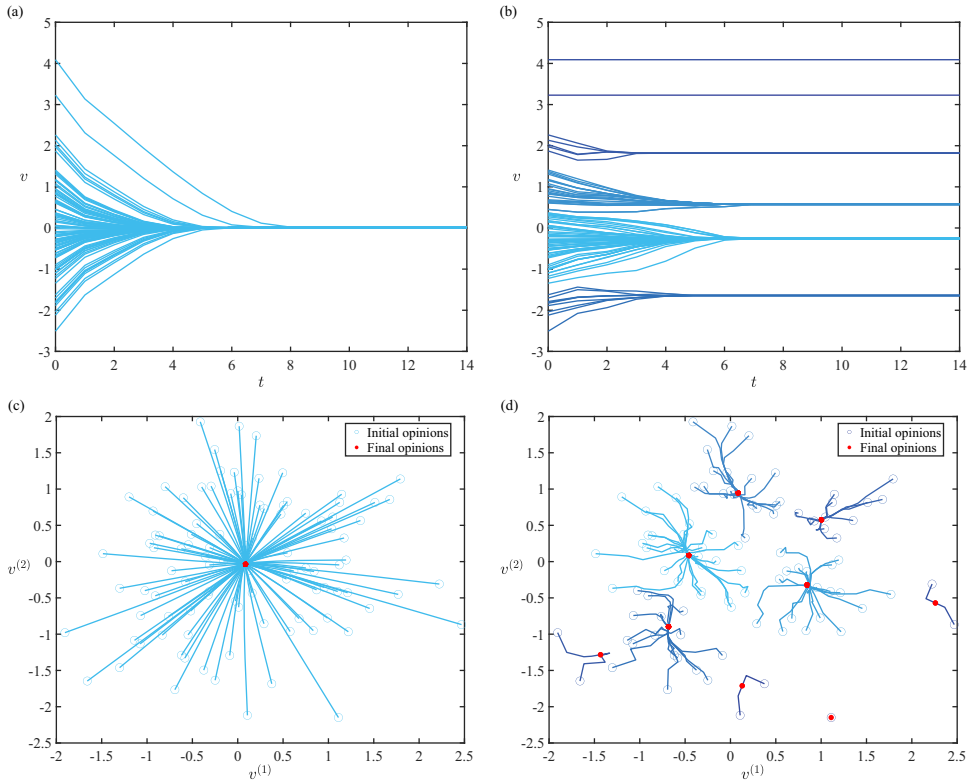


Figure 2. Examples of one- and two-dimensional opinion trajectories of $N = 100$ agents evolving according to (3), with memory capacity $\mu = 2$. Panels (a,b): One-dimensional dynamics from a common initial state (sampled from a standard normal distribution), with different universal thresholds: $\rho = 0$ (a) and $\rho = 0.8$ (b). Panels (c,d): Two-dimensional dynamics from a common initial state (the two dimensions sampled from two independent standard normal distributions), with different universal thresholds: $\rho = 0$ (c) and $\rho = 0.8$ (d); $v^{(1)}$ and $v^{(2)}$ denote the first and second dimensions of the opinions, respectively. Consensus is reached in (a,c), while (b) exhibits segregation with 6 distinct clusters, and (d) shows segregation with 9 distinct clusters.

The simulations produced two distinct types of phenomena. The system reaches either a steady state of consensus, in which there is exactly one cluster (see, [Figure 2a,c](#)), or a steady state of *segregation*, where more than one clusters co-exist ([Figure 2b,d](#)). In particular, whenever consensus is formed, the consensus opinion equals the initial mean opinion of the population, as [Proposition 2.2](#) predicts. When segregation is reached under a high value of ρ , it is typical that some agents never alter their opinion for all $t > 0$ ([Figure 2b,d](#)). This stubbornness is exhibited only by agents whose initial opinions are ‘extreme,’ i.e. far from $\mathbf{0}$. Since the universal threshold is high, everyone listens only to a small number of others, and it is likely that those who hold initial opinions far from everyone else will never be influenced by anyone. In the example of [Figure 2d](#), the set of connections among all agents, or the *connectome* of the population, evolves in the manner displayed by [Figure 3](#). The figure shows

that even at the initial $t = 0$, one agent is ‘their own island’: not connected to anyone. This agent never alters their opinion while the agents who have connections evolve their positions. It is a common feature of the model that as the opinions evolve, the connectome becomes more disconnected in the graph-theoretical sense: more isolated ‘islands’ appear. In the example illustrated by [Figure 3](#), the nine clusters that constitute the population’s final steady state have almost stabilized by $t = 7$, at which time only a few connections remain while the majority of initial connections have been severed. The cutting of a connection occurs if the pairwise affinity drops below the relevant threshold, and affinity decays over time if two agents keep failing to agree with each other. The model dictates that agents always try to align with neighbors; the difference between their succeeding in coming together and failing to do so (before their connection is cut) gives rise to the difference between consensus and segregation.

The rate at which connections are lost is strongly dependent on the population’s memory capacity, μ . Comparing [Figure 3](#) to [Figure 4](#), we see that given identical initial opinions and other parameters, the connectome evolves more quickly when the memory capacity is small. That is, if agents quickly forget past discrepancies, then the connectome gets rewired dramatically at each step, and the system takes few steps to stabilize. This result is reminiscent of a recent success story in mathematical sociology. After multiple theoretical models predicted that the rapid rewiring of a social network promotes cooperative behavior (Fu et al., 2008; Hanaki et al., 2007; Santos et al., 2006), the phenomenon was observed in a human experiment by Rand et al. (2011). In the current model, faster rewiring of the connectome accompanies not only faster stabilization of the population, but also the formation of fewer, larger clusters (see, [Figure 5](#)). This effect is most pronounced when the dimensionality of opinion space is $D = 2$ or 3 and when the universal threshold is high ($\rho \approx 0.8$). For example, in two-dimensional simulations with $\rho = 0.8$, the mean number of stable clusters formed is 19 if $\mu = 10$ and 12 if $\mu = 2$, the latter scenario having necessarily larger cluster sizes on average ([Figure 5c,d](#)). Even more dramatically, in three-dimensional simulations with $\rho = 0.8$, the mean number of stable clusters is 48 if $\mu = 10$ and 35 if $\mu = 2$ ([Figure 5e,f](#)). By interpreting the large clusters (which are always close to the neutral $\mathbf{0}$ position of opinion space) as cooperative groups, and the small clusters (which are always on the periphery of opinion space) as ‘defectors’ in the language of Rand et al. (2011), we are able to understand the dynamics presented here as a process of seeking cooperation. Note that in order to make such identifications, we need to assume that cooperation is the neutral, or default, position; that a randomly sampled population will position themselves in a normal distribution around it. The smaller the memory capacity (or, the more ‘forgetful’ the agents), the more quickly the network gets rewired and cooperative clusters are formed,

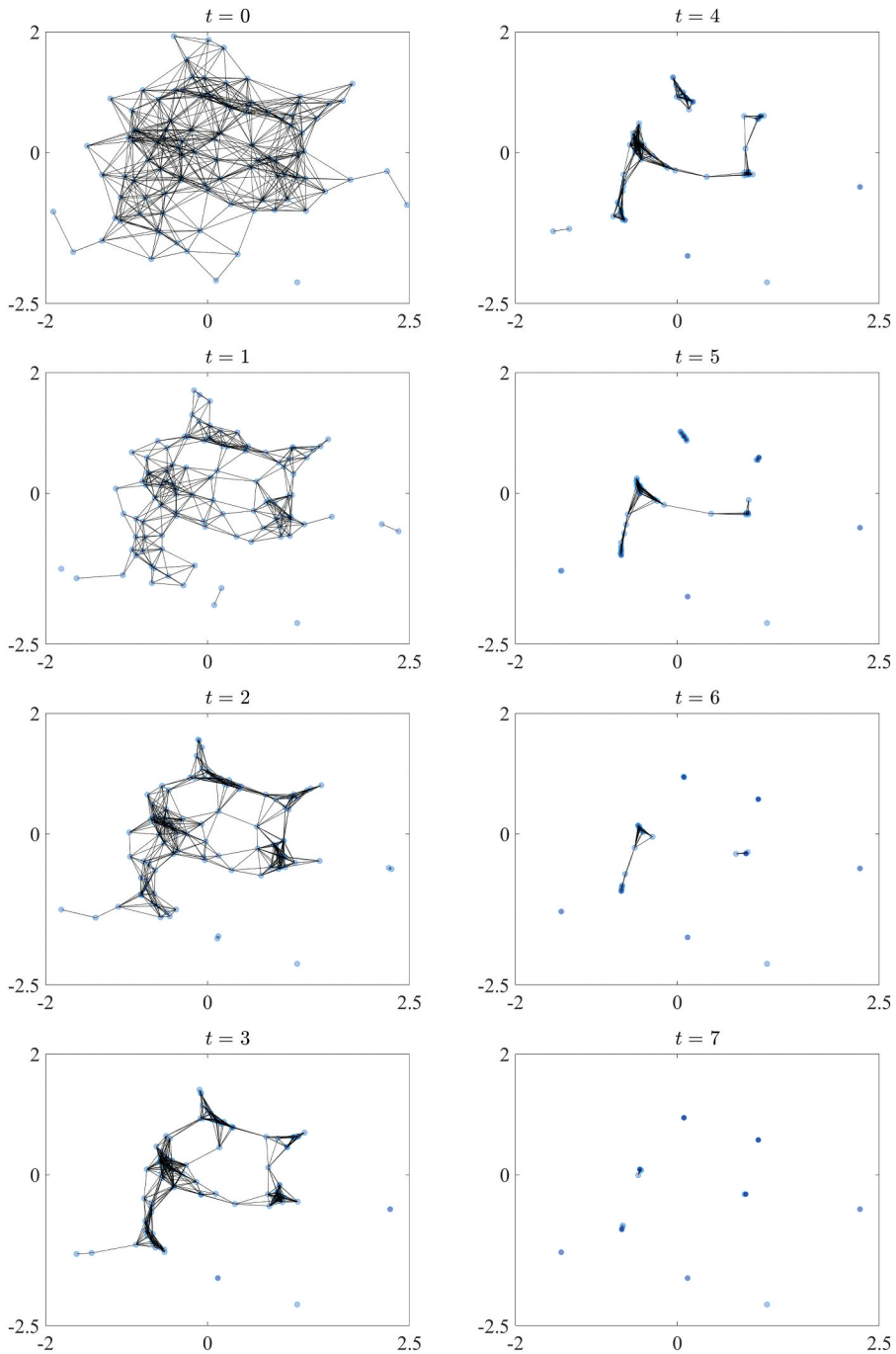


Figure 3. Evolution of the connectome of the population as the opinions evolve in the manner of Figure 2d, with parameters $D = 2$, $\mu = 2$, $\rho = 0.8$, $a = 0$. Opinions are represented by (blue) dots while connections between agents are (black) lines. The opacity of the dots increase as agents overlap, so that the larger the cluster, the darker the dots. Since the threshold ρ is universal, every connection is bidirectional: the pair of agents influence each other. The opacity of the dots increase as agents overlap, so that the larger the cluster, the darker the dots.

