Journal of Complex Networks (2020) **2**, Advance Access Publication on 5 May 2020
doi: 10.1093/comnet/cnaa010

The polarization within and across individuals: the hierarchical Ising opinion model

HAN L. J. VAN DER MAAS[†]
Institute of Advanced study, University of Amsterdam, Amsterdam, Netherlands and Department of Psychology, University of Amsterdam, Building G, Room 0.37, Nieuwe Achtergracht 129-B, 1018WS Amsterdam, Netherlands
[†]Corresponding author. Email: h.l.j.vandermaas@uva.nl

AND

JONAS DALEGE AND LOURENS WALDORP
Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

Edited by: Yamir Moreno

[Received on 8 October 2019; editorial decision on 4 February 2020; accepted on 10 February 2020]

Polarization of opinions is a societal threat. It involves psychological processes as well as group dynamics, a popular topic in statistical physics. However, the interaction between the within individual dynamics of attitude formation and across person polarization is rarely studied. By modelling individual attitudes as Ising networks of attitude elements, and approximating this behaviour by the cusp singularity, we developed a fundamentally new model of social dynamics. In this hierarchical model, agents behave either discretely or continuously depending on their attention to the issue. At the individual level, the model reproduces the mere thought effect and resistance to persuasion. At the social level, the model implies polarization and the persuasion paradox. We propose a new intervention for escaping polarization in bounded confidence models of opinion dynamics.

Keywords: attitudes; polarization; persuasion; Ising model; cusp catastrophe.

1. Introduction

The polarization of opinions is an important and increasing societal problem [1, 2]. Polarization across individuals leads to the formation of distinct camps which prevents us from reaching consensus on issues such as health care, education and climate [3]. Individual cases of polarization, radicalization for instance, may lead to harmful extremist behaviours [4].

Various scientific disciplines study these types of processes. Psychology studies the formation of attitudes in individuals, while sociology and political science are concerned with the collective properties of polarization. These collective properties have also become popular topics in statistical physics and computer science. Over the last decades, the statistical physics of social dynamics, or sociophysics, has become a field in itself, with many different approaches to the formal modelling and simulation of social phenomena [5, 6].

As both the individual and collective processes are extremely complex, models of the individual tend to ignore or greatly simplify group processes and vice versa. In statistical physics models of opinion dynamics, for instance, individuals are often reduced to binary state systems that tend to switch to the majority opinion in their local environment [7]. Simplifications are indeed required but a richer model of the individual agent within models of social dynamics may provide new insights in the dynamics of

© The authors 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

polarization. This article presents such a new model, the hierarchical Ising opinion model (HIOM), for the integrative study of individual and collective polarization.

Specifically, we will first provide a new answer to a famous question posed by Axelrod [8]: 'If people tend to become more alike in their beliefs, attitudes, and behaviour when they interact, why do not all such differences eventually disappear?' There are different ways to model this lack of consensus in dynamic opinion models. Generally, this is modelled in terms of limited interaction between agents. In Axelrod's model, for instance, this was due to selective interaction between agents. In the HIOM, this effect is due to hysteresis (explained below) within each agent, which is possible because we use a more realistic model of the agents. Second, we explicate polarization between agents in a new way. In the HIOM attempts of activists to influence, the common sense occasionally create opposition and thereby polarization. For this effect, the interaction of within person and between person dynamics is essential. Third, we derive a new prediction on avoiding polarization in cases where polarization is caused by selective interaction.

The article is organized as follows. We first introduce a formal model for individual attitudes in which attitudes are conceptualized as networks of interacting attitude elements. Among other things, this within agent model clarifies why it is difficult to change highly involved persons' opinions. We then show that the complex dynamics of such attitude networks can be summarized by a stochastic cusp model. This reduced model is used in the second part of the article to model the individual-level behaviour in an agent-based model of opinion dynamics. By simulations, we show that enriching the behaviour of agents in this way yield many new interesting avenues for models of social dynamics.

2. The Ising model of attitudes

In the understanding, modification and prediction of human behaviour, the concept of attitudes plays a central role. Attitudes serve multiple purposes, such as guiding our behaviour and organizing knowledge [9]. Attitudes are formed through several different processes [10–12] and they range from being highly stable and impactful to fluctuating and inconsequential [13]. Stable and consistent attitudes are thought to be essential for human functioning. The empirical data base on attitudes in social psychology is massive and there exists a rich variety of theories on what the impact of attitudes are on behaviour [14], if and how they change as a result of exposure to persuasive messages [15, 16], and how people strive for consistency in their attitudes [17].

Formal models of attitudes are rare but, recently, a network model of attitudes, based on the Ising model has been introduced [18–21]. The Ising model was originally proposed for ferromagnetism but has been applied to various other phenomena, such as image segmentation [22], voice recognition [23] and spatial statistics [24]. This type of analogical modelling is often highly useful in social science research, as it allows borrowing formalisms from more advanced research fields [25]. The limitation is that a full quantitative application is out of scope, but as we will show here, there are many interesting qualitative properties that can be derived from such a model. The most important qualitative property of the Ising model, modelled on dimensions higher than one, is that it exhibits a first order phase transition.

In the Ising attitude model, an attitude is defined as a network of many interacting nodes (see Fig. 1, top left panel). These nodes represent beliefs (meat production impacts climate), feelings (loves steak) and behaviours (eat burger) that relate to the attitude object (meat eating). Adopting the Ising model comes with advantages, as we will show below, but also requires simplifying assumptions, which may be judged untenable.

The first assumption is that nodes (representing the attitude elements) can be characterized as two-valued systems (e.g. 'on' versus 'off', 'pro' versus 'con'). Whether this is reasonable depends on the definition of nodes. We would probably express global descriptions of behaviour, such as one's vegetarian

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

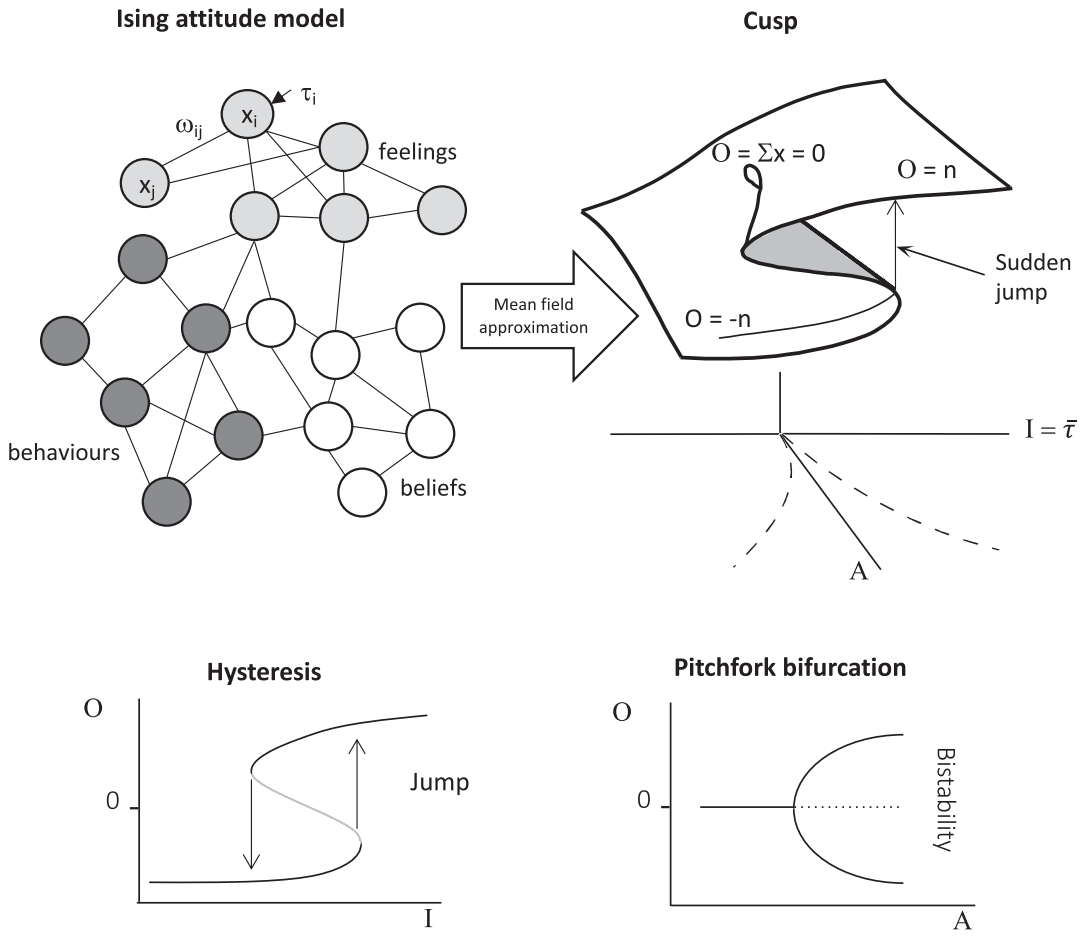

FIG. 1. The Ising attitude model and its approximation by the cusp singularity. The network consists of binary nodes x representing feelings, beliefs and behaviours towards the attitude object. Dispositions, such as external influences, are denoted by τ. Connections ω between nodes are symmetrical. In the cusp, the sum of all node values, Σx or O (opinion), has either one or three fixed points. In the case of three fixed points, the middle one is unstable (located in the grey area). The number of fixed points depend on the control variables I = mean(τ), the informational external influences, and A, the attention paid to the attitude object. Two intersections of the cusp are hysteresis and the pitchfork bifurcation. Hysteresis possibly explains resistance to persuasion; the pitchfork describes individual polarization.

behaviour on some ordinal scale (from not at all to very consistent), but on a more fine-grained level (consumes horse meat, wears leather shoes, etc.) a binary scale often suffices. Non-binary, multi-categorical, nodes are possible and are captured by related models [26].

The second assumption of the standard Ising model is that the connections between nodes are symmetrical (undirected). Many connections between attitude nodes are probably symmetric but this might sometimes be unrealistic. For instance, the connection between belief nodes (meat is expensive) and behaviours (buying meat) are probably asymmetrical. In Appendix A, we discuss the directed Ising

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

model and show by simulation that allowing for asymmetrical positive connections weights will not change the results in our case.

