Classification-Based Opinion Formation Model Embedding Agents' Psychological Traits

Devia Pinzon, C.A.; Giordano, G.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Classification-Based Opinion Formation Model Embedding Agents' Psychological Traits

Carlos Andrés Devia[1] and Giulia Giordano[1, 2]

[1]*Delft Center for Systems and Control, Delft University of Technology, Mekelweg, 2 - 2628 CD Delft (Zuid-Holland), The Netherlands*
[2]*Department of Industrial Engineering, University of Trento, Via Sommarive, 9 - 38123 Povo (Trento), Italy*
Correspondence should be addressed to *C.A.DeviaPinzon@tudelft.nl*

Abstract: We propose an agent-based opinion formation model characterised by a two-fold novelty. First, we realistically assume that each agent cannot measure the opinion of its neighbours about a given statement with infinite resolution and accuracy, and hence it can only perceive the opinion of others as agreeing *much more*, or *more*, or *comparably*, or *less*, or *much less* (than itself) with that given statement. This leads to a classification-based rule for opinion update. Second, we consider three complementary agent traits suggested by significant sociological and psychological research: *conformism*, *radicalism* and *stubbornness*. We rely on World Values Survey data to show that the proposed model has the potential to predict the evolution of opinions in real life: the classification-based approach and complementary agent traits produce rich collective behaviours, such as polarisation, consensus, and clustering, which can yield predicted opinions similar to survey results.

Keywords: Agent-Based Social Simulation, Agent-Based Model, Opinion Formation, Opinion Dynamics, Real Data Validation

● Introduction

1.1 The development and analysis of opinion formation models has been an active field of research since the introduction of the first opinion formation models (French Jr. 1956; Harary 1959; Harary et al. 1965; DeGroot 1974). Increasingly more sophisticated models have been developed by embedding different concepts such as *susceptibility* (Friedkin 1986; Friedkin & Johnsen 1999), *stubbornness* (Hegselmann & Krause 2015; Masuda 2015), *leaders* (Kacperski & Holyst 1999, 2000), *emotions* (Sobkowicz & Sobkowicz 2010; Chmiel et al. 2011), *trust* (Yin et al. 2019; Krawczyk et al. 2010), *bounded confidence* (Hegselmann & Krause 2002), *coevolving networks* (Su et al. 2014; Sobkowicz 2009), *biases* (Sobkowicz 2018; Dandekar et al. 2013; Banisch & Shamon 2021), *polarity* (Lorenz et al. 2021), *assimilation* (Fu & Zhang 2016; Lorenz et al. 2021), *tolerance* (Duggins 2017), *mass media* (Chattoe-Brown 2014), *controversy* (Baumann et al. 2020), *weighted balance theory* (Schweighofer et al. 2020), among others. Although there may be different reasons to construct mathematical models of opinion formation (Epstein 2008), the ultimate goal is typically to capture the mechanisms behind opinion change in society and accurately predict the evolution of real-life opinions (Thompson & Derr 2009; Troitzsch 2009).

1.2 Agent-based models (ABMs), such as the French-DeGroot model (DeGroot 1974), are very common in the opinion formation literature. In an ABM, every individual holds a different opinion (or vector of opinions) and interacts with the other agents according to a given function over a network that can be directed, weighted, or signed. Some notable examples of agent-based models are those by Hegselmann & Krause (2006), Salzarulo (2006), and Deffuant et al. (2002), among many others (Urbig et al. 2008; Afshar & Asadpour 2010; Deffuant 2006; Mckeown & Sheehy 2006; Urbig et al. 2003). An extensive literature (Mastroeni et al. 2019) proposes and analyses opinion formation models for different types of agent interactions and network characteristics.

1.3 This paper proposes an agent-based model characterised by two novel features.

1. **Finite resolution communication:** Even if the agents communicate, and openly express their real opinion, it is impossible for an agent to exactly measure and quantify the opinion of another. To account for this, the model introduces a classification-based approach, supported by the empirical finding that the assessment of the opinion of others depends on the perceived distance to those others (Schweighofer et al. 2020): each agent classifies its neighbours in different groups according to their *perceived* opinion, distinguishing between those that agree *much more*, or *more*, or *comparably*, or *less*, or *much less* (than itself) with a given statement.