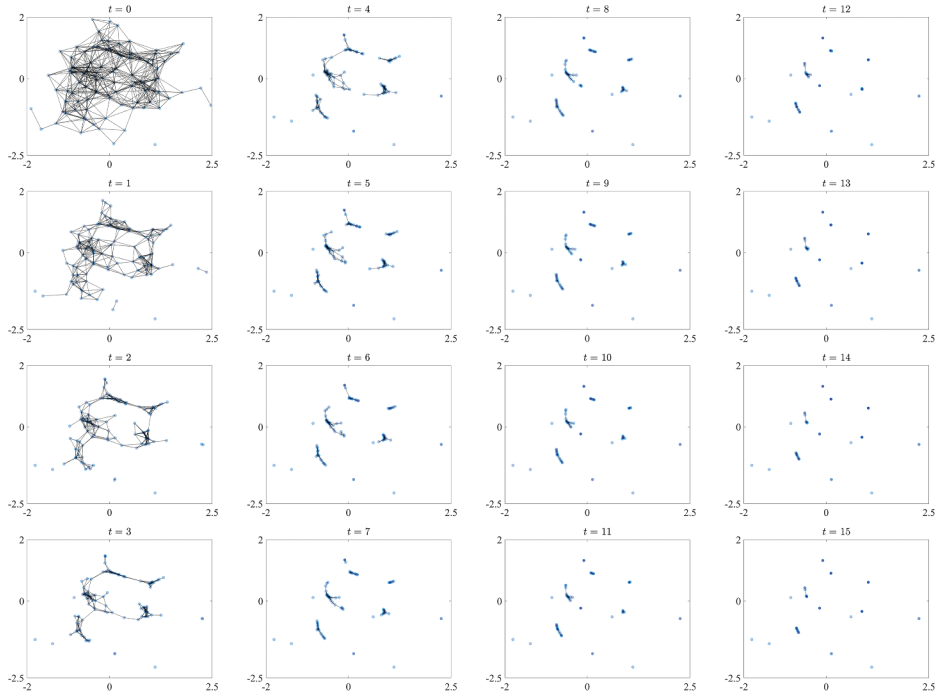


Figure 4. Evolution of the connectome of the population as the 2-dimensional opinions evolve from the same initial state as in [Figure 2d](#), with parameters $\mu = 10, \rho = 0.8, \alpha = 0$. Opinions are represented by (blue) dots while connections between agents are (gray) lines. The opacity of the dots increase as agents overlap, so that the larger the cluster, the darker the dots. Since the threshold ρ is universal, every connection is bidirectional: the pair of agents influence each other.

and the larger those clusters. This finding is consistent with [Rand et al. \(2011\)](#) and the preceding theoretical predictions.

The remainder of this section focuses on the effects of the parameters D, ρ and μ on the simulation results, particularly on clustering and segregation. We reiterate that these results are contingent on the assumption of normally distributed initial opinions.

For any given initial state, the system reaches consensus if ρ is sufficiently small, and segregation if ρ is sufficiently large (all other parameters being fixed). That is to say, if everyone is sufficiently amenable, then consensus will be formed; otherwise, there will be segregation. A deeper investigation of this phenomenon reveals a key feature of the model. For any fixed D and μ , the number of clusters in the steady state tends to increase with ρ ; in fact, the mean number of clusters formed over 1000 simulations is a monotonic function of ρ (see, [Figure 5](#)). If $\rho \leq \rho_c$ for some ρ_c which depends on D and μ , the only outcome over 1000 simulations is consensus. For example, if $(D, \mu) = (1, 2)$ then $\rho_c = 0.29$ ([Figure 5a](#)); if $(D, \mu) = (1, 10)$ then $\rho_c = 0.2$ ([Figure 5b](#)); and if $(D, \mu) = (2, 2)$ then $\rho_c = 0.27$ ([Figure 5c](#)). We find that ρ_c is a decreasing

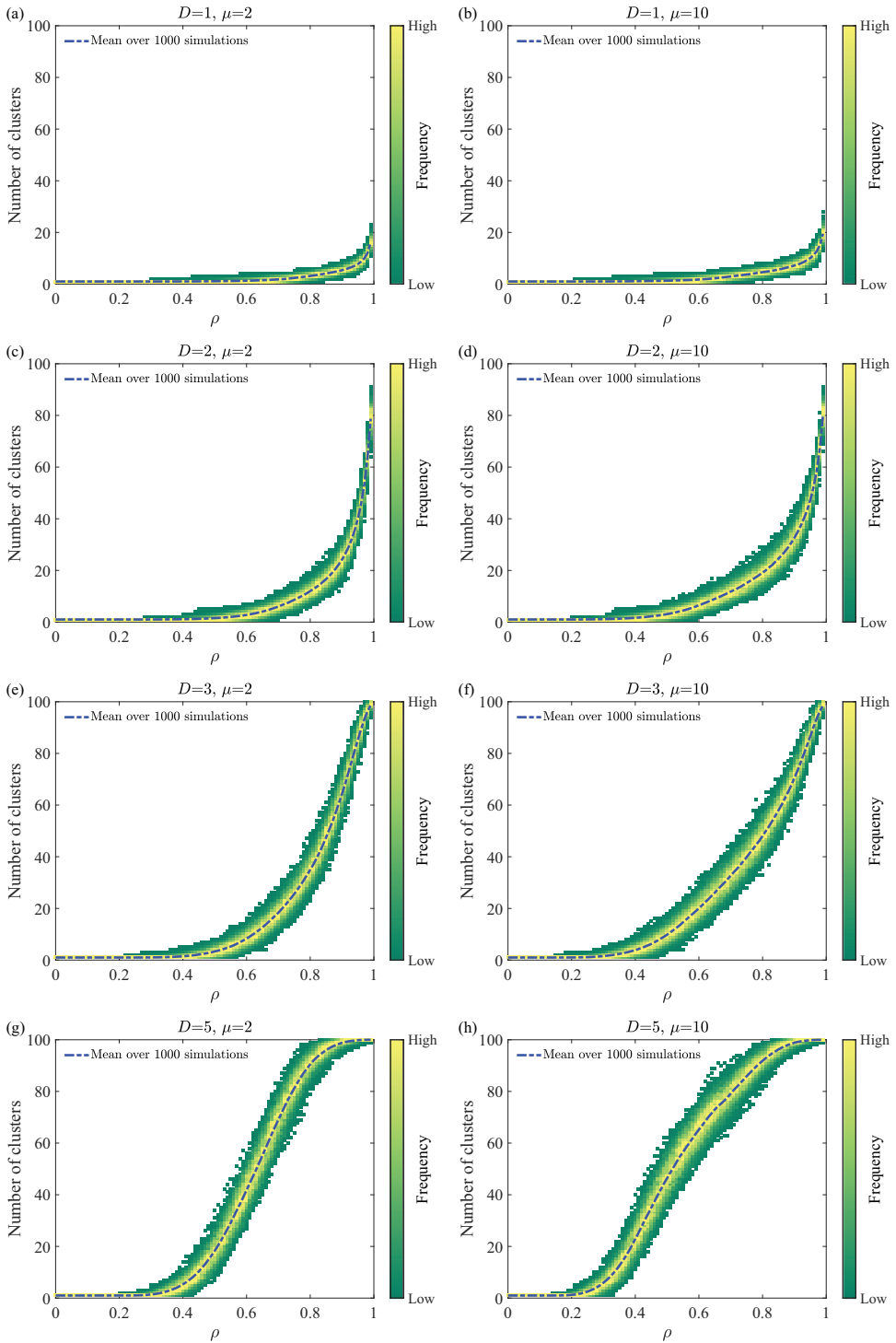


Figure 5. Number of clusters formed for $N = 100$ agents with universal threshold ρ taking values as per Table 1; 1000 simulations per value of ρ . A common set of 1000 initial states are used. Dimensionality $D = 1$ (a,b), 2 (c,d), 3 (e,f), and 5 (g,h). Memory capacity $\mu = 2$ (a,c,e,g), and 10 (b,d,f,h).

function of both D and μ : the more high-dimensional the opinions, or the larger the collective memory capacity, the more amenable everyone must be in order to form a consensus. All these simulation results are consistent with the sufficient condition (19) for consensus. If $D \geq 3$, we find that some initial states lead to steady states with as many clusters as there are agents: every agent holds their own unique opinion and will not change them. We call such a steady state *maximum segregation*. These states are achievable (over the 1000 simulations that we ran) only if $\rho \geq \rho_s$ for some ρ_s which depends on D (but its dependence on μ is negligible). For example, if $D = 3$ then $\rho_s = 0.97$ (Figure 5ef); and if $D = 5$ then $\rho_s = 0.83$ (Figure 5gh). We find that ρ_s is a decreasing function of D : the more high-dimensional the opinions, the easier it is for the system to reach maximum segregation. In particular, for $D = 5$, the mean number of clusters formed resembles a sigmoid function of ρ where, for $\rho \geq 0.93$, even the mean number is greater than 99.5, indicating that maximum segregation is extremely likely.