The third assumption is that nodes have dispositions. These dispositions represent the probability of a given attitude node to be endorsed or not when there were no effects of other nodes in the network. One way to think about these dispositions is in terms of external social influences (cf. external field), such as the common sense on an issue like the environmental effects of meat consumption. Dispositions can be node specific or equal for all nodes.

Variables associated with elements or nodes are denoted by $x_i \in \{+1, -1\}$, connections weights by ω_{ij} and dispositions by τ_i (i in 1..n). We will summarize the mean effect of the dispositions $\bar{\tau}$ by the concept of information $I(I = \bar{\tau})$. In the between subject model, explained in Section 6, agents exchange information, meaning that they influence each other's dispositions of attitude nodes.

Based on these assumptions, a measurement theory for attitudes has been developed and applied to several datasets [20]. While psychologists often invoke latent variables to explain correlations between indicators or symptoms of constructs like attitudes, the Ising model of attitudes does not require such an explanation. In the Ising model of attitudes, stable attitudes are emergent phenomena. Main concepts in the attitude literature, such as attitude strength, ambivalence and cognitive consistency can be precisely defined in network terms [18].

Several papers describe the formal relationship between the Ising model and popular statistical models, such as the log-linear model, the logistic regression and item response theory models [27]. These relationships are another reason to accept the simplifying assumptions of the standard Ising model when applied to psychological constructs.

3. Entropy reduction

Given these first assumptions it is useful to define the micro- and macrostate of an attitude. The configuration $\mathbf{x} = (x_1, \ldots, x_n)$ of the attitude elements constitutes the microstate of the attitude, whereas the macrostate is defined as the number of positive versus negative attitude elements $\Sigma\mathbf{x}$ (similar to the definition of magnetization). The global evaluation of an attitude object, or opinion, can be defined as a situation-dependent weighted sum score [21]. Here, we simply define opinion O as $\Sigma\mathbf{x}$, the variable of interest in the between subject model that will be described later.

The relation between the microstates and the macrostate has important implications for the dynamics of attitudes. A disorganized attitude, consisting of many random microelements, is associated with 'a close to zero' macrostate, implying an inconsistent attitude. The Boltzmann entropy (Sb) of the network can thus be understood as the inconsistency of an attitude in the sense that it relates to the number W of microstates that can realize the macrostate (Sb = lnW). The Boltzmann entropy therefore describes how (dis)organized an attitude with a given macrostate is.

The second law of thermodynamics, a famous law in the general theory of these types of systems, implies that the entropy of an isolated system always increases (although not necessarily monotonically [28]). If humans existed in isolation, it can be stated, based on this law, that the inconsistent and unstable state (high attitudinal entropy) is the natural state of an attitude. An important question is thus how stable and consistent attitudes are developed and maintained [29].

To model the development of stable, low entropy, attitudes, we need one additional assumption. The fourth assumption is that the probability of an element to behave in accordance with the state of the network depends on how much one focuses attention on or thinks about the attitude or attitude object. The state is defined by the values of neighbouring elements, the strength of connections and the dispositions of elements. Thus, we assume that attention, denoted by A, has an analogous effect on

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

attitude representations as inverse temperature has on thermodynamic systems. If attention to the attitude object is low elements tend to behave randomly; if attention is high, then elements are aligned. How much one attends to or thinks about an attitude object depends on other factors, such as involvement or importance [30].

To state these assumptions mathematically, we adopt the standard Hamiltonian energy function associated with this type of Ising network:

$$H(x) = -\sum_{i} \tau_i x_i - \sum_{\langle i,j \rangle} \omega_{ij} x_i x_j. \tag{1}$$

Thus, states of nodes that are incongruent to their associated dispositions cause higher amounts of energy. The same holds for positively connected nodes that are in incongruent states. The energy determines the probability of states according to the following equation:

$$Pr(X = x) = \frac{exp(-AH(x))}{Z}, \tag{2}$$

where Z is an integration constant such that the probabilities sum to one and A represents attention. Note that if $A = 0$, all states are equally probable and entropy will be at its maximum.

These equations are instrumental in simulation dynamics for the Ising model. In every iteration of these Glauber dynamics [31], a random node is first selected. It is then computed how much the energy of the network changes if the node is flipped (ΔH). The probability of such spin flips follows from the Boltzmann factor and is $1/(1 + \exp(-A\Delta H))$. Thus, for $A = 0$ flips are random, and for high A the lower energy state is almost always preferred. In the present case, this means that with sufficient attention people tend to reduce inconsistency in their attitude networks. We note that this setup of the model is remarkably consistent with the root psychological theory of attitudes, the theory of cognitive dissonance [17].

Note that at least two types of high entropy attitudes exist. The first type consists of 'unattended' attitudes, cases where one simply has never thought about this issue. The second type, in contrast, concerns issues that one cares about, but consistency is hard to achieve due to an irregular configuration of dispositions and connection weights. For many smokers, for instance, the urge to smoke (the disposition for the main behavioural node) conflicts with the opposite influence of the neighbouring nodes (smoking causes cancer). The behaviour of these high entropy attitudes is intriguing and could involve a sudden transition to a new equilibrium state, as explained in the next section.

4. The dynamics of attitudes

The behaviour of the Ising model has been studied in great detail. We are specifically interested in the dynamics of the macrostate, the sum of all microstates (nodes), because it determines the overall opinion. It has been shown, using the mean field approximation, that the equilibrium behaviour of the Ising model, under rather general assumptions, can be described by the cusp catastrophe [32, 33]. Formally, the cusp is a second order Taylor approximation of the Curie Weiss solution [34]. It is approximately correct for many instances of the Ising model. In Appendix A, this is further discussed.

The equilibria of the cusp catastrophe are given by the cubic equation:

$$Y^3 - aY - b = 0, \tag{3}$$

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

where Y is called the behaviour variable, a is called the splitting variable and b is called the normal variable. Many phase transitions can be modelled using the cusp [35]. For instance, for the phase transition between states of water (solid to liquid), a approximately equals pressure and b temperature.

Also, for the Ising attitude model the definitions of the normal and splitting axes are straightforward. The behavioural variable relates to opinion, defined as the sum of the node values ($Y \sim O = \Sigma x$). The normal variable or axis relates to information ($b \sim I = \bar{\tau}$) and the splitting axis represent attention ($a \sim A$). This makes the understanding of the dynamics of the Ising attitude model relatively simple. Figure 1 displays the cusp, its relation to the Ising model and two intersections.

Starting at the back of the cusp, at $I = A = 0$, increasing attention A leads to a pitchfork bifurcation, as the middle state $O = 0$ becomes unstable and O bifurcates to either strongly positive or negative values. This is the first main prediction of the Ising model of attitudes. Increasing attention or thought leads to polarization of attitudes. There is ample evidence for this so-called mere thought effect [36].

For values of A close to zero, at the back of the cusp, changing information I from negative to positive and vice versa leads to smooth changes in O. However, at the front of the cusp, when A is positive and behaviour is polarized, the effect of changing I is more intriguing. The change in I has hardly any effect until the original stable state disappears and a sudden jump to the alternative stable state takes place. This applies to increases and decreases in I but the sudden jumps 'up' and 'down' do not take place at the same value of I. This delay in the sudden jump is called hysteresis. This is the second main prediction. Informing people about the validity of the other attitude position (thus manipulating I) may not result in attitudinal change. Hysteresis possibly explains why persuasion is often so hard, especially when people are highly attentive to and involved in the issue [21, 37].

This explanation is an important result. As Eagly and Chaiken [38] noted: 'explaining why people are so often effective at resisting efforts to change their strong attitudes remains one of the core issues of attitude theory' (p. 680). One interesting implication of this prediction is that effective persuasion might require a decrease in A before I is changed. A mediator in a conflict should first lower attention (often a function of involvement) before a fruitful exchange of arguments can take place [39]. This might also explain why difficult negotiations often require prolonged meetings. When tiredness has diminished involvement, common ground can be reached. Note that the two main predictions are not built into the model. They follow from the elementary assumptions 1–4.

This is an advantage over earlier work in which the cusp catastrophe is introduced as a phenomenological model of attitudes [40, 41]. In this earlier work, the cusp is not based on a micromodel of first principles. Yet, this earlier work, showing for instance an increase of bimodality in attitude distributions with increasing involvement, is fully consistent with the approach in this article.

5. Networks of networks

The Ising model of attitudes focuses on single attitudes and models them without taking other attitudes into account. Each person, however, holds a large number of interrelated attitudes, which probably make up a huge network of attitude elements. In these networks, cliques of highly connected nodes define attitudes. A model that specifies these dynamics is as yet out of reach, and we defer this important question to future work. Here, we focus on another extension, that of between person interactions.

Whereas formal approaches to the within person dynamics of attitudes are rare, formal approaches to the between person dynamics are numerous. Opinion dynamics models are built on three types of assumptions [5].

The first assumption concerns the agents, especially the agent's opinions. Opinions can be continuous or categorical (often binary). For instance, in models such as the voter model, the Sznajd model and social

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

impact theory, opinions are binary, in the Axelrod model they are categorical, and in the Deffuant and the Hegselmann–Krause models opinions are continuous [42].

Second, the topology of interactions has to be chosen. In many models, individuals are located regularly in some space, often a two-dimensional grid (lattice) where each individual only interacts with its neighbours. Alternatively, individuals may interact with random partners or in some complex (growing) social network.

Third, the interactions between agents need to be defined. Again, there are many options. In Ising type models, an agent conforms with some probability to the majority opinion in its neighbourhood [43]. In the prototypical Ising models of opinion dynamics, agents have binary opinions (-1,1), are organized on a two-dimensional grid, and conform, applying the Glauber dynamics, to their neighbours depending on the (social) temperature parameter. In the majority rule model, groups of nodes conform to the majority of the group at once. In social impact theory [44], impact depends on distance between agents, and persuasiveness and supportiveness of agents. In models with continuous opinions bounded confidence plays a role: agents compromise in opinions but only when their original opinions are sufficiently similar.

A systematic overview of the various approaches to the modelling of social dynamics is beyond the scope here. Castellano *et al.* [5] speak of a real explosion of new models. In the majority of these models, the agents are relatively simple and one-dimensional. One exception is social impact theory where three variables describe the state of the agent but two of these variables are fixed parameters. In the multidimensional Hegselmann–Krause model [45], opinion is multidimensional but these are independent variables. Martins [46] studies a model in which opinions are continuous but actions are discrete. Another interesting case is the model of Sobkowicz [47]. He assumes the cusp model for the individual agent, using emotions and information as control variables. Interactions between agents lead to changes in opinion and the control variables. For instance, agitated agents make other agents also agitated. In his opinion model, Sobkowicz reduces the cusp dynamics to a three-state system, opinions are either -1, 0 or 1. We will not adopt this simplification as one of the central predictions of our model holds that whether opinions are discrete or continuous depends on the agent's attention to the attitude object.