   The fact that agents don't have access to the exact opinion of their neighbours with infinite resolution and accuracy has been taken into account by models with quantised opinions (Guo & Dimarogonas 2013; Ceragioli & Frasca 2018), while threshold models (Granovetter 1978; Granovetter & Soong 1986) could be seen as adopting a classification approach because the opinion update law depends on the number of neighbours expressing a particular opinion or action. Our classification-based approach is based not on the opinion of an agent's neighbours, but on the weighted difference between the opinion of the agent and of its neighbours, accounting for the finite resolution with which agents perceive the opinions of their neighbours.

   Also in opinion formation models with private and public opinions (Ye et al. 2019; Anderson & Ye 2019; Shang 2021; Duggins 2017; Banisch & Olbrich 2019) the agents cannot have perfect access to the real opinion of their neighbours. However, there is a critical difference. In these models, the agents can choose which public opinion they show, with certainty that it will be the opinion perceived by others, and hide their true private opinion: the misperception is intentional. Also in the Continuous Opinions and Discrete Actions model (Martins 2008) the mismatch between real and perceived opinions is intentional and due to the agents purposefully hiding their actual opinion to others (each agent controls the action it takes and consequently how its opinion is perceived by its neighbours).

   Conversely, in our model, the misperception is unintentional and unavoidably caused by the impossibility to communicate with infinite accuracy in view of the rich communication process that is characteristic of humans, including different interpretation of words, subtle clues, cultural aspects, social frames, and additional factors like non-verbal communication. In the proposed model, the agents wish to show as openly as possible their opinion, which still cannot be perceived with infinite resolution, and the other agents can only perceive the range in which the opinion falls, which depends on both the agent that expresses the opinion and the one that assesses it. Therefore an agent cannot know with certainty how its opinion is perceived by others. The misperception of communicated opinions is a consequence of the subjective nature of opinions and of the interpretation of verbal and non-verbal communication. In fact, the problem of measuring opinions and attitudes is so complex and nuanced that it is the main object of study of psychometrics (Coaley 2014), a whole field of study within psychology.

   To reflect imperfect opinion perception in the model, our proposed solution of classifying the opinion of others in one of five categories is inspired by the field of psychometrics: in questionnaires, the responses quantify opinions according to discrete scales. Likert scales are a standard psychometric scale used to conduct surveys, which in turn are the typical approach to measure the opinions of individuals in a population. In our model, the process of agent i assessing the opinion of agent j yields, at each time step, the answer to a five-point Likert question, which asks how much agent j agrees with a statement, compared with agent i, where the possible answers are: *Agrees much more*, *Agrees more*, *Agrees the same*, *Agrees less*, and *Agrees much less*. For certain specific questions and specific social groups and connections, the perception may be sharper, while in other cases it may be less sharp; also, some agents may have a sharper perception than others. Five levels are chosen as a compromise resolution to account for the perception skills of the *average agent* interacting with an *average neighbour*. Still, the model could be modified to consider more than five levels, thus accounting for agents with a sharper average perception, and differentiating the sharpness of perception for different agents could also be interesting; however, this goes beyond the scope of the manuscript and is left for future work.

2. **Complementary agent traits:** Each agent behaves according to a combination of three *internal traits* based on well studied sociological and psychological concepts: conformisms, radicalism, and stubbornness.

   - Conformism: agents tend to agree with their neighbours. This behaviour was first shown in the conformity experiments by Asch (1961), Asch (1955), Asch (1956) and evolved into social conformity theory (Larsen 1974). A similar behaviour is supported by the cognitive dissonance theory (Festinger 1957; Matz & Wood 2005).

- Radicalism: agents do not care if their opinion is different from their neighbours'. On the contrary, their opinion is strengthened by the presence of agents with a similar opinion, which reinforce their beliefs; this is known as the persuasive argument theory, which supports several polarisation models (Mäs & Flache 2013; La Rocca et al. 2014; Liu et al. 2015; Fu & Zhang 2016; Pinasco et al. 2017).

- Stubbornness: agents refuse to change their opinion; this type of behaviour has been often present in opinion formation models starting from those by Friedkin & Johnsen (1999), Friedkin & Johnsen (2011).