We also find that if $\rho \geq \rho_{nc}$ for some ρ_{nc} which depends on D and μ , then the population never forms a consensus. For example, if $(D, \mu) = (1, 2)$ then $\rho_{nc} = 0.85$ (see, Figure 5a); if $(D, \mu) = (1, 10)$ then $\rho_{nc} = 0.78$ (Figure 5b); and if $(D, \mu) = (2, 2)$ then $\rho_{nc} = 0.71$ (Figure 5c). If $(D, \mu) = (5, 10)$, then ρ_{nc} becomes as small as 0.34. The more high-dimensional the opinions, or the larger the collective memory capacity, the easier it is for consensus to be impossible.

The *convergence time*, t_c (defined in Section 3.1), is strongly dependent on the memory capacity, μ (see, Figure 6). When $\mu = 2$, no simulations take more than 50 steps to converge, and 95% of simulations take fewer than 25 steps to converge (Figure 6a,c,e,g). Raising the memory capacity to $\mu = 10$ approximately doubles the convergence time (Figure 6b,d,f,h). The mean convergence time is maximized by a ρ -value that is negatively correlated with both D and μ . When $D = 3$ or 5, simulations with large ρ can yield zero convergence time (Figure 6e,f,g,h). Indeed, if the affinity threshold is so high that there are no interactions between agents in the initial state, then no agent would ever deviate from their initial opinion, leading to maximum segregation with 100 distinct clusters (see, Figure 5e,f,g,h).

We define the *opinion drift* of a system as the Euclidean distance from the initial mean opinion of the population to the steady-state mean opinion. The simulations reveal that the mean opinion drift over all simulations is maximized at some $\rho = \rho_d$ which depends on D and μ (see, Figure 7). While ρ_d is a decreasing function of both D and μ , the maximum value of mean opinion drift increases with D and μ , reaching approximately 0.11 when $(D, \mu) = (5, 10)$. The opinion drift is zero for sufficiently small ρ , a result consistent with the fact that (19) is a sufficient condition for convergence to the mean initial opinion. The phenomenon of opinion drift demonstrates that

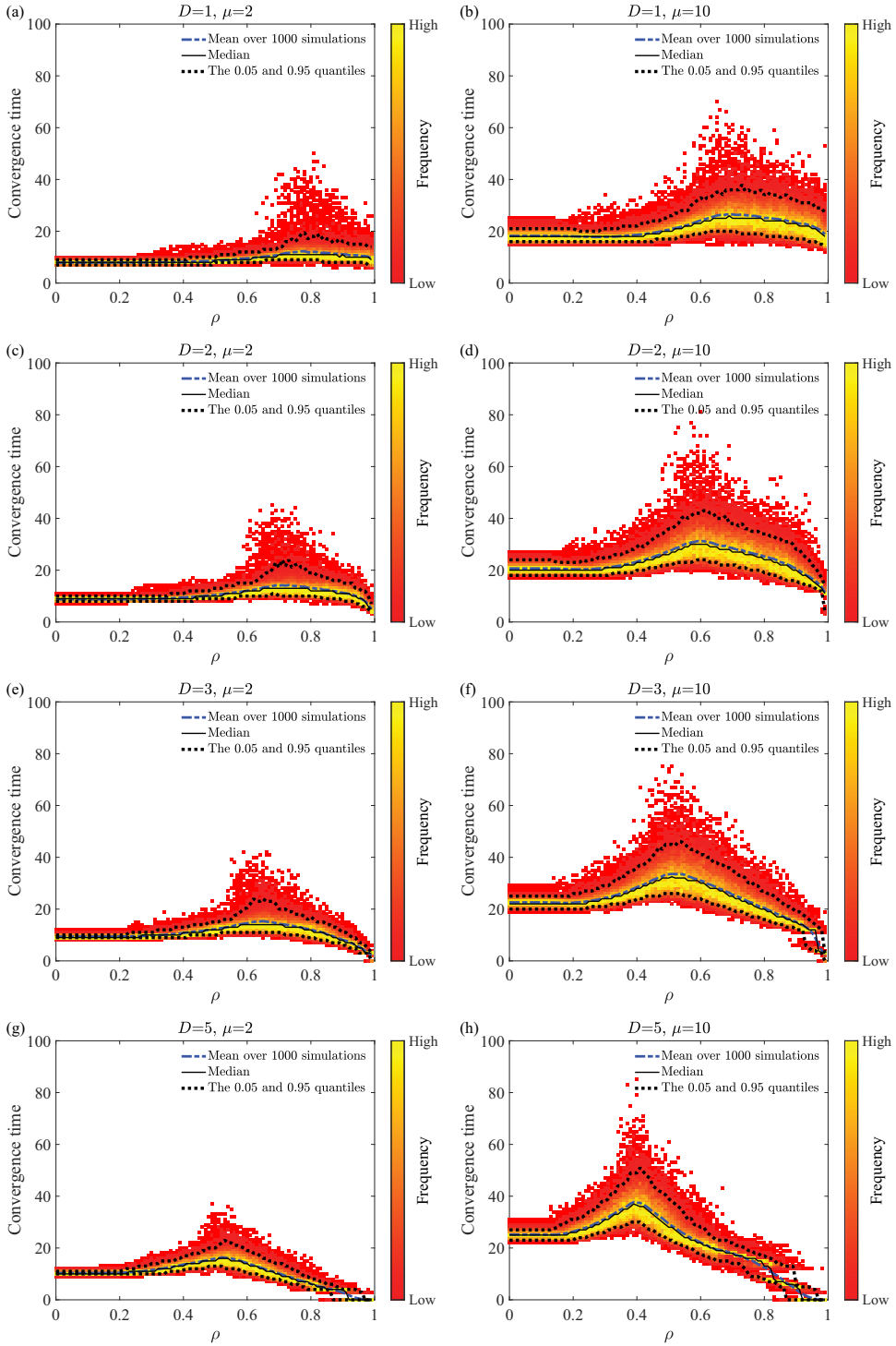


Figure 6. Convergence time for $N = 100$ agents with universal threshold ρ taking values as per Table 1; 1000 simulations per value of ρ . A common set of 1000 initial states are used. Dimensionality $D = 1$ (a,b), 2 (c,d), 3 (e,f), and 5 (g,h). Memory capacity $\mu = 2$ (a,c,e,g), and 10 (b,d,f,h).

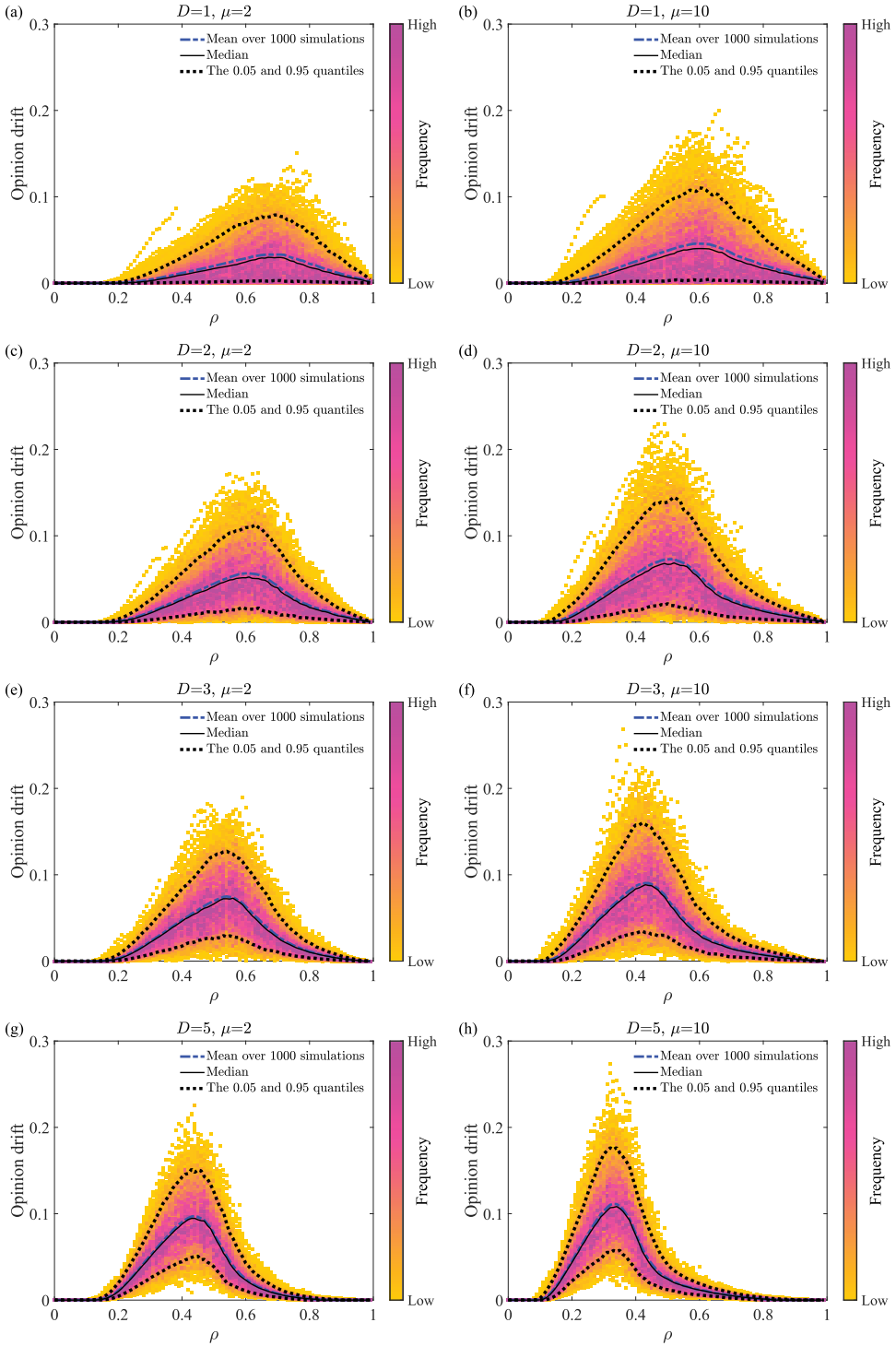


Figure 7. Opinion drift for $N = 100$ agents with universal threshold ρ taking values as per Table 1; 1000 simulations per value of ρ . A common set of 1000 initial states are used. Dimensionality $D = 1$ (a,b), 2 (c,d), 3 (e,f), and 5 (g,h). Memory capacity $\mu = 2$ (a,c,e,g), and 10 (b,d,f,h). The grouping precision of opinion drift is 0.001; that is, the histogram bins are the intervals $[0, 0.001)$, $[0.001, 0.002)$, and so on.

the population's average opinion tends to change over time as the agents evolve into clusters, and it tends to change more for more complex systems (recall that the system is most complex at intermediate values of ρ). The simplest systems, with extreme values of ρ , tend to exhibit very small amounts of opinion drift as the agents either form a consensus (small ρ) or barely adjust their opinions (large ρ). A similar fact holds for the convergence time: the more complex systems tend to take longer to reach steady state (Figure 6).