6. The HIOM

As we argue in the first part of this article, the attitude of an agent is a multidimensional construct, which under some simplifying assumptions can be modelled by an Ising network. A combined model of the within and between attitude/opinion dynamics could thus be modelled as a hierarchical Ising model. Hierarchical or multi-layered network models, such as multi-layered neural networks and multi-layered voter models [48], are studied in many fields [49]. By using the mean field approximation of the Ising model within persons, such a model is tractable in the form of a network of cusps.

We describe each agent by a cusp model based on three variables, opinion O, attention A and information I (see Fig. 1). Note that I is a broad concept in this context—it could, for instance, consist of scientific facts, rumours or shared fears. O, A and I are continuous variables. Negative O and I reflect a contra opinion. A is constrained to be non-negative. In our model, the dynamics of O are described by a cuspoid stochastic differential equation [50, 51]:

$$dO_i = -\left(O_i^3 - (A_i + A^{\min})O_i - I_i\right)dt + s_O dW_i(t) \text{ for agent } i \text{ in} 1...N. \quad (4)$$

Change in opinion depends on opinion itself, attention and information. The rate of change depends on dt. The last term represents Wiener (Brownian) noise with variance s_O, incorporating effects on O not

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

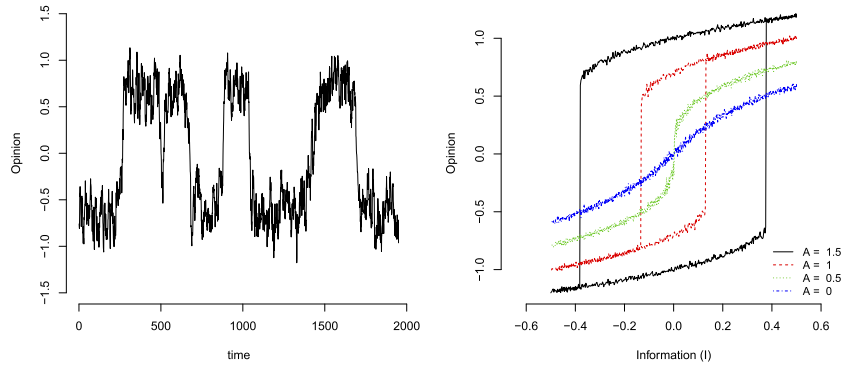

FIG. 2. The left panel (a) shows a typical times series for the 'ambivalent case' for fixed A and I for one agent (no interactions). In this case, attention is high ($A >> 0$) but information pro and contra balance out ($I = 0$). Due to noise, spontaneous sudden jumps between the two alternative stable states occur. The right panel (b) shows hysteresis in O as function of I for different values of A.

included in A and I. To illustrate the rich behaviour of this equation, Fig. 2a displays a typical bistable case for one agent ($s_O = 0.15, dt = 0.1, A = 1.5, I = 0, A^{min} = -0.5$). Figure 2b shows the hysteresis effect in opinion O as function of information I for different levels of attention ($s_O = 0.01, dt = 0.1$). To incorporate close to linear change in O as function of I, we set $a = A + A^{min}$, where $A^{min} = -0.5$ and $A \geqslant 0$.

Agents interact with other agents in a network. For the HIOM, the exact choice of social network topology is not essential. We experimented with various scale free and small-world networks, with similar qualitative results. In our main simulations, we apply the Watts–Strogatz model [52], the stochastic block model [53, 54] and the lattice model [55], which are all popular in social network analysis.

A central postulate of the HIOM is that effects of social interactions on opinion are not direct, but operate via A and I. The dynamics of A and I are governed by three assumptions.

The first assumption is that the probability that an agent initiates an interaction with a randomly selected neighbour depends on the agent's attention to the attitude object. We expect an agent who frequently attends to the attitude object (e.g. a highly involved agent) to initiate interactions more often than uninvolved agents (see [56] for a similar assumption). This asynchronous weighted agent selection is implemented by:

$$Pr\left(selectagent_i\right) = A_i / \sum_i A_i \text{ If } \forall A = 0 \text{ no agent is selected.} \tag{5}$$

The second assumption concerns the dynamics of attention. We assume that attention slowly decays over time if the agent is not involved in any interactions but increases due to interactions. This assumption is based on the notion that involvement or interest in an issue slowly decays. Interactions, a discussion for instance, generally increases attention to the object. This is formalized as follows:

$$dA_i = -\frac{2d_A}{N^2} A_i + d_A u_i \left(2A^* - A_i\right), \tag{6}$$

where $u_i = 1$ if the agent is involved in an interaction. Constant d_A determines the rate of change in A. A^* is the equilibrium state. Figure 3 shows an example of the time course of involvement ($d_A = 0.2, p(u = 1) = 0.05, A^* = 1$).

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

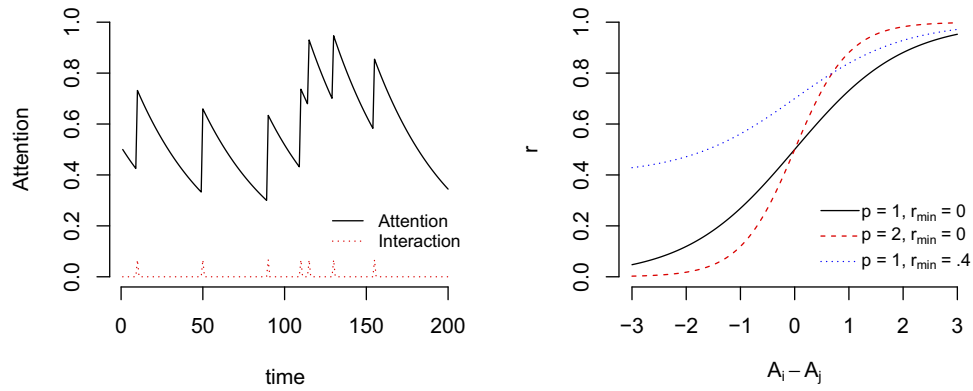

FIG. 3. The left panel demonstrates the dynamics of attention to an attitude: attention decreases slowly but increases when interactions (red spikes) take place. The right panel concerns the information update function. It shows resistance, r, as function of the difference in attention of agents i and j, modified by persuasion, p, and base resistance, r_{min}.

The third and last assumption concerns the update of information. We assume that information changes during interactions between agents. We expect that this exchange of information is an averaging process weighted by attention. If agent i is less attending to the attitude object than agent j, i moves more to j than j moves to i. This update of information is modelled by:

$$I_i = rI_i + (1 - r)\, I_j + \mathrm{N}\,(0, s_I)\,, \text{ where } r = r_{\min} + \frac{1 - r_{\min}}{1 + e^{-p(A_i - A_j)}}, \tag{7}$$

where I_i and I_j denote the information of the two agents i and j. Resistance, r in [0,1], determines the relative impact of agent j on agent i and is a logistic function of the difference in attention in agents i and j. If $A_i << A_j$, r will be close to zero and the information in agent i will change to the value of I in agent j. The strength of this effect is determined by the steepness, p, of the logistic function. We interpret p as a persuasion parameter. If p is high small changes in attention lead to large changes in information. Parameter r_{min} determines the minimal value of r, and functions as a base resistance parameter. If r_{min} is high, r will be high and agents will stick to their informational state. If $s_I > 0$ some normally distributed noise is added. Adding noise prevents the variance in information to converge to zero.

With these three assumptions the HIOM is complete. We specified the opinion dynamics of agents (Equation 4), a topology and dynamics of interactions (Equations 5–7). With regard to the second assumption, we note that we could not find empirical underpinning for this seemingly noncontroversial assumption. The role of involvement in attitudes has been extensively studied but not the dynamics of involvement itself. With regard to the third assumption, we note that this dynamic is clearly a gross simplification of what can happen in social interactions [57]. The weighted averaging of information can however be justified by analysing the effect of coupling two Ising networks of attitude elements. The higher involved and attentive agent will be more consistent and thus strongly influence the (inconsistent) states of the uninvolved agent.

To summarize, more realistic (and complicated) assumptions and definitions are possible, but our current setup suffices to reproduce important polarization phenomena and to derive new predictions from the HIOM.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

7. Hypotheses

The main novelty of the HIOM is that agents may behave continuously or categorically depending on their attention to the subject. When agents behave categorically, they may display hysteresis. As far as we know hysteresis within agents is a new way to explain polarization and a new answer to Axelrod's question on prevailing differences between people even when they are alike in their underlying beliefs (e.g. information). Simulation 1, reported below, demonstrates that polarization is an enduring effect even when the original cause of polarization (differences in information) evaporated. We show that hysteresis indeed causes this effect.

We hypothesized that hysteresis within agents may also lead to polarization in rather unexpected ways. An interesting puzzle in the dynamics of opinion change is that some opinions remain fairly neutral for a long time but then suddenly polarize [58]. An example of such dynamics is the Dutch black Pete discussion [59]. The black Pete character is part of the annual children's feast of St. Nicholas, celebrated on the evening of 5 December in the Netherlands. For a long time, people's attitudes towards black Pete could be characterized as slightly positive and by low involvement and thus low attention to the attitude object. In the last 15 years, the debate about black Pete (whether or not it is a racial stereotype) became tremendously polarized. Somehow, the activists did not convince the whole uninvolved majority but created fierce opposition in some people of the majority. We call this the persuasion paradox.

This persuasion paradox, related to, for instance, the boomerang effect [60] and the backfire effect [61], refers to the phenomenon that an attempt to persuade someone sometimes results in the adoption of an opposing position. The theory of psychological reactance [62, 63] explains this effect using the concept of freedom; if freedom is threatened, people become motivated to restore it. Trevors *et al.* [64] list many examples of this effect and provide an explanation in terms of negative emotions (cf. involvement). The HIOM produces this paradox in a similar but formal way. In the second simulation, we demonstrate this persuasion paradox in the HIOM.

Our last application of the HIOM concerns a solution or escape from polarization in continuous opinion models with bounded confidence. Polarization in such models is generally attained by restrictions on the interactions between agents [65].