3.3. Evolving heterogeneous thresholds: extremisation and oscillations

In this second line of investigation, we allow agents to evolve their thresholds from some baseline value, ρ , according to Equation (6; see, Figure 1). Recall that the reinforcement rate, $\alpha > 0$, determines how sharply one's threshold increases as one's opinion becomes more extreme. Agents with more extreme views will have higher thresholds and therefore be less inclined to listen to other agents, thus making those extreme agents appear 'stubborn.' This correlation between extremeness of views and stubbornness has been studied in formal models and observed in real data (Kozitsin, 2020; Tian et al., 2021). For simplicity, we fix the dimensionality of opinion space at $D = 2$ throughout this section.

Unlike the scenario with a universal threshold (which can be recovered in the limit $\alpha \rightarrow 0$) where every initial state leads to a steady state, we find that when α is sufficiently large, not all initial states induce a steady state (see, Figure 8). The number of failures to reach steady state in 1000 simulations, $F_{\mu,\alpha}(\rho)$, is negatively correlated with the baseline threshold, ρ , and positively correlated with the memory capacity, μ . Given any combination of (μ, α, ρ) within the range as per Table 1, the number of simulations that reach steady state is always at least 950, providing a suitably large pool of results to analyze. We consider the cases that fail to converge, and the collective dynamics that arise, in more detail in Section 3.4.

For every setting of (μ, α, ρ) , the $1000 - F_{\mu,\alpha}(\rho)$ simulations that do reach steady state provide us with results on cluster formation and on convergence time, enabling comparisons with corresponding results in the case of universal thresholds. Firstly, the mean number of clusters formed is an increasing function of ρ , α and μ (see, Figure 9a,b), and consistently higher than the counterpart under a universal threshold (Figure 5c,d). Thus, a system where agents become more stubborn as their opinions become extreme tends to become more segregated than a system with a universal threshold. Meanwhile, for sufficiently small ρ , the mean convergence time is much larger under evolving heterogeneous thresholds than under universal thresholds (compare Figure 9c,d with Figure 6c,d). A larger reinforcement rate α is therefore

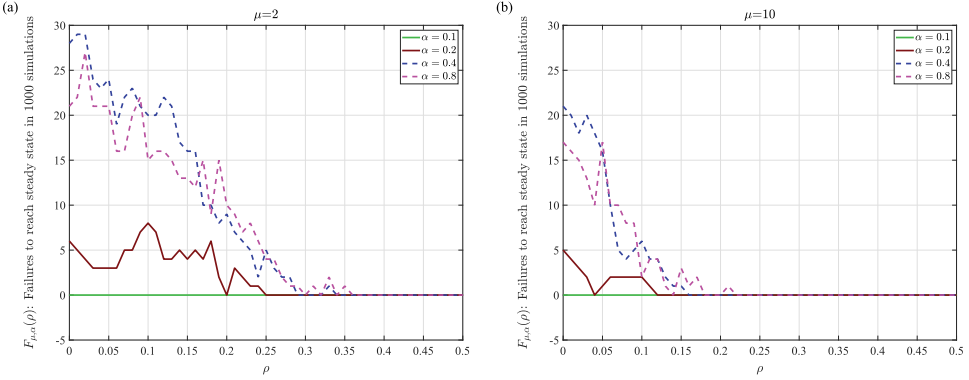


Figure 8. Under evolving heterogeneous thresholds, with reinforcement rate $\alpha > 0$, a small number of simulations out of the total 1000 fail to produce a steady state. That number, $F_{\mu,\alpha}(\rho)$, depends on α , the memory capacity μ , and the baseline threshold ρ . Panel (a): $(D, \mu) = (2, 2)$. Panel (b): $(D, \mu) = (2, 10)$.

responsible not only for more splintering of the population, but also for longer times taken by any sub-population to reach an agreement.

The most striking result that we observe from simulations relates to the extremisation of opinions. We define the *extremisation measure* of the system as the difference between two Euclidean norms:

$$\text{Extremisation measure} = \left\| \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i(t_c) \right\| - \left\| \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i(0) \right\|, \quad (20)$$

where N is the population size (always 100 in this study), $\mathbf{v}_i(t)$ are the opinions and t_c is the convergence time. Recall that the origin in D -dimensional opinion space represents the neutral opinion, and that the Euclidean norm of any position in the opinion space is a measure of how extreme it is. Thus, the extremisation measure represents the extent to which the population's average view becomes more extreme over the course of the opinion dynamics; a positive (negative) value indicates that the average view becomes more extreme (more moderate). Note that extremisation is unlikely to be negative when we generate the initial opinions from normal distributions, which necessarily results in an initial mean close to $\mathbf{0}$. Nevertheless, the fashion in which positive extremisation occurs is illuminating, as we now proceed to demonstrate.

In many instances, we observe that the mechanism by which the average view becomes more (or less) extreme over time is a *collective drift* (see, [Figure 10](#)), in which a large group of agents form an unstable drifting cluster with more members than any stable cluster. These drifting agents first coalesce around some neutral opinion, before collectively moving away from it, being drawn to a small number of fringe agents. The drifting cluster eventually

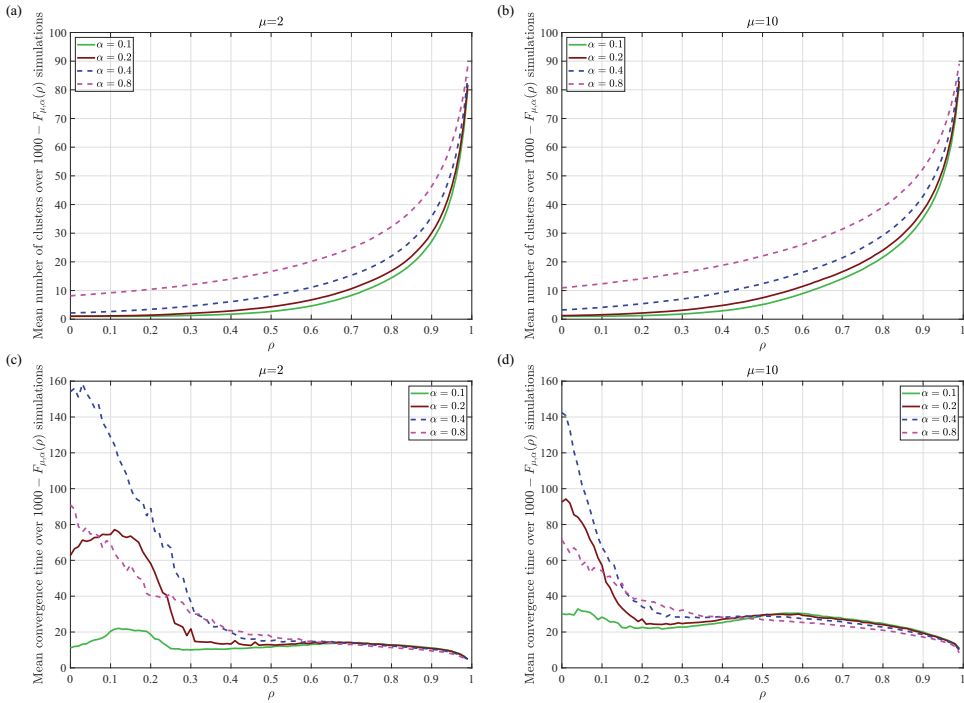


Figure 9. Under evolving heterogeneous thresholds, the mean number of clusters formed (a,b) and the mean convergence time (c,d) are taken, for each parameter setting (μ, α, ρ) , from all simulations that result in steady states. Panels (a,c): $\mu = 2$. Panels (b,d): $\mu = 10$. Other parameters used for each panel: $D = 2$, $\alpha = 0.1, 0.2, 0.4, 0.8$, and $\rho = 0, 0.01, 0.02, \dots, 0.99$.

stabilizes, merging with the fringe attractors, so the population reaches a steady state. The drift toward the extremities of the opinion space equates to a positive extremisation measure for the population. This phenomenon where fringe agents exert great influence over the moderate amenable majority, pulling their opinions to the extremes, has been widely studied in the context of radicalization. For example, it has been observed that when university students without strong existing social identities are exposed to a large variety of strong views, they become at high risk of radicalization (Hollewell & Longpré, 2022). More generally, it has been proposed that fair-minded individuals become radicalized through deepening engagement with extremists on a gradually narrowing ‘Staircase to Terrorism’ (Moghaddam, 2005).

A detailed view of the dynamics depicted in Figure 10a is presented in Figure 11. We see that three fringe clusters have formed by the time $t = 6$, after which they exert influence over the relatively neutral majority without moving their own positions. At a much later time, one of the fringe groups begins moving under the influence of the majority due to its close proximity, and eventually merges with the majority, stabilizing the entire population.

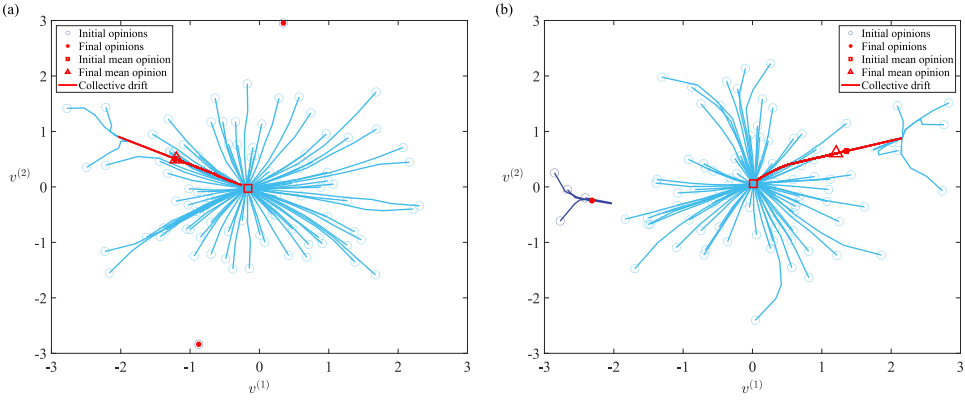


Figure 10. Two examples of collective drift of opinions under evolving heterogeneous thresholds. The first and second dimensions of the opinions are denoted by $v^{(1)}$ and $v^{(2)}$, respectively. Drifting begins at some t which is chosen to best illustrate the trajectory (rather than rigorously defined). Final opinions are taken at the convergence time, t_c . Parameters: $\mu = 2, \rho = 0, \alpha = 0.8$.

The simulations show that a short memory capacity ($\mu = 2$) tends to induce larger extremisation measures than a long one ($\mu = 10$), suggesting that a population who takes a long history of itself into account is less likely to become extremised (comparing Figure 12a,c,e,g with b,d,f,h). This finding supports the theory that, the more strongly one's recent memory influences one's online behavior, the more rapidly one tends to become sympathetic to extremist views (Z. Z. Cao et al., 2018). If the baseline threshold ρ is close to 1, then almost all simulations produce extremisation measures close to zero, simply because these systems tend not to induce any changes in opinions at all. If the reinforcement rate α is small, then the majority of simulations produce zero extremisation (even though outliers with enormous extremisation skew the mean value away from the median; see, Figure 12a,b). If α is suitably large and ρ sufficiently small (a population where the neutral agents are highly amenable but the fringe agents are highly stubborn), then the mean and median values of extremisation measure closely align, and we infer that the population's most likely behavior is high extremisation (Figure 12e,f,g,h). In such cases, for every fixed (μ, α) pair, the mean/mode extremisation measure is maximized by $\rho = 0$. In particular, for $(\mu, \alpha, \rho) = (2, 0.4, 0)$, the mean/mode extremisation measure is just over 1 (Figure 12e), which is a substantial distance in the normalized opinion space. That is to say, the agents tend to move a long way from their initial positions to become their extremised final selves.

All the extremisation results mirror the well-known socio-psychological effect of group polarization, where a group moves toward a view more extreme than most individual views that were held before their exposure to social influence (Moscovici & Zavalloni, 1969; Myers & Lamm, 1976). A similar

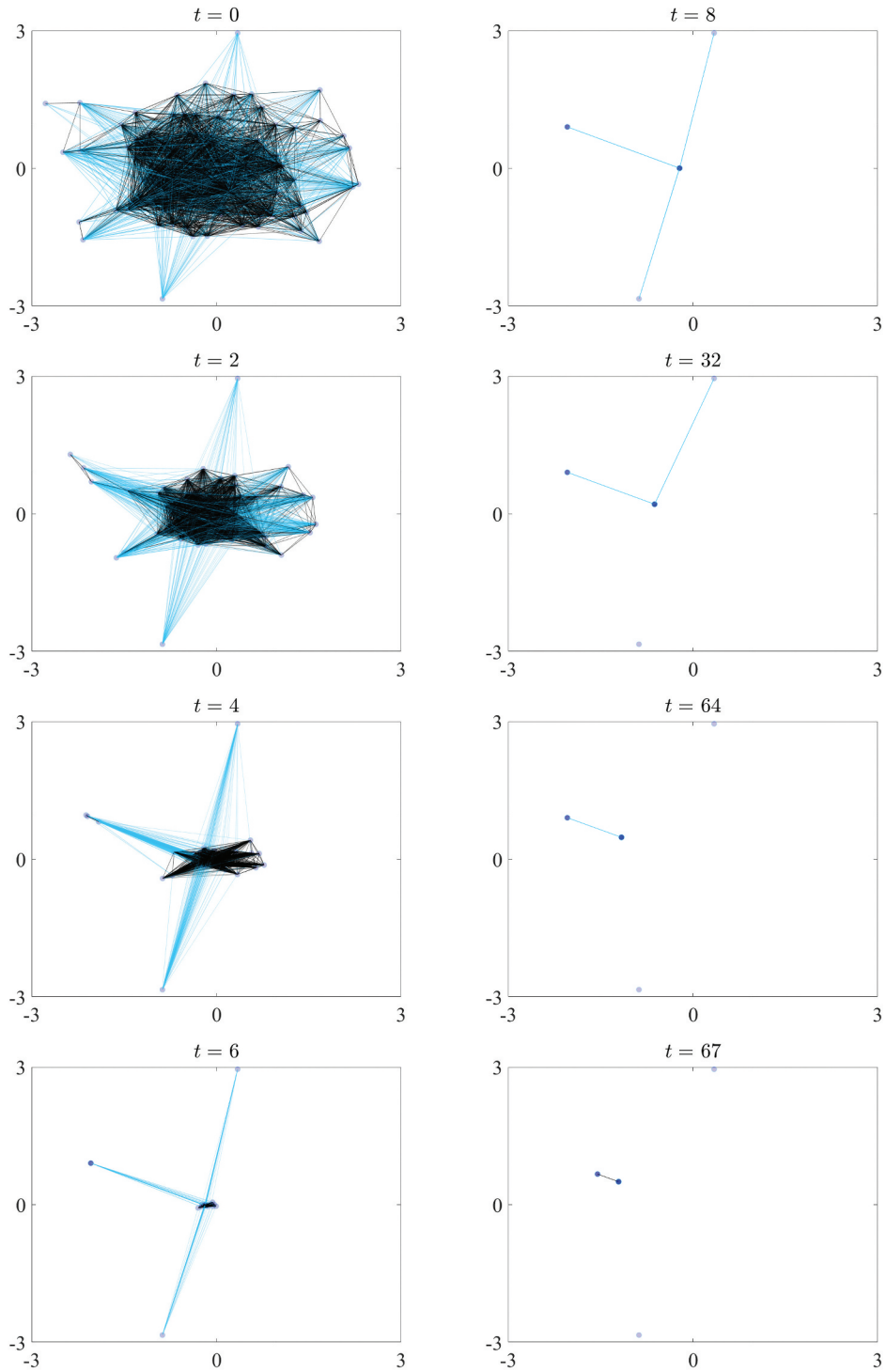


Figure 11. Key steps in the evolution of the connectome of the population as the opinions evolve in the manner of Figure 10a, with parameters $D = 2$, $\mu = 2$, $\rho = 0$, $\alpha = 0.4$. Opinions are represented by (blue) dots, bidirectional connections are dark (gray) lines and unidirectional connections are light (blue) lines. The opacity of the dots increase as agents overlap, so that the larger the cluster, the darker the dots. Since the threshold ρ is heterogeneous, unidirectional connections may exist, where agent j influences agent i without reciprocation.

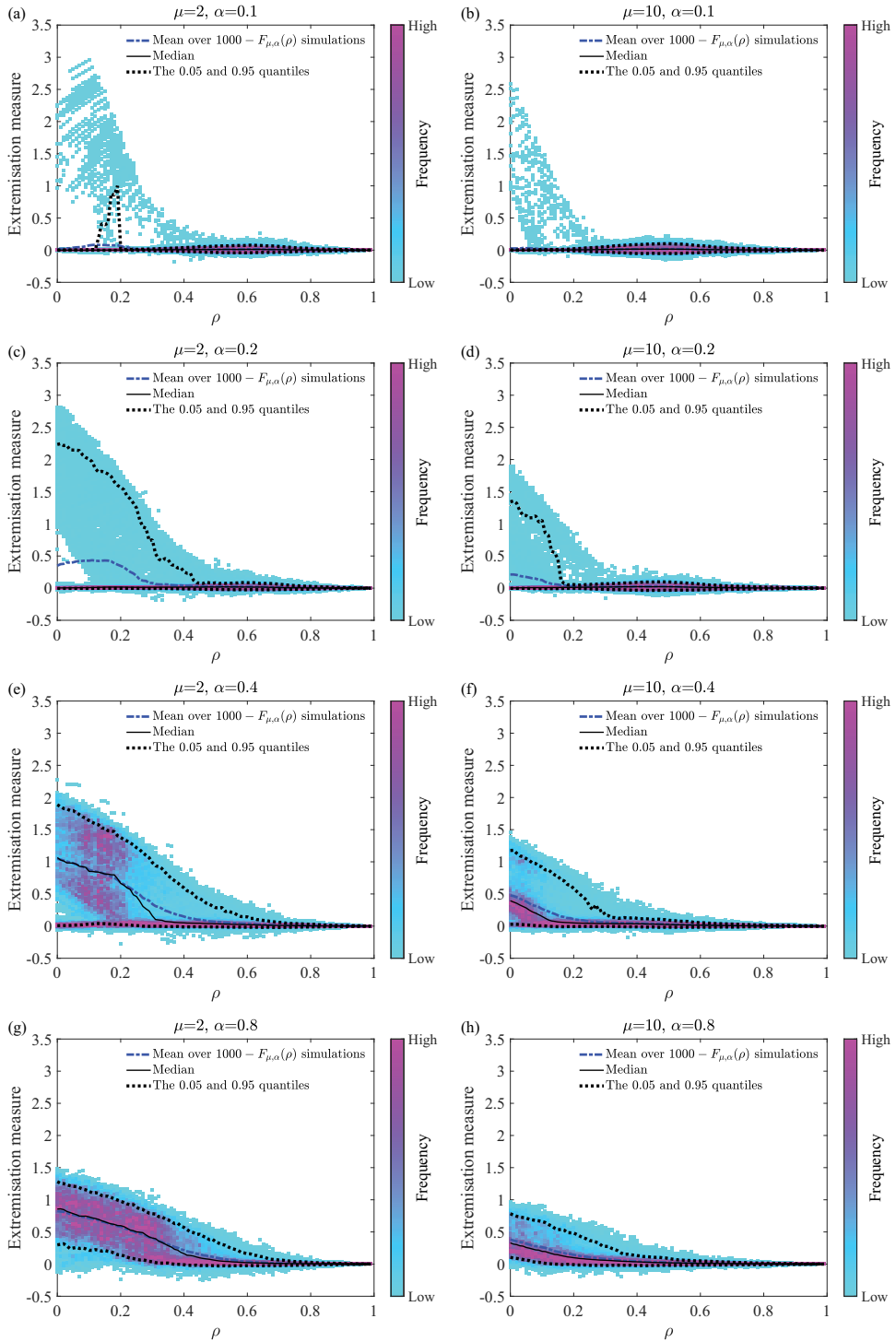


Figure 12. Extremisation measure for $N = 100$ agents with evolving heterogeneous thresholds $\rho_i(t)$; the baseline threshold ρ takes values as per Table 1. A thousand simulations are performed per value of ρ and a common set of 1000 initial states are used for each ρ . Dimensionality $D = 2$. Memory capacity $\mu = 2$ (a,c,e,g), and 10 (b,d,f,h). Reinforcement rate $\alpha = 0.1$ (a,b), 0.2 (c,d), 0.4 (e, f), and 0.8 (g,h). The grouping precision of extremisation measure is 0.01; that is, the histogram bins are intervals of size 0.01, giving the same number of bins as in Figure 7.

effect has been observed in the increasing polarization of the US senate over time (Liu & Srivastava, 2015). The present model provides a detailed view of the mechanics underlying the group polarization effect; for example, we have described the collective drift mechanism, where the majority abandon their moderate initial agreement and become extremised by fringe agents. A sociologically significant lesson arising from these results is that, if the fringe agents, who hold extreme views to begin with, were more amenable to change (i.e. if α were smaller in the model), then such collective extremisation would not occur.