Hegselmann and Krause [66] introduced a typology of continuous opinion models. In the simplest form, the interaction weights are fixed and independent of O (De Groot model). In the Friedkin–Johnsen model [67] a susceptibility parameter is added. If susceptibility is low, agents are not influenced socially and stick to their initial opinion. The base resistance parameter in equation 7 is based on a similar idea (susceptibility being equal to 1 - r_{min}). The third and most important form is bounded confidence models. In these models, examples being the Deffuant–Weisbuch model and the Hegselmann–Krause model, opinion adjustments only take place when the absolute difference in opinions is below some threshold t_O. This idea resembles the concept of latitude of rejection in social judgment theory [68, 69]. We incorporate bounded confidence by adding the condition $|O_i - O_j| < t_O$ to the model. If this condition is not met neither A nor I is updated.

Hegselmann and Krause provide analytical proof and simulations on the convergence to consensus of these models. Generally, they conclude that consensus is difficult when (a) (subgroups of) agents are unconnected, (b) susceptibility is low or (c) the bounded confidence threshold on interactions is low.

The HIOM is similar to the De Groot model when A is small and equal for all agents. In this case $O \approx I$, as O is approximately a linear function of I (see Fig. 2b, blue line) and consensus is to be expected. If we introduce bounded confidence in the HIOM, the dynamics become fundamentally different. Depending on the threshold, a number of opinion groups may emerge. The bounded confidence model is a very active field of research [70] and many variants have been proposed.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

The standard solution for polarization in these models is to raise the threshold, such that opposing agents exchange information. This, however, may not be so easy in practice. Here we offer a new mechanism for reaching consensus in bounded confidence models. This mechanism is based on a counter-intuitive intervention that makes use of the special dynamics of the HIOM. In this intervention activists are temporarily set to the opposite (conservative) opinion while their underlying attention and information is preserved. They may stay in the opposite opinion for some time because of the hysteresis effect, but eventually they jump back. However, in the mean time they can exchange information with agents holding conservative views, because they now pass the bounded confidence threshold. As these activists' attention is generally higher their impact on these less involved agents is strong. In Simulation 3, we show that this intervention leads to a substantial opinion switch.

8. Measures

The quantification of polarization is not a simple matter [71]. A popular but conservative test of bimodality is Hartigan's dip statistic [72], which we apply for formal testing. Its values vary roughly between 0 and 0.05, where higher values indicate less unimodality. Additionally, we quantify the behaviour of the HIOM with two simple measures, the proportion of positive opinions and variance in opinion. The proportion of opinions reveals whether opinions converged to one of the extremes while the variance of opinions indicates polarization. Additionally, we compute Cardan's discriminant (CD = $27I^2 - 4A^3$), which is negative when agents reside in the bifurcation set (bistable area) of the cusp.

Our last measure concerns correlation between opinions of neighbouring agents. The assortativity coefficient [73] measures the clustering of opinions in the network. If the coefficient is high, connected vertices tend to have the same values.

9. Simulations

The HIOM algorithm can be summarized as follows.

- Choose a network topology, such as a stochastic block model
- Set model parameters N, dt (0.01), A_{min} (-0.5), A^* (1), s_O (0.01), s_I, d_A, p, r_{min}, and t_O. If values are given in parentheses these are default values in all our simulations.
- Initialize agents, set I_{init}, O_{init} and A_{init}.
- Iterate
  - Randomly choose one agent, weighted with attention A (eq. 5)
  - Randomly choose a neighbour as partner in the interaction
  - If the opinion difference between these two agents is less than t_O:
    - Add attention to both agents (eq. 6, $d_A u_i(2A^* - A_i)$)
    - Exchange information (eq. 7)
  - Apply decay in A to all agents (eq. 6, $-\frac{2d_A}{N^2}A_i$)
  - Update opinion O in all agents (eq. 4).

In each of the simulations we specify the specific parameter settings. R code for all simulations are publicly available at: https://github.com/hvdmaas/HIOM.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

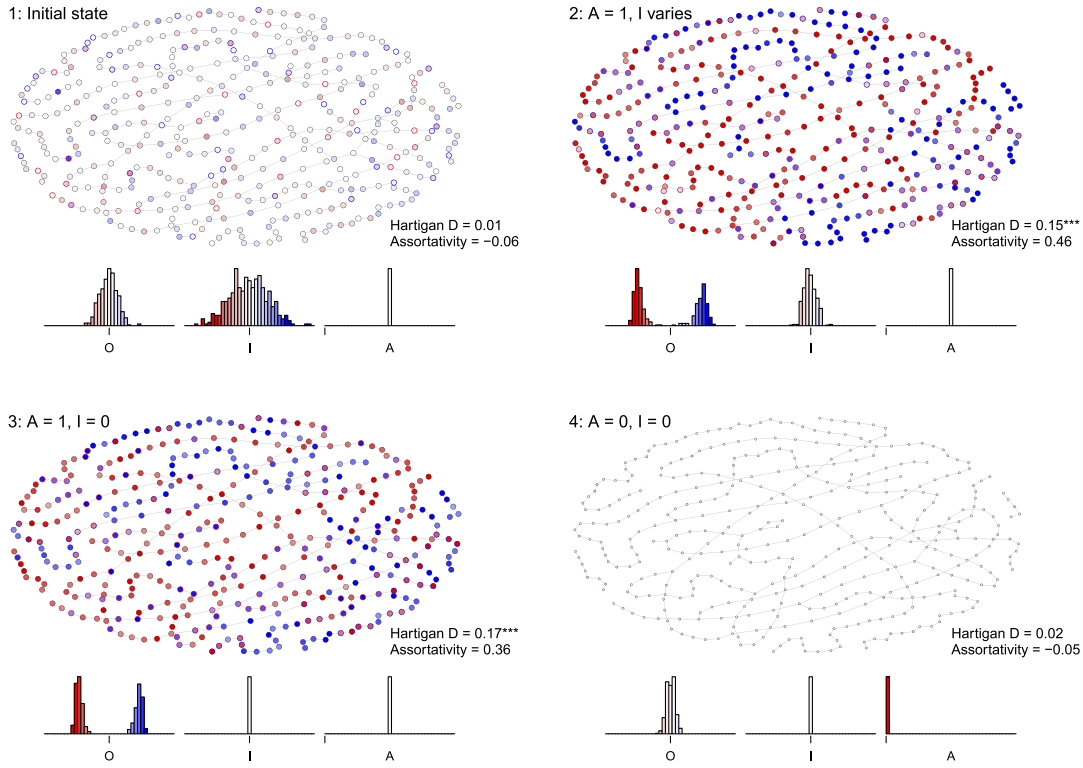

FIG. 4. Polarization in the multi-agent model. Polarization in the form of clusters emerges spontaneously and persists when information underlying opinions is equal and neutral for all agents. When attention also diminishes, opinions become neutral. Negative, neutral and positive opinion (information) are displayed on a colour scale from red to white to blue. Node borders of agents are black, except when an agent is ambivalent ($O_iI_i < 0$ and $CD_i < 0$). For ambivalent agents, border colours correspond to their informational state. Below the networks the distributions of O, I and A are displayed. Hartigan's D is given and is significant (*) when the opinion distribution is not unimodal. The assortativity measure is correlational measure of clustering of opinions.

9.1 *Polarization due to hysteresis*

To test for Axelrod polarization effect, we run the following simulation. We choose a Watts–Strogatz small-world model [52], with a rewiring probability of 0.02. We start the system with randomly selected I and O ($I_{init} = N(0, 0.3), O_{init} = N(0, 0.3)$). All agents are and stay highly attentive ($A_{init} = 1, d_A = 0$). Furthermore, we set $N = 400, dt = 0.01, p = 1, r_{min} = 0, s_I = 0$ and t_O to an arbitrary high value such that bounded confidence plays no role.

The first panel of Fig. 4 shows the initial random state, the second panel shows the polarization in clusters after 5,000 iterations. After 5,000 iterations, we let the information I shrink to 0 for all agents. Thus, all agents become equal ($A_i = A = 1, I_i = I = 0$). But instead of consensus to a neutral opinion a polarized landscape remains (Panel 3, after 10,000 iterations). This remaining polarization is due to the hysteresis effect. Agents stick to their 'old' opinion which is possible due to their high attention. After 10,000 iterations, we let attention shrink to zero too, resulting in convergence to a neutral opinion (panel 4, after 15,000 iterations). This convergence occurs because at low attention hysteresis is absent.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

9.2 *Opposition to activism: the case of black Pete*

We simulate the persuasion paradox by setting up an initial state in which almost all agents (conservatives) are moderately positive ($I_{init} = 0.1$) and lowly attentive ($A_{init} = 0$), except for two individuals. These two subjects (activists) are negative ($I_{init} = -0.5$) and highly involved and attentive ($A_{init} = 1$). Persuasion is set to $p = 2$, such that when an highly attentive activist and a lowly attentive conservative interact, the activist tends to 'win the debate' and copies his I to the conservative agent. We apply weighted agent selection, thus only highly attentive agents initiate interactions. We further set $dt = 0.01, d_A = 0.1$, $r_{min} = 0.1$, $s_I = 0.0005$ and t_O to an arbitrary high value. In this simulation, we use a stochastic block model ($N = 400$) with 10 clusters as social network (within and between connectivity probabilities of 0.001 and 0.2, respectively).

The naive expectation for this simulation is that activists quickly spread. They are the only ones initiating interactions and win all debates. They copy their informational state to any conservative they interact with. But Fig. 5 shows otherwise. Initially, activists spread quickly but they also create their own opposition, resulting in strong polarization.

Several parameters settings affect this effect. Two important parameters are the rate of change in A (d_A) and the base resistance of agents (r_{min}). We expect that for activist to be successful d_A and r_{min} should be low, such that change in information dominates change in attention. We simulated data for different combinations of d_A and r_{min} for the stochastic block model used in Fig. 5 and for a two-dimensional lattice as network topology. Figure 6 displays the outcomes for the probability of positive opinions and Hartigan's dip test. The results indicate an interaction effect. For base resistances larger 0.3, an increase in the rate of change in attention indeed lowers the success of activists. Their actions lead to polarization, as indicated by Hartigan's D statistic, or, when base resistance is very high (> 0.6) to dominance of the conservative opinion. For base resistances lower than 0.3, an increase in the rate of change in attention leads to dominance of the activist's opinion. We note that the results for both network topologies are similar, except for the spread of activists for low values of d_A and r_{min}, which is faster in the lattice model.

9.3 *A solution to polarization: the meat-eating vegetarian*

The last application of the HIOM concerns the persuasion problem in continuous opinion models. In these models, bounded confidence is a crucial cause of polarization. If opinions of agents differ above some threshold, no interaction and no exchange of information take place. The HIOM offers a new way out of this polarization dilemma.

In the HIOM, it is possible to have agents that have a mismatch between I and O (due to the hysteresis effect). An example would be a meat-eating vegetarian, a person with information consistent with the vegetarian point of view, but with non-vegetarian behaviour. Consider the following scenario. We setup a social network with 400 agents, as in the previous simulation. Agents belong to either group V (vegetarian, 20%) or group M (meat eating, 80%). The V agents are highly involved and attentive ($A_{init} = 1$) and have extreme negative (contra) information ($I_{init} = -0.4$), the M agents are less attentive ($A_{init} = 0.1$) and have low positive I ($I_{init} = 0.1$). Other parameters are set to: $p = 2, r_{min} = 0, d_A = 0.01, dt = 0.01, s_O = 0.0001$ and $s_I = 0.0$. Note the low value of d_A. This setting limits the persuasion paradox effect in this simulation. The bounded confidence threshold t_O is set to 0.2.

In this setup, due to bounded confidence, no transitions between states happen. Only interactions between very similar agents occur with no effect on the polarized state. But now, as intervention, we introduce meat-eating vegetarians (i.e. agents, who have negative I, but positive O). In each iteration,

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

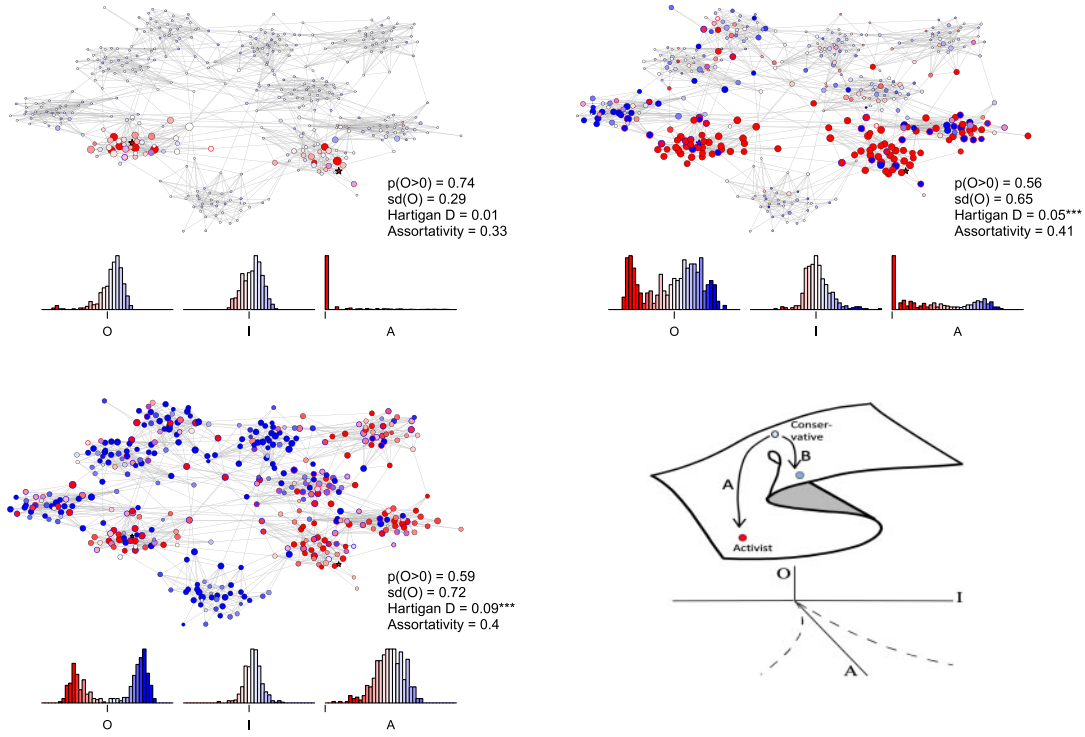

FIG. 5. The black Pete case. The stars indicate the two first activists. Top panels: Initially, activist (red) quickly persuade lowly involved conservatives (light blue). However, they not only spread their point of view (information) but also increase attention in interaction partners. Some conservatives radicalize in the opposite direction. The lower left panel shows a strongly polarized state after 25,000 iterations. Right bottom panel provides the explanation of this effect. Activists attract conservatives to their informational position because, due to their higher attention to the attitude object, they 'win' the debates. Thus, conservatives move along the information (I) axis to the activist position (path A). However, due to interactions, the attention to the attitude object in conservatives increases too. If the change in attention is too fast, some conservatives may become anti-activists, resulting in polarization. They will follow path B instead of path A.

with a small probability (0.0005), we reverse the opinion of some V agents (with $O < 0, I < 0$ and $A > 0$). This changes vegetarians into meat-eaters. Note that we only change O and not I. Due to the hysteresis effect, this new state is temporarily stable (see Fig. 2a). Agents will stay in this new state but will probably flip back after some iterations. The point of this intervention is that these transformed agents are 'trusted' by the M agents because they have similar opinions. The effect of the succeeding interaction is large because the I update is weighted by attention which is higher in the V agents. In this way, the meat-eating vegetarians change the common opinion. This is demonstrated in Fig. 7.

To study the robustness of this effect, we repeated this simulation for a combination of values of the probability of perturbations, p(perturbation) and the bounded confidence threshold t_O. Figure 8 displays the effects of these parameters on the probability of a positive opinion ($p(O > 0)$ and Hartigan's D statistic.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

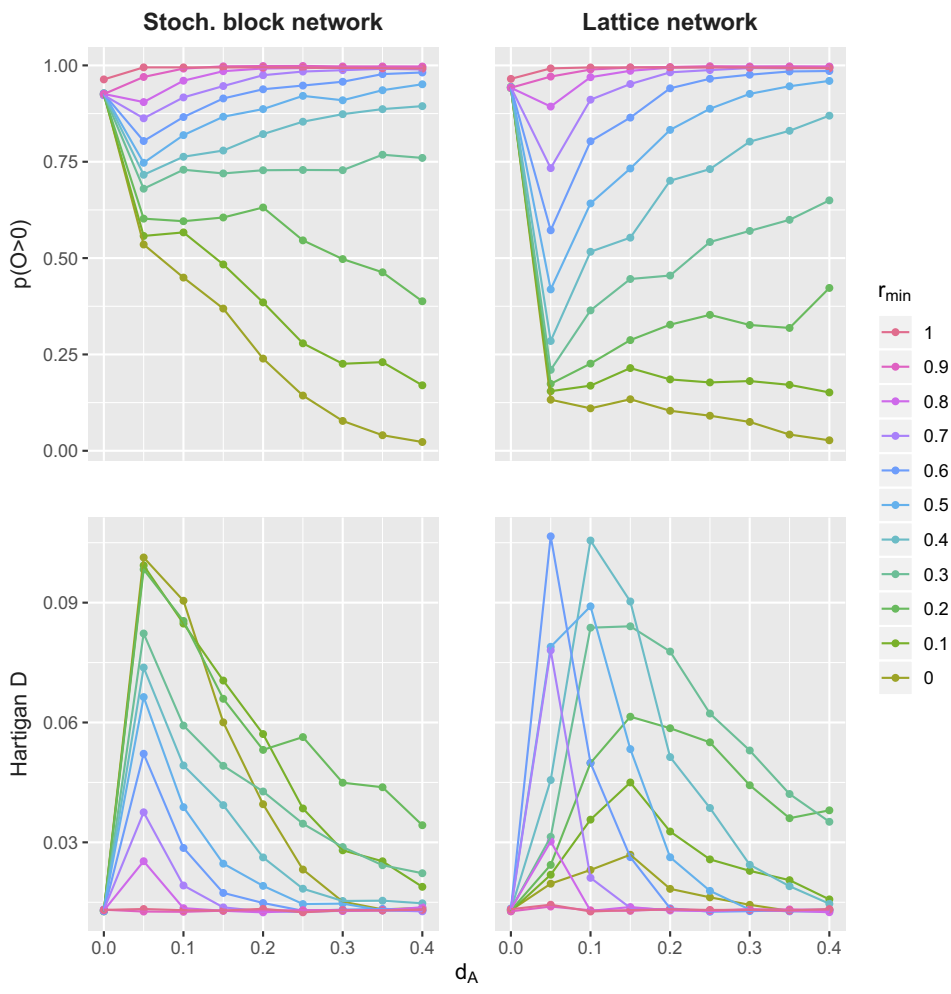

FIG. 6. Opposition to activism as function of rate of change in attention (d_A) and base resistance (r_{min}). The figure shows measures (averages based on 5 replications) of the HIOM after 25,000 iterations of the network displayed in Fig. 5. Intermediate values of d_A and base resistance results in strong polarization as indicated by proportion of positive opinions ($P(O > 0)$ and Hartigan's D statistic. The lower panels show a replication of these results for a network where agents are placed on a two-dimensional lattice.

It is remarkable that we perturb vegetarian agents in the direction of the opposite attitude (we temporarily make them meat-eaters or flexitarians), resulting in a long-term conversion to vegetarianism in the population.