3.4. Failure to converge: collective oscillations

As seen in Figure 8, when agents possess evolving heterogeneous thresholds, a small number of simulations fail to converge to a steady state. Before presenting the dynamics produced by the numerical results, we will first explicitly construct a system with evolving heterogeneous thresholds as per eq. (6), which fails to converge to any steady state and instead exhibits oscillatory dynamics.

Consider $N \geq 3$ agents in $D = 1$ dimension, with opinions denoted by v_i for $i = 1, 2, \dots, N$. Let the memory capacity $\mu = 1$ and baseline threshold $\rho = 0$. At $t = 0$, let $v_1 = L < 0$, $v_2 = R > 0$. We require $R + L \neq 0$ and assume without loss of generality that $R + L > 0$, then define

$$C = \frac{R + L}{2} > 0. \quad (21)$$

Let the initial $v_3 = v_4 = \dots = v_N = v_*$ for some $v_* \in (0, C)$. The following facts about the affinities a_{N1} and a_{N2} are easily established through elementary calculus.

- (1) a_{N1} is a strictly decreasing, smooth, positive function of $v_* \in (0, C)$;
- (2) a_{N2} is a strictly increasing, smooth, positive function of $v_* \in (0, C)$;
- (3) $a_{N2} < a_{N1}$ for all $v_* \in (0, C)$, with $a_{N1}(v_* \rightarrow C) = a_{N2}(v_* \rightarrow C) = \frac{1}{\sqrt{1+X^2}}$, where we have defined the half-distance between R and L , $X = (R - L)/2 > -L$.

Whatever R and v_* are, we choose a reinforcement rate $\alpha > 0$ such that the threshold ρ_N coincides with a_{N2} ; that is,

$$1 - e^{-\alpha v_*} = \frac{1}{\sqrt{1 + (R - v_*)^2}}, \quad (22)$$

which we rearrange to give

$$e^{-\alpha} = \left(1 - \frac{1}{\sqrt{1 + (R - v_*)^2}} \right)^{1/v_*}. \quad (23)$$

We therefore have $\rho_3 = \rho_4 = \dots = \rho_N = a_{N2} < a_{N1}$, meaning that when agents $3, 4, \dots, N$ are at position v_* , they listen to agent 1 and do not listen to agent 2. As a corollary, since $1 - e^{-\alpha R} > 1 - e^{-\alpha v_*}$ and $1/\sqrt{1 + (R - v_i)^2} < 1/\sqrt{1 + (R - v_*)^2}$ for all $v_i < v_*$, agent 2 (while at position R) listens to no opinions less than or equal v_* . We take R and v_* to be such that α satisfies the constraint

$$1 - e^{-\alpha|L|} \geq \frac{1}{\sqrt{1 + L^2}} \Leftrightarrow e^{-\alpha} \leq \left(1 - \frac{1}{\sqrt{1 + L^2}} \right)^{1/|L|}. \quad (24)$$

which ensures that agent 1 (while at position L) listens to no opinions greater than or equal to 0.

We proceed to find further conditions under which, for agents $3, 4, \dots, N$ initialized at v_* , the subsequent dynamics are periodic: $v_3(t > 0) = \dots = v_N(t > 0) = \{0, v_*, 0, v_*, \dots\}$. To begin, we seek to make their common opinion zero at $t = 1$; that is,

$$0 = v_* + \frac{1}{n} a_{N1}(L - v_*) = v_* + \frac{L - v_*}{n\sqrt{1 + (L - v_*)^2}}, \quad \text{where } n = N - 1, \quad (25)$$

which implies the quadratic equation for L ,

$$(n^2 v_*^2 - 1)L^2 + (2v_* - 2n^2 v_*^3)L + ((n^2 - 1)v_*^2 + n^2 v_*^4) = 0. \quad (26)$$

Equation (26) has real solutions if and only if

$$\begin{aligned} v_* \neq \pm \frac{1}{n} \text{ and } 0 \leq (2v_* - 2n^2 v_*^3)^2 - 4(n^2 v_*^2 - 1)((n^2 - 1)v_*^2 + n^2 v_*^4) \\ = 4n^2 v_*^2 - 4n^4 v_*^4, \end{aligned} \quad (27)$$

if and only if $-1/n < v_* < 1/n$. Since $v_* > 0$ by construction, we use the constraint $0 < v_* < 1/n$. Thus, (26) has exactly one negative solution, which also solves (25):

$$L = v_* \left(1 - \frac{n}{\sqrt{1 - n^2 v_*^2}} \right). \quad (28)$$

According to (28), L is a strictly decreasing function of v_* ; for all $v_* \in (0, 1/n)$, we have $L < (1 - n)v_* < -v_*$. Next, we make $v_3(t = 2) = \dots = v_N(t = 2) = v_*$. Since their common threshold when $v_3 =$

$\dots = v_N = 0$ is 0, all those agents listen to both agent 1 and agent 2, so we require

$$\frac{1}{N} \left(\frac{R}{\sqrt{1+R^2}} + \frac{L}{\sqrt{1+L^2}} \right) = v_*. \quad (29)$$

It is clear that for all $v_* > 0$, any $R > 0$ and $L < 0$ satisfying (29) must be related by $R > -L$. To ensure that (29) has a real R solution, we impose the constraint

$$Nv_* - L < 1, \quad (30)$$

which implies $Nv_* - L/\sqrt{1+L^2} < 1$ (since $L < 0$), and therefore $R/\sqrt{1+R^2} = Nv_* - L/\sqrt{1+L^2}$ can be solved for R . Now, using (28) to write L in terms of v_* in (30) yields

$$nv_* \left(1 + \frac{1}{\sqrt{1-n^2v_*^2}} \right) < 1, \quad (31)$$

which translates to $v_* < m/n$, where

$$m = \frac{1}{2} \left(1 - \sqrt{2} + \sqrt{\sqrt{8} - 1} \right) \approx 0.469. \quad (32)$$

Note that $v_* < m/n$ is a stricter condition than $v_* < 1/n$.

So far, we have established that any $v_* \in (0, m/n)$, and the corresponding value of $L < -v_*$ determined by (28), guarantee the existence of some $R > -L$ satisfying (29). The question remains as to whether for some such v_* , the reinforcement rate α according to (23) is able to satisfy the constraint (24). To that end, we need

$$\left(1 - \frac{1}{\sqrt{1+(R-v_*)^2}} \right)^{|L|/v_*} \leq 1 - \frac{1}{\sqrt{1+L^2}} = 1 - \cos(\arctan L). \quad (33)$$

We will prove that (33) holds if n is sufficiently large and v_* appropriately defined in terms of n . Let

$$nv_* = \sin \phi, \quad (34)$$

for some ϕ which satisfies

$$\sin \phi + \tan \phi = \cos \phi. \quad (35)$$

The left-hand side of (35) is a strictly increasing function of $\phi \in [0, \arcsin m]$, with $\sin(0) + \tan(0) = 0$ and $\sin(\arcsin m) + \tan(\arcsin m) = m(1 + 1/\sqrt{1-m^2}) = 1$ by definition of m ; while the right-hand side is strictly decreasing from $\cos(0) = 1$ to $\cos(\arcsin m) < 1$. Therefore, (35) has exactly

one solution $\phi \in (0, \arcsin m)$, so that $v_* \in (0, m/n)$. Putting (34) into (28), we find

$$L = \frac{\sin \phi}{n} - \tan \phi, \tag{36}$$

and hence $|L|/v_* = n \sec \phi - 1$. Using the identity $R/\sqrt{1+R^2} \equiv \sin(\arctan R)$, re-arranging (29) yields

$$R = \tan \left(\arcsin \left(Nv_* - \frac{L}{\sqrt{1+L^2}} \right) \right), \tag{37}$$

and using the identities $1/\sqrt{1+R^2} \equiv \cos(\arctan R)$ and $\cos(\arcsin Z) \equiv \sqrt{1-Z^2}$, we further deduce

$$\begin{aligned} \frac{1}{\sqrt{1+R^2}} &= \sqrt{1 - \left(Nv_* - \frac{L}{\sqrt{1+L^2}} \right)^2} \\ &> \sqrt{1 - (Nv_* - L)^2} \\ &= \sqrt{1 - (\sin \phi + \tan \phi)^2}, \end{aligned} \tag{38}$$

where the final equality follows from (34) and (36). By (35), we then find

$$\frac{1}{\sqrt{1+R^2}} > \sin \phi. \tag{39}$$

Since $v_* \in (0, R)$, it then follows that

$$\begin{aligned} \left(1 - \frac{1}{\sqrt{1+(R-v_*)^2}} \right)^{|L|/v_*} &< \left(1 - \frac{1}{\sqrt{1+R^2}} \right)^{|L|/v_*} < (1 - \sin \phi)^{n \sec \phi - 1} : \\ &= F(n). \end{aligned} \tag{40}$$

We find that $F(n)$ is a strictly decreasing function of $n \geq 0$ with $F(\cos \phi) = 1$ and $\lim_{n \rightarrow \infty} F(n) = 0$. Moreover, we have

$$1 - \cos(\arctan L) = 1 - \cos \left(\arctan \left(\frac{\sin \phi}{n} - \tan \phi \right) \right) := G(n), \tag{41}$$

which is a strictly increasing function of $n \geq \cos \phi$ with $G(\cos \phi) = 0$ and $\lim_{n \rightarrow \infty} G(n) = 1 - \cos \phi > 0$. Therefore there exists some $n_{\min} \geq \cos \phi$ such that, for all $n \geq n_{\min}$, we have $F(n) \leq G(n)$. Thus, (33) holds for all $n \geq n_{\min}$.