The HIOM thus predicts, under the assumption of bounded confidence, that persuasion is more effective when individuals with the same opinion or behaviour but different information interact. To our best knowledge, there is no empirical research on this prediction although it relates to the dual identity effect [74]. A related prediction has been put forward in [75]. This prediction might thus be an interesting avenue for future research.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

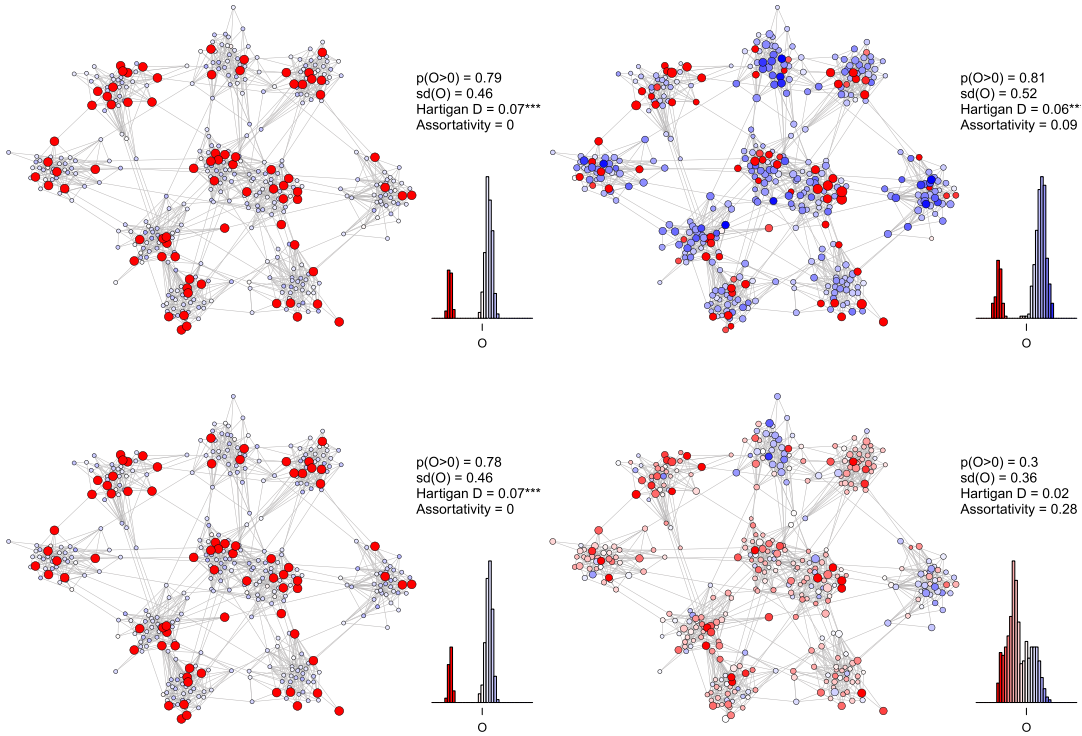

FIG. 7. The meat-eating vegetarian. The left panels show two equal initial states. In the top panels bounded confidence prevents effective interactions between two groups (vegetarians = red; meat eaters = blue). The right top panel shows a polarized state after 30,000 iterations. However, in the lower panels occasionally vegetarians are perturbed in the direction of meat eating. These perturbed agents have $I < 0$ and $O > 0$, which is possible since $A > 0$ (hysteresis). Because $O > 0$ these agents can exchange information with meat eaters, which leads to slow convergence to the V state in the population (lower right panel).

10. Discussion

We propose the hierarchical opinion model (HIOM). This formal approach unites the rich empirical database of studies in psychology and sociology and the rapidly developing field of formal modelling of social interaction in computer science and statistical physics.

Its unique character is due to how the agents are modelled. The agent's attitude or opinion is conceptualized as a network of feelings, thoughts and behaviours towards an object or issue. We made some simplifying assumption such that the attitude network resembles the well-known Ising model. Using the mean field approximation of the Ising network, we derived a stochastic cusp description of the individual agent. In this cusp model, opinion is based on attention and information. Highly attentive agents are polarized. Changes in information may lead to sudden jumps and hysteresis. In lowly attentive agents change in opinion is continuous. Their opinion can be easily influenced but their attitudes are highly unstable.

Though still simplistic, this agent model is more advanced than the agent models used in most statistical physics models of social dynamics. Such models are generally divided in two broad classes,

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

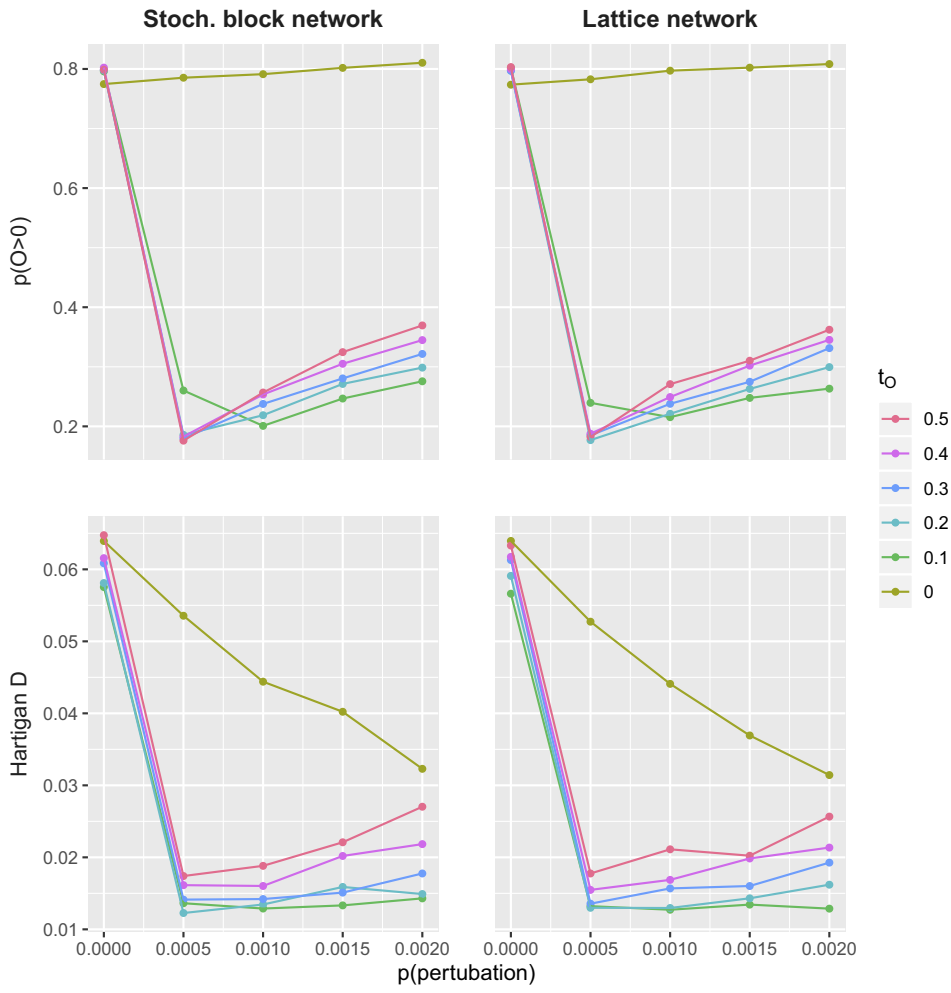

FIG. 8. The diminishing of polarization by perturbations as function of probability of perturbations, p(perturbation), and the bounded confidence threshold t_O. The figure shows measures (averages based on 5 replications) of the HIOM after 20,000 iterations of the network displayed in Fig. 7. For $t_O = 0$ and p(perturbation) = 0 the initial polarized state persists. For other values polarization diminishes (decrease in Hartigan's D) and the V agents take over (as indicated by the decrease in $p(O > 0)$. A low value of p(perturbation) is optimal for this effect.

based on whether opinions of agents are discrete or continuous. As the HIOM included both classes, it integrates these two main branches of sociophysics. Additionally, because agent's behaviour displays hysteresis, history, the path to the current state, could come into play.

We have shown that the HIOM reproduces both polarization within and between individuals and, additionally, derived some novel predictions. The main prediction derived from the HIOM is the persuasion paradox. The attempt to persuade other agents, even in ideal circumstances, may fail if the attention in other agents increases such that these agents polarize in the opposite direction. Secondly, we introduced a new way to escape polarization in bounded confidence models, based on the hysteresis effect in agents.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

The HIOM is based on some reasonable assumptions on the effects of interaction on attention and information of agents. We assume (a) that higher attentive (higher involved) agents initiate interactions more often, (b) that information sharing is weighted by attention and (c) that attention to the attitude object increases during interaction but otherwise decays.

We note that the three assumptions on the interactions between agents are sufficient to model these phenomena, but perhaps not necessary. Modifications of the assumptions on information sharing are possible [76]. With regard to dynamics of attention, we note that not much is known about the way attention or involvement spread in populations. More empirical research on the tenability of the assumptions on interaction in the HIOM is required and possible [77].

On the other hand, for our main effects hysteresis in agents is necessary. For instance, in a cusp model without hysteresis, by applying the so-called Maxwell convention [78], in which systems seek the state that globally minimizes the potential, agents immediately recover from perturbations. In such a case, perturbations have no effect at all and there is no escape from polarization.

We clearly did not explore all the new possibilities of the HIOM. In the field of social dynamics, many other interesting model ideas have been introduced which could be incorporated in the HIOM. Examples are learning in social networks of opinions [79], the role of mass media [80], the role of reputation [81] and the effect of leaders on group opinions [82].

Also, many other empirical phenomena studied in social psychology could be investigated in the HIOM. Social psychologists distinguish between different types of involvement, for instance, claiming different effects on persuasion [83]. We already discussed many links with the empirical literature, but further integration with approaches in social science research is required. The HIOM emphasizes the importance of attention and involvement, in addition to information, in opinion change. It suggests that persuasion requires a delicate, intermediate, level of involvement. If involvement is too high, attitudes are polarized due to hysteresis and very hard to influence with new information. If involvement is too low, attitudes are easily influenced but highly unstable. The lowly involved agent goes where the winds blows.

Finally, we also expect that more analytical work on the HIOM is possible. Further analysis may, for instance, require simplification of the weighted agent selection. This idea has ecological validity in our view, but it considerably complicates the dynamics of the HIOM.

Given that the HIOM unifies the two broad classes of opinion spread models, integrates individual-level and groups-level polarization, and provides several novel predictions, it is our view that the HIOM represents a significant advancement in the opinion spread literature. We hope that our work will inspire both (a) empirical research on the HIOM to further integrate the psychology and sociology of opinion spread and (b) further analytical work on the HIOM to advance this promising model. We expect that such work will illuminate several interesting properties at the intersection of psychology, sociology and statistical physics.

References

1. Abramowitz, A. I. & Saunders, K. L. (2008) Is polarization a myth? *J. Polit.*, **70**, 542–555.
2. Iyengar, S., Lelkes, Y., Levendusky, M., et al. (2019) The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.*, **22**, 129–146.
3. McCright, A. M., Xiao, C. & Dunlap, R. E. (2014) Political polarization on support for government spending on environmental protection in the USA, 1974-2012. *Soc. Sci. Res.*, **48**, 251–260.
4. Gill, P., Horgan, J. & Deckert, P. (2014) Bombing alone: tracing the motivations and antecedent behaviors of lone-actor terrorists. *J. Forensic Sci.*, **59**, 425–435.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

5. Castellano, C., Fortunato, S. & Loreto, V. (2009) Statistical physics of social dynamics. *Rev. Mod. Phys.*, **81**, 591–646.
6. Helbing, D. (2010) *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes*, 2nd edn. Berlin Heidelberg: Springer.
7. Galam, S. (2008) Sociophysics: a review of galam models. *Int. J. Mod. Phys. C*, **19**, 409–440.
8. Axelrod, R. (1997) The dissemination of culture: a model with local convergence and global polarization. *J. Confl. Resolut.*, **41**, 203–226.
9. Katz, D. (1960) The functional approach to the study of attitudes. *Public Opin. Q.*, **24**, 163–204.
10. Fazio, R. H. (2007) Attitudes as object-evaluation associations of varying strength. *Soc. Cogn.*, **25**, 603–637.
11. Fazio, R. H. (1995) Attitudes as object-evaluation associations: determinants, consequences, and correlates of attitude accessibility. *Attitude Strength: Antecedents and Consequences* (R. E. Petty & J. A. Krosnick eds). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc., pp. 247–282.
12. Zanna, M. P. & Rempel, J. K. (1988) Attitudes: a new look at an old concept. *The Social Psychology of Knowledge* (D. Bar-Tal & A. W. Kruglanski eds). Paris, France: Editions de la Maison des Sciences de l'Homme, pp. 315–334.
13. Krosnick, J. A. & Petty, R. E. (1995) Attitude strength: an overview. *Attitude Strength: Antecedents and Consequences* (R. E. Petty & J. A. Krosnick eds). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc., pp. 1–24.
14. Ajzen, I. (1991) The theory of planned behavior. *Organ Behav. Hum. Decis. Process.*, **50**, 179–211.
15. Chaiken, S., Liberman, A. & Eagly, A. H. (1989) Heuristic and systematic information processing within and beyond the persuasion context. *Unintended Thought*. New York, NY, USA: Guilford Press, pp. 212–252.
16. Petty, R. E. & Cacioppo, J. T. (1986) The elaboration likelihood model of persuasion. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (R. E. Petty & J. T. Cacioppo eds). New York, NY, USA: Springer, pp. 1–24.
17. Festinger, L. (1962) A theory of cognitive dissonance. Stanford, CA: Stanford University Press.
18. Dalege, J., Borsboom, D., van Harreveld, F., et al. (2016) Toward a formalized account of attitudes: the Causal Attitude Network (CAN) model. *Psychol. Rev.*, **123**, 2.
19. Dalege, J., Borsboom, D., van Harreveld, F. & van der Maas, H. L. (2017) Network analysis on attitudes: a brief tutorial. *Soc. Psychol. Personal. Sci.*, **8**, 528–537.
20. Dalege, J., Borsboom, D., van Harreveld, F., et al. (2017) Network structure explains the impact of attitudes on voting decisions. *Sci. Rep.*, **7**, 4909.
21. Dalege, J., Borsboom, D., van Harreveld, F. & van der Maas, H. L. J. (2018) The attitudinal entropy (AE) Framework as a general theory of individual attitudes. *Psychol. Inq.*, **29**, 175–193.
22. Wainwright, M. J. & Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Found. Trends® Mach. Learn.,* **1**, 1–305.
23. Graves, A., Mohamed, A. & Hinton, G. (2013) Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (R. Ward & L. Deng eds). Piscataway, NY: IEEE, pp. 6645–6649.
24. Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Methodol.*, **36**, 192–236.
25. Haig, B. D. (2005) An abductive theory of scientific method. *Psychol. Methods*, **10**, 371–388.
26. Potts, R. B. (1952) Some generalized order-disorder transformations. *Math. Proc. Camb. Philos. Soc*., **48**, 106–109.
27. Marsman, M., Borsboom, D., Kruis, J., et al. (2018) An introduction to network psychometrics: relating Ising network models to item response theory models. *Multivar. Behav. Res.*, **53**, 15–35.
28. Frigg, R. & Werndl, C. (2017) Equilibrium in Boltzmannian statistical mechanics. *Proceedings of the EPSA15 Conference* (M. Massimi & J. W. Romeijn eds). Berlin and New York: Springer, p. 19.
29. Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.*, **11**, 127–138.
30. Zaichkowsky, J. L. (1986) Conceptualizing involvement. *J. Advert.*, **15**, 4–34.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

31. MARTINELLI, F. (1999) Lectures on glauber dynamics for discrete spin models. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probailités de Saint-Flour XXVII – 1997* (J. Bertoin, F. Martinelli, Y. Peres & P. Bernard eds). Berlin, Heidelberg: Springer, pp. 93–191.

32. ABE, Y., ISHIDA, M., NOZAWA, E., et al. (2017) Cusp singularity in mean field Ising model. *Eur. J. Phys.*, **38**, 065102.

33. STANLEY, H. E. (1987) *Introduction to Phase Transitions and Critical Phenomena*. Oxford: Oxford University Press.

34. Friedli, S. & Velenik, Y. (2017) *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge: Cambridge University Press.

35. POSTON, T. & STEWART, I. (1978) *Catastrophe Theory and Its Applications*. San Francisco: Pitman.

36. TESSER, A. (1978) Self-generated attitude. *Advances in Experimental Social Psychology* (L. Berkowitz ed.). New York: Academic Press, pp. 289–338.

37. AHLUWALIA, R. (2000) Examination of psychological processes underlying resistance to persuasion. *J. Consum. Res.*, **27**, 217–232.

38. EAGLY, A. H. & CHAIKEN, S. (1993) *The Psychology of Attitudes*. Orlando, FL, USA: Harcourt Brace Jovanovich College Publishers.

39. ZARTMAN, I. W. & FAURE, G. O. (2005) The dynamics of escalation and negotiation. *Escalation and Negotiation in International Conflicts* (I. W. Zartman & G. O. Faure eds). Cambridge: Cambridge University Press.

40. LATANÉ, B. & NOWAK, A. (1994) Attitudes as catastrophes: from dimensions to categories with increasing involvement. *Dynamical Systems in Social Psychology* (R. R. Vallacher & A. Nowak eds). San Diego, CA, USA: Academic Press, pp. 219–249.

41. VAN DER MAAS, H. L. J., KOLSTEIN R. & VAN DER PLIGT, J. (2003) Sudden transitions in attitudes. *Sociol. Methods Res.*, **32**, 125–152.

42. HU, H. (2017) Competing opinion diffusion on social networks. *R. Soc. Open Sci.*, **4**, 171160.

43. CHEON, T. & GALAM, S. (2018) Dynamical Galam model. *Phys. Lett. A*, **382**, 1509–1515.

44. NOWAK, A., SZAMREJ, J. & LATANÉ, B. (1990) From private attitude to public opinion: a dynamic theory of social impact. *Psychol. Rev*., **97**, 362–376.

45. LORENZ, J. (2007) Continuous opinion dynamics under bounded confidence: a survey. *Int. J. Mod. Phys. C*, **18**, 1819–1838.

46. MARTINS, A. C. R. (2008) Continuous opinions and discrete actions in opinion dynamics problems. *Int. J. Mod. Phys. C*, **19**, 617–624.

47. SOBKOWICZ, P. (2012) Discrete model of opinion changes using knowledge and emotions as control variables. *PLoS One*, **7**, e44489.

48. MASUDA, N. (2014) Voter model on the two-clique graph. *Phys. Rev. E*, **90**, 012802.

49. BOCCALETTI, S., BIANCONI, G., CRIADO, R., et al. (2014) The structure and dynamics of multilayer networks. *Phys. Rep.*, **544**, 1–122.

50. COBB, L. & ZACKS, S. (1985) Applications of catastrophe theory for statistical modeling in the biosciences. *J. Am. Stat. Assoc.*, **80**, 793–802.

51. WAGENMAKERS, E.-J., MOLENAAR, P. C. M., GRASMAN, R. P. P. P., et al. (2005) Transformation invariant stochastic catastrophe theory. *Phys. Nonlinear Phenom.,* **211**, 263–276.

52. WATTS, D. J. & STROGATZ, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440.

53. KARRER, B. & NEWMAN, M. E. J. (2011) Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, **83**, 016107.

54. NOWICKI, K. & SNIJDERS, T. A. B. (2001) Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, **96**, 1077–1087.

55. NAKAMARU, M. (2006) Lattice models in ecology and social sciences. *Ecol. Res.*, **21**, 364–369.

56. BALDASSARRI, D. & BEARMAN, P. (2007) Dynamics of political polarization. *Am. Sociol. Rev.*, **72**, 784–811.

57. REED, C. (1998) Dialogue frames in agent communication. *Proceedings of the 3rd International Conference on Multi Agent Systems* (Y. Demazeau ed.). Washington, DC, USA: IEEE Computer Society, p. 246.

58. RAMOS, M., SHAO, J., REIS, S. D. S., et al. (2015) How does public opinion become extreme? *Sci. Rep.,* **5**, 10032.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

59. Hilhorst, S. & Hermes, J. (2016) 'We have given up so much': passion and denial in the Dutch Zwarte Piet (Black Pete) controversy. *Eur. J. Cult. Stud.,* **19**.

60. Hovland, C. I., Janis, I. L. & Kelley, H. H. (1953) *Communication and Persuasion: Psychological Studies of Opinion Change*. New Haven, CT, USA: Yale University Press.

61. Nyhan, B. & Reifler, J. (2010) When corrections fail: the persistence of political misperceptions. *Polit. Behav.*, **32**, 303–330.

62. Brehm, S. & Brehm, J. W. (1981) *Psychological Reactance: A Theory of Freedom and Control.* New York: Academic Press.

63. Rosenberg, B. D. & Siegel, J. T. (2018) A 50-year review of psychological reactance theory: do not read this article. *Motiv. Sci.*, **4**, 281–300.

64. Trevors, G. J., Muis, K. R., Pekrun, R., et al (2016) Identity and epistemic emotions during knowledge revision: a potential account for the backfire effect. *Discourse Process.*, **53**, 339–370.

65. Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. (2000) Mixing beliefs among interacting agents. *Adv. Complex Syst.*, **03**, 87–98.

66. Hegselmann, R. & Krause, U. (2002) Opinion dynamics and bounded confidence: models, analysis and simulation. *J. Artif. Soc. Soc. Simul.*, **5**, 1–33.