We have now shown that the common opinion of agents $3, 4, \dots, N$ moves from $v_* \neq 0$ at $t = 0$, to 0 at $t = 1$, back to v_* at $t = 2$. In the meantime, agents 1 and 2 do not move since that they are too ‘stubborn’ to listen to any opinions in $[0, v_*]$. Thus, the system has returned at $t = 2$ to its original state, and will continue to oscillate with period 2. We have therefore constructed an N -body system, with explicitly specified parameters and initial condition, which follows periodic dynamics. It is interesting that this particular construction is possible only if the number of agents sharing the oscillatory opinion is sufficiently large, i.e. $n - 1 \geq n_{min} - 1$.

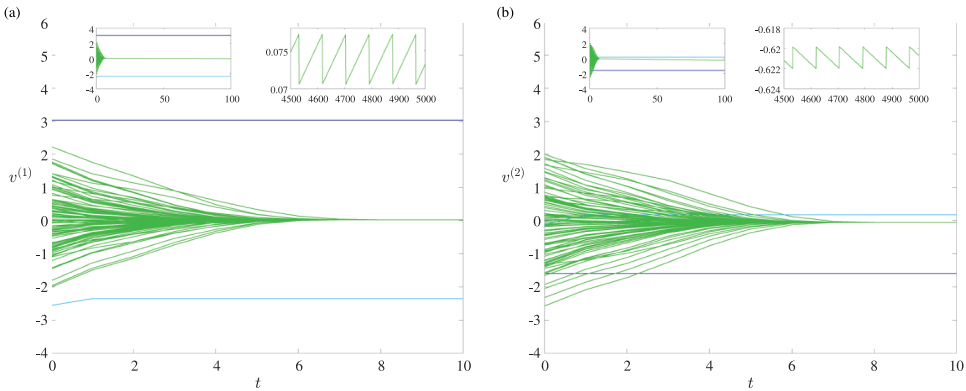


Figure 13. Example of a system that fails to reach a steady state under evolving heterogeneous thresholds. The time periods 0–10, 0–100, and 4500–5000 are shown. In particular, the top-right inset in each panel shows only the large-time dynamics of the oscillatory cluster. Panels (a,b) show the first and second dimensions of the opinions, respectively. Each cluster is shown in a different color. Parameters: $D = 2, \mu = 2, \rho = 0, a = 0.4$.

This condition is borne out by our numerical simulations of the model (even in higher dimensions and with larger memory capacities), where we see oscillations of the ‘neutral majority’ being pulled back and forth by a small number of extreme agents. In our simulations, whenever a system fails to reach a steady state, a number of stable clusters are formed, while the remaining agents form an unstable cluster that oscillates collectively by small amounts $\sim \mathcal{O}(10^{-3})$ along each dimension (see, [Figure 13](#)). These collective oscillations have a long timescale compared to the memory capacity of the population. Moreover, the oscillatory cluster is always the majority, having more members than any of the stable clusters. The oscillations are facilitated by the majority agents’ evolving thresholds. As exemplified by [Figure 13](#), while the majority cluster near position $(0.07, -0.62)$ moves toward the neutral $(0, 0)$ position due to an attraction to the fringe cluster near position $(-2.35, 0.18)$, the majority agents’ thresholds decrease according to Equation (6). When these thresholds become sufficiently low, the fringe agents further away ‘on the other

side' become able to exert influence on the majority, pulling them back toward the other extreme. While the majority move away from the neutral position, their thresholds increase again until they become so high that only the fringe cluster closest to them, near position $(-2.35, 0.18)$, can exert influence. This oscillatory process continues indefinitely. While the moderate majority swing from one position to another, failing to settle, the peripheral agents hold firm their positions, having such high thresholds that they fail to listen to any other cluster.

4. Conclusions and future directions

We have presented a novel agent-based model of opinion dynamics capable of mimicking many socio-psychological phenomena. The model extends several existing frameworks through bespoke elements such as an agent's interaction threshold (generalizing the confidence bound), a measure of pairwise affinity between agents, and a system-wide memory capacity. The resulting dynamics is a non-Markovian, nonlinear process of opinion updating. We have analyzed the mathematical properties of the model, and explored the rich variety of simulated behavior that emerges from the dynamics, focusing on consensus, segregation, and extremisation.

The agents' interaction thresholds are assigned in one of two ways: either prescribing a universal and constant threshold for all agents, or allowing each agent to evolve their own threshold such that the more extreme agents are less susceptible to change. When all agents are given a universal threshold, the system achieves a steady state of either consensus or segregation. We have proved that if all agents are assigned a sub-critical universal threshold $\rho < \rho_*$, where ρ_* is dependent on parameters μ and D as per (13), then consensus is formed regardless of the initial configuration of opinions, and the consensus view equals the average (mean) opinion of the initial state. The system transitions from consensus to segregation as the interaction threshold increases. Through numerical simulations, we have investigated the effects of the model parameters on the opinion clustering, convergence time, and opinion drift. It is found that a high universal threshold promotes segregation in generic D -dimensional opinion space, extending similar findings by Hegselmann and Krause (2002) in one-dimensional opinion space. The simulations also reveal that the connectome of the population becomes more disconnected as the opinions evolve, and the rate at which the connectome rewires itself is strongly dependent on the system's memory capacity. The opinion dynamics can be seen to represent a process of seeking cooperation, reflecting recent theoretical and experimental results (Rand et al., 2011).

In the case where the agents individually evolve their thresholds with some reinforcement rate (a model parameter controlling the rate at which agents

become more stubborn), we have examined the system's clustering behavior. Steady states are not always achieved in this case. By explicitly constructing an N -body system that forms an oscillatory cluster near the neutral position, we have proven that the model admits periodic solutions. Extreme agents 'on either side' of the cluster exert their influence in turn, resulting in the oscillations. The construction shows that periodic solutions are possible only if the oscillatory cluster is sufficiently large. Numerical simulations reveal oscillatory behavior of large clusters under various parameter settings. Both the analytic and numerical results in [Section 3.4](#) demonstrate the power of stubborn fringe agents over the neutral majority. By introducing an extremisation measure, we have quantified the extent to which the collective opinion becomes more extreme over time. Extremisation is maximized when the baseline threshold (of entirely neutral agents) is small but the reinforcement rate is large. A population that takes a longer history of itself into account (larger memory capacity) is less likely to become extremised than a population that quickly forgets the past. These results echo the socio-psychological phenomena of group polarization (Moscovici & Zavalloni, 1969; Myers & Lamm, 1976) and online extremism (Z. Z. Cao et al., 2018), providing a mechanistic explanation for the behaviors. When extremisation is large, it tends to involve a process of collective drift, where a large cluster of moderate agents moves toward a small cluster of extremists. The fact that extremisation occurs when fringe agents have a low tolerance to others corroborates the theory of Deffuant et al. (2000).

For simplicity of methodology and ease of interpretation, we have assumed that the initial opinions in each dimension of opinion space follow a normal distribution. It is worth reiterating that the system's subsequent behaviors are rich in variety *despite* the simplistic initial states. We expect an even richer range of phenomena to emerge from more sophisticated initial opinion distributions that may be better fits for real-world scenarios. For example, when a new political issue arises and a population forms initial opinions on the matter, those opinions may already be polarized rather than normally distributed, especially if media-driven tribalisation encourages immediate segregation (Llewellyn & Cram, 2016; Meredith & Richardson, 2019). The current model is capable of simulating the opinion dynamics in this context; one simply needs to input the appropriate data describing the initial opinions of the population. Moreover, when modeling multi-dimensional issues, it may be appropriate to sample initial opinions from correlated distributions, rather than independent distributions as we have done in this paper (Bartels, 2018).

It is also worth noting that other frameworks for performing a stability analysis on state dependent networks exist (Etesami, 2019; Proskurnikov & Tempo, 2018). In particular, the paper by Etesami (2019) contains some mathematical tools that could be used in the future for analysis of the model we have proposed. Other potential extensions to the model may include: a repulsive force, where low-affinity pairs do not merely ignore each other

but actively move away from each other's views; stochastic fluctuations in the agents' interaction thresholds, representing externally-driven variations in one's openness to other people; and a hierarchical population where some agents are assigned a much higher interaction threshold than the majority, describing powerful individuals exerting influence with little reciprocation. Overall, the modeling framework developed in this study generates various sociologically relevant phenomena under simple assumptions, while being sufficiently versatile to suit more elaborate contexts, and integration with experimental data in future work will help to further enhance the theory.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project is supported by the British Academy, grant number SRG1920_101649. The computer simulations are performed on the BlueBEAR HPC system at the University of Birmingham, UK. BMS is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/S022945/1. JL thanks Samuel Johnson (University of Birmingham) for useful discussions and thanks the University of Birmingham for fellowship funding.