67. Friedkin, N. E. & Johnsen, E. C. (1990) Social influence and opinions. *J. Math. Sociol.*, **15**, 193–206.

68. Sherif, M. & Hovland, C. I. (1961) *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*. Oxford, England: Yale University Press.

69. Siero, F. W. & Doosje, B. J. (1993) Attitude change following persuasive communication: integrating social judgment theory and the elaboration likelihood model. *Eur. J. Soc. Psychol.*, **23**, 541–554.

70. Flache, A., Mäs, M., Feliciani, T., et al (2017) Models of social influence: towards the next frontiers. *J. Artif. Soc. Simul.*, **20**, 2.

71. Pfister, R., Schwarz, K. A., Janczyk, M., et al. (2013) Good things peak in pairs: a note on the bimodality coefficient. *Front. Psychol.*, **4**, 700.

72. Hartigan, J. A. & Hartigan, P. M. (1985) The dip test of unimodality. *Ann. Stat.*, **13**, 70–84.

73. Newman, M. E. J. (2003) Mixing patterns in networks. *Phys. Rev. E*, **67**, 026126.

74. González, R. & Brown, R. (2006) Dual identities in intergroup contact: group status and size moderate the generalization of positive attitude change. *J. Exp. Soc. Psychol.*, **42**, 753–767.

75. Hegselmann, R. & Krause, U. (2015) Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: a simple unifying model. *Netw. Heterog. Media*, **10**, 477.

76. Hegselmann, R. & Krause, U. (2005) Opinion dynamics driven by various ways of averaging. *Comput. Econ.*, **25**, 381–405.

77. Takács, K., Flache, A. & Mäs, M. (2016) Discrepancy and disliking do not induce negative opinion shifts. *PLoS One*, **11**, e0157948.

78. Gilmore, R. (1981) *Catastrophe Theory for Scientists and Engineers.* New York: Wiley.

79. Acemoglu, D. & Ozdaglar, A. (2011) Opinion dynamics and learning in social networks. *Dyn. Games Appl.*, **1**, 3–49.

80. Mckeown, G. & Sheehy, N (2006) Mass media and polarisation processes in the bounded confidence model of opinion dynamics. *J. of Arti. Soc. and Soc. Sim.*, 9. http://jasss.soc.surrey.ac.uk/9/1/11.html. Accessed 23 December 2019.

81. Gross, J. & Dreu, C. K. W. D. (2019) The rise and fall of cooperation through reputation and group polarization. *Nat. Commun.*, **10**, 1–10.

82. Zhao, Y., Kou, G., Peng, Y. & Chen, Y. (2018) Understanding influence power of opinion leaders in e-commerce networks: an opinion dynamics theory perspective. *Inf. Sci.,* **426**, 131–147.

83. Johnson, B.T. & Eagly, A. H. (1989) Effects of involvement on persuasion: a meta-analysis. *Psychol. Bull.,* **106**, 290–314.

84. Lipowski, A., Ferreira, A. L., Lipowska, D. & Gontarek, K. (2015) Phase transitions in Ising models on directed networks. *Phys. Rev. E*, **92**, 052811.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

85. Jin, S., Sen, A., Guo, W. & Sandvik, A. W. (2013) Phase transitions in the frustrated Ising model on the square lattice. *Phys. Rev. B*, **87**, 144406.
86. Bianconi, G. (2002) Mean field solution of the Ising model on a Barabási-Albert network. *Phys. Lett. A*, **303**, 166–168.
87. Schriesheim, C. A. & Hill, K. D. (1981) Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. *Educ. Psychol. Meas.*, **41**, 1101–1114.
88. Waldorp, L. & Marsman. M. (2019) Intervention in undirected Ising graphs and the partition function. ArXiv190511502 Stat.[AQ]
89. Dalege, J., van Harreveld, F., Borsboom, D. & van der Maas, H. L. J. (2019) The Learning Ising Model of Attitude (LIMA): entropy reduction by Hebbian learning. Manuscript Preparation.
90. Rojas, R. (1996) *Neural Networks: A Systematic Introduction.* Berlin Heidelberg: Springer.
91. Monroe, B. M. & Read, S. J. (2008) A general connectionist model of attitude structure and change: the ACS (Attitudes as Constraint Satisfaction) model. *Psychol. Rev.,* **115**, 733–759.
92. Van Overwalle, F. & Jordens, K. (2002) An adaptive connectionist model of cognitive dissonance. *Personal. Soc. Psychol. Rev.*, **6**, 204–231.

Appendix A

As we use the cusp as a mean field approximation for the Ising attitude network of individuals, it is important to know under which conditions this approximation holds. It is known to hold for the Curie Weiss (CW) model exactly [34], where all nodes are connected to all other nodes. In the Ising attitude model, several deviations of the Curie Weiss are plausible. First, we do not expect a fully connected network, a certain percentage of connections will be absent. Second, connection weights will probably not all have the same value. Third, connections may be asymmetrical. Each of these deviations could have a large impact on the dynamics [32, 33, 84–86].

However, two properties of the Ising attitude model are of importance. The first is that the connections may not be all equal, they will be mostly positive (after rescaling). That is, we conjecture that it is generally possible to define all relevant nodes (for instance, regarding the consumption of meat) such that all positive values represent a pro attitude and all negative values represent a contra attitude. This is standard practice in the analysis of attitude questionnaires [87]. We can show that if the variation in connectivity strength is sub-Gaussian, then the normalizing constant (and hence the probability) is guaranteed to be close to the true one with exponential rate as a function of the size of the graph [88].

Secondly, we have no reason to believe that attitude networks are extremely sparsely connected. There will be many positive connections between attitude nodes. In [89], we model the development of Ising attitude networks using Hebb rule, which says that what fires together wires together. As soon as nodes display congruent behaviour over time, the Hebb rule will increase the values of the interaction parameters, leading to positive connections. Connections decay (depending on d_ω), but this is a slow process. An Ising model with Hebbian learning is known as the Boltzmann machine or stochastic Hopfield neural network [90]. Several related modelling approaches of attitudes based on connectionist networks have been proposed [91, 92]. Under these two conditions, the mean field approximation for the Curie Weiss probably holds for the Ising attitude model.

In Fig. A.1, we report a simulation study in which we investigate these cases. Each row in this figure represents a test for the presence of the pitchfork bifurcation (by increasing A, keeping I at zero) and the hysteresis effect (by subsequently increasing and decreasing I, keeping A at 2), which are typical for the cusp catastrophe.

The first row shows the CW case. The network ($n = 40$) is fully connected with all connection weights equal to 0.2. As expected, both the pitchfork bifurcation and hysteresis emerge. The other rows represent various deviations from the CW. In the second row connections weights are not equal but sampled from

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021

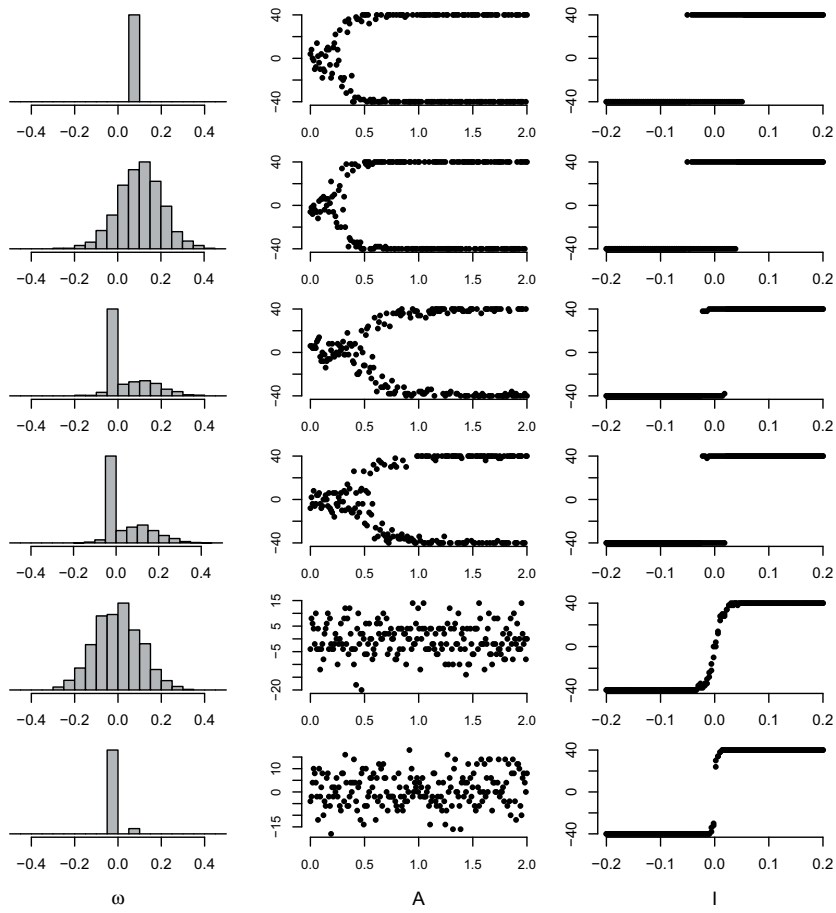

FIG. A.1. The robustness of the cusp approximation of the Ising attitude model for various deviations the basic Curie Weis model (first row). The first column shows the distributions of the interactions ω for the five cases described in the text. The second column displays the effect of increasing A (attention) on opinion ($O = \Sigma x$), which should give a pitchfork bifurcation. The third column shows hysteresis as a function of successive increasing and decreasing in information I.

N(0.1,0.1), implying a substantial amount of variations in connections weights, some being negative. In the third row, additionally, 50% missing links are introduced. In fourth row, additionally, we allow for asymmetric connections. This case combines three severe deviations of the CW. In all these three cases the cusp approximation seems to hold.

The last two rows show cases that do not work. The fifth row is equal to the second row but connections weights are sampled from N(0,0.1), such that they are not mostly positive. In the six row, the connections are too sparse (95% missing links). In both cases, the pitchfork and hysteresis do not appear. Note, however that these two cases are not expected under the Ising attitude model due to Hebbian learning. In the ideal case, Ising attitude networks converge, due to learning, to the CW case.

This simulation supports our claim that the dynamics of the Ising attitude model can be approximated with the cusp catastrophe if the interactions are not too sparse and are mostly positive. If so, connections may be asymmetric and vary in value.

Downloaded from https://academic.oup.com/comnet/article/8/2/cnaa010/5823576 by University of Amsterdam user on 08 January 2021