ORCID

Jingxi Luo  <http://orcid.org/0000-0003-3032-9950>

References

- Alizadeh, M., Cioffi-Revilla, C., & Crooks, A. (2015). The effect of in-group favoritism on the collective behavior of individuals' opinions. *Advances in Complex Systems*, 18(01n02), 1550002. <https://doi.org/10.1142/S0219525915500022>
- Anderson, B. D., & Ye, M. (2019). Recent advances in the modelling and analysis of opinion dynamics on influence networks. *International Journal of Automation and Computing*, 16(2), 129–149. <https://doi.org/10.1007/s11633-019-1169-8>
- Artime, O., Peralta, A. F., Toral, R., Ramasco, J. J., & San Miguel, M. (2018). Aging-induced continuous phase transition. *Physical Review E*, 98(3), 032104. <https://doi.org/10.1103/PhysRevE.98.032104>
- Banisch, S., & Olbrich, E. (2019). Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2), 76–103. <https://doi.org/10.1080/0022250X.2018.1517761>.
- Bartels, L. M. (2018). Partisanship in the Trump era. *The Journal of Politics*, 80(4), 1483–1494. <https://doi.org/10.1086/699337>
- Benatti, A., de Arruda, H. F., Silva, F. N., Comin, C. H., & da Fontoura Costa, L. (2020). Opinion diversity and social bubbles in adaptive Sznajd networks. *Journal of Statistical*

- Mechanics: Theory and Experiment*, 2020(2), 023407 <https://doi.org/10.1088/1742-5468/ab6de3>.
- Bentley, R. A., Ormerod, P., & Batty, M. (2011). Evolving social influence in large populations. *Behavioral Ecology and Sociobiology*, 65(3), 537–546. <https://doi.org/10.1007/s00265-010-1102-1>
- Blondel, V. D., Hendrickx, J. M., Olshevsky, A., & Tsitsiklis, J. N. (2005). Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the 44th IEEE conference on decision and control* (pp. 2996–3000).
- Cao, M., Morse, A. S., & Anderson, B. D. (2008). Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM Journal on Control and Optimization*, 47(2), 575–600. <https://doi.org/10.1137/060657005>
- Cao, Z., Zheng, M., Vorobyeva, Y., Song, C., & Johnson, N. F. (2018). Complexity in individual trajectories toward online extremism. *Complexity*, 2018, 3929583. <https://doi.org/10.1155/2018/3929583>
- Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2), 591. <https://doi.org/10.1103/RevModPhys.81.591>
- Cheng, C., & Yu, C. (2019). Opinion dynamics with bounded confidence and group pressure. *Physica A: Statistical Mechanics and Its Applications*, 532, 121900. <https://doi.org/10.1016/j.physa.2019.121900>
- Cucker, F., & Smale, S. (2007). Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5), 852–862. <https://doi.org/10.1109/TAC.2007.895842>
- Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15), 5791–5796. <https://doi.org/10.1073/pnas.1217220110>
- Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04), 87–98. <https://doi.org/10.1142/S0219525900000078>
- DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345), 118–121. <https://doi.org/10.1080/01621459.1974.10480137>
- Etesami, S. R. (2019). A simple framework for stability analysis of state-dependent networks of heterogeneous agents. *SIAM Journal on Control and Optimization*, 57(3), 1757–1782. <https://doi.org/10.1137/18M1217681>.
- Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. *The Journal of Mathematical Sociology*, 35(1–3), 146–176. <https://doi.org/10.1080/0022250X.2010.532261>
- Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 4. <https://doi.org/10.18564/jasss.3521>
- French, J. R., Jr. (1956). A formal theory of social power. *Psychological Review*, 63(3), 181–194. <https://doi.org/10.1037/h0046123>
- Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3–4), 193–206. <https://doi.org/10.1080/0022250X.1990.9990069>
- Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*, 78(2), 026117. <https://doi.org/10.1103/PhysRevE.78.026117>
- Galesic, M., & Stein, D. L. (2019). Statistical physics models of belief dynamics: Theory and empirical tests. *Physica A: Statistical Mechanics and Its Applications*, 519, 275–294. <https://doi.org/10.1016/j.physa.2018.12.011>
- Hanaki, N., Peterhansl, A., Dodds, P. S., & Watts, D. J. (2007). Cooperation in evolving social networks. *Management Science*, 53(7), 1036–1050. <https://doi.org/10.1287/mnsc.1060.0625>

- Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5, 3. <http://jasss.soc.surrey.ac.uk/5/3/2.html>.
- Hendrickx, J. M., Shi, G., & Johansson, K. H. (2014). Finite-time consensus using stochastic matrices with positive diagonals. *IEEE Transactions on Automatic Control*, 60(4), 1070–1073. <https://doi.org/10.1109/TAC.2014.2352691>
- Hollewell, G. F., & Longpré, N. (2022). Radicalization in the social media era: Understanding the relationship between self-radicalization and the Internet. *International Journal of Offender Therapy and Comparative Criminology*, 66(8), 896–913. <https://doi.org/10.1177/0306624X211028771>
- Holley, R. A., & Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *The Annals of Probability*, 3(4), 643–663. <https://doi.org/10.1214/aop/1176996306>
- Huet, S., Deffuant, G., & Jager, W. (2008). A rejection mechanism in 2d bounded confidence provides more conformity. *Advances in Complex Systems*, 11(4), 529–549. <https://doi.org/10.1142/S0219525908001799>
- Kononovicius, A. (2021). Supportive interactions in the noisy voter model. *Chaos, Solitons & Fractals*, 143, 110627. <https://doi.org/10.1016/j.chaos.2020.110627>
- Kozitsin, I. V. (2020). Formal models of opinion formation and their application to real data: Evidence from online social networks. *The Journal of Mathematical Sociology* 46, (2), 120–147. <https://doi.org/10.1080/0022250X.2020.1835894>.
- Kurahashi-Nakamura, T., Mäs, M., & Lorenz, J. (2016). Robust clustering in generalized bounded confidence models. *Journal of Artificial Societies and Social Simulation*, 19(4), 7. <https://doi.org/10.18564/jasss.3220>
- Lewis, A. D. (2010). A top nine list: Most popular induced matrix norms. *Queen's University, Kingston, Ontario, Tech. Rep.* 1–13 <https://mast.queensu.ca/~andrew/notes/pdf/2010a.pdf>.
- Liu, C. C., & Srivastava, S. B. (2015). Pulling closer and moving apart: Interaction, identity, and influence in the U.S. Senate, 1973 to 2009. *American Sociological Review*, 80(1), 192–217. <https://doi.org/10.1177/0003122414564182>
- Llewellyn, C., & Cram, L. (2016). Brexit? Analyzing opinion on the UK-EU referendum within Twitter. *Proceedings of the International Aaai Conference on Web and Social Media*, 10, 760–761 <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/viewPaper/13119>.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109. <https://doi.org/10.1037/0022-3514.37.11.2098>
- Lorenz, J. (2005). A stabilization theorem for dynamics of continuous opinions. *Physica A: Statistical Mechanics and Its Applications*, 355(1), 217–223. <https://doi.org/10.1016/j.physa.2005.02.086>.
- Mariano, S., Morărescu, I., Postoyan, R., & Zaccarian, L. (2020). A hybrid model of opinion dynamics with memory-based connectivity. *IEEE Control Systems Letters*, 4(3), 644–649. <https://doi.org/10.1109/LCSYS.2020.2989077>
- Meredith, J., & Richardson, E. (2019). The use of the political categories of Brexiter and Remainer in online comments about the EU referendum. *Journal of Community & Applied Social Psychology*, 29(1), 43–55. <https://doi.org/10.1002/casp.2384>
- Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist*, 60(2), 161. <https://doi.org/10.1037/0003-066X.60.2.161>
- Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of personality and social psychology*, 12(2), 125–135 <https://psycnet.apa.org/doi/10.1037/h0027568>.

- Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, 83(4), 602–627. <https://doi.org/10.1037/0033-2909.83.4.602>
- Nedić, A., & Liu, J. (2016). On convergence rate of weighted-averaging dynamics for consensus problems. *IEEE Transactions on Automatic Control*, 62(2), 766–781. <https://doi.org/10.1109/TAC.2016.2572004>
- Noorazar, H., Vixie, K. R., Talebanpour, A., & Hu, Y. (2020). From classical to modern opinion dynamics. *International Journal of Modern Physics C*, 31(7), 2050101 <https://doi.org/10.1142/S0129183120501016>.
- Proskurnikov, A. V., & Tempo, R. (2018). A tutorial on modeling and analysis of dynamic social networks. part ii. *Annual Reviews in Control*, 45, 166–190. <https://doi.org/10.1016/j.arcontrol.2018.03.005>
- Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48), 19193–19198. <https://doi.org/10.1073/pnas.1108243108>
- Ren, W., & Beard, R. W. (2005). Consensus seeking in multiagent systems under dynamically changing interaction topologies. *IEEE Transactions on Automatic Control*, 50(5), 655–661. <https://doi.org/10.1109/TAC.2005.846556>
- Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006). Cooperation prevails when individuals adjust their social ties. *PLoS Computational Biology*, 2(10), e140. <https://doi.org/10.1371/journal.pcbi.0020140>
- Schweighofer, S., Schweitzer, F., & Garcia, D. (2020). A weighted balance model of opinion hyperpolarization. *Journal of Artificial Societies and Social Simulation*, 23(3), 5. <https://doi.org/10.18564/jasss.4306>
- Stadtfeld, C., Takács, K., & Vörös, A. (2020). The emergence and stability of groups in social networks. *Social Networks*, 60, 129–145. <https://doi.org/10.1016/j.socnet.2019.10.008>
- Stark, H.-U., Tessone, C. J., & Schweitzer, F. (2008a). Decelerating microdynamics can accelerate macrodynamics in the voter model. *Physical Review Letters*, 101(1), 018701. <https://doi.org/10.1103/PhysRevLett.101.018701>
- Stark, H.-U., Tessone, C. J., & Schweitzer, F. (2008b). Slower is faster: Fostering consensus formation by heterogeneous inertia. *Advances in Complex Systems*, 11(4), 551–563. <https://doi.org/10.1142/S0219525908001805>
- Tian, Y., Jia, P., Mirtabatabaei, A., Wang, L., Friedkin, N. E., & Bullo, F. (2021). Social power evolution in influence networks with stubborn individuals. *IEEE Transactions on Automatic Control* 67(2) doi:10.1109/TAC.2021.3052485.
- Turner, M. A., & Smaldino, P. E. (2018). Paths to polarization: How extreme views, miscommunication, and random chance drive opinion dynamics. *Complexity*, 2018, 2740959. <https://doi.org/10.1155/2018/2740959>
- Vicsek, T., & Zafeiris, A. (2012). Collective motion. *Physics Reports*, 517(3–4), 71–140. <https://doi.org/10.1016/j.physrep.2012.03.004>.
- Ye, M., Qin, Y., Govaert, A., Anderson, B. D., & Cao, M. (2019). An influence network model to study discrepancies in expressed and private opinions. *Automatica*, 107, 371–381. <https://doi.org/10.1016/j.automatica.2019.05.059>