In the model, the behaviour of each agent is determined by a convex *combination* of these three traits: in reality people are not completely conformist, radical, or stubborn, but everyone is characterised by a peculiar blend of these three traits. By allowing each individual behaviour to be an outcome of a particular mix of traits, rather than a 'fixed type', the model generates a continuum of distinct agent types, each with its peculiar psychological and sociological profile, in an effort to mimic the complexity of different personalities in real life. For instance, an agent that is 50% conformist and 50% radical can be thought of as a *persuader*: thanks to its radical traits, the agent will tend to move to an extreme opinion (which a completely conformist agent would never do), but at the same time, thanks to its conformist traits, the agent will take into account also neighbours that think differently (which a completely radical agent would never do), thus capturing the complex nuances of real behaviours. Hence, the model can produce richer collective dynamics and have more flexibility, without increasing the complexity by adding more agent types. The inclusion of the radical trait can be seen as an extension of the model by Friedkin & Johnsen (1999), Friedkin & Johnsen (2011), which includes both conformist and stubborn traits.

1.4 The proposed model evolves over an invariant, directed, signed and unweighted network. Signed edges are interpreted as in structural balance theory: an edge from agent j to agent i is positive if agent i approves, trusts, or follows agent j, whereas it is negative if agent i disapproves, distrusts, or antagonises agent j (Altafini 2013; Xia et al. 2016; Cartwright & Harary 1956).

1.5 Despite significant research efforts in developing and analysing opinion formation models, empirical validation is often lacking, and has been identified as one of the frontiers of opinion modelling (Flache et al. 2017). In most cases, just an analytical or numerical characterisation of possible opinion evolutions is provided and, with some exceptions (most notably the model by Friedkin & Johnsen 1999, 2011), there are no systematic comparisons with real world behaviours. The problem of identifying individual-level parameters (in our case, agent inner traits) from population-level data (in our case, survey results) is known as the *inverse problem* (Kandler & Powell 2018) and arises, in the context of opinion dynamics, for any agent-based model, also when estimating agent interactions (Lu et al. 2021) and underlying networks (Hassanibesheli & Donner 2019) from data. An approach to solve the inverse problem using survey results relies on evolutionary algorithms (Duggins 2017); other papers taking into account survey results or empirical data in the study of opinion formation models include those by Banisch & Shamon (2021), Chattoe-Brown (2014), Baumann et al. (2020), Martins (2008).

1.6 Here, we assess the potential of our model to recreate opinion evolution in real-life settings using data from the World Values Survey (Haerpfer et al. 2010, 2015), a global research project that studies people's values and beliefs over time, conducting surveys every five years. The results of these surveys are classified by 'waves'. We use the results from wave 5 (years 2005 to 2009) and wave 6 (years 2010 to 2014). The answers of wave 5 are used as initial opinions that are evolved, according to the model dynamics, so as to produce *final* opinions that are compared with the survey results from wave 6. Our main purpose is to present a new opinion formation model; through the comparison with real data, we identify parameter choices showing that the model has the *potential* to accurately predict real opinions starting from a variety of different initial opinions, but this does not fully or univocally solve the inverse problem (Kandler & Powell 2018).

1.7 The paper is structured as follows. First, it introduces the model and its key parameters. Then, four types of simulation results are presented: simulations with simple parameters and digraphs, to gain intuition on the model behaviour; parameter sensitivity analysis, to explore the effect of different parameters on the opinion evolution; characterisation of model outcomes, also using distributional measures (Lorenz et al. 2021); model validation with real data, to assess the predictive potential of the model by choosing the parameters through an optimisation problem. The Appendices include a comparison with the French-DeGroot (DeGroot 1974) and the Friedkin-Johnsen (Friedkin & Johnsen 1999, 2011) models, as well as more details about the considered optimisation approach and simulation results, and the data from the WVS.

● The Classification-Based Model

2.1 In our proposed Classification-based (CB) model, the set $V = \{1, 2, \ldots, n\}$ indexes the agents. The *opinion* of agent $i \in V$ at time k, representing its level of agreement with a statement, is denoted by $x_i[k] \in [-1, 1]$. The opinions $x_i = 1$, $x_i = 0$, and $x_i = -1$ represent complete agreement, indifference, and complete disagreement respectively. The vector of all opinions at time k is denoted by $x[k]$.

2.2 The agent opinions evolve in discrete time due to opinion exchanges occurring over a signed digraph, represented by the matrix $W \in \{-1, 0, 1\}^{n \times n}$, whose entries are constant and, in particular, not opinion-dependent. The self-confidence of each agent is expressed by $w_{ii} = 1$ for all i. The coefficient w_{ij} represents the influence of agent j over agent i. If $w_{ij} = 0$, then agent i is not influenced by agent j. If $w_{ij} \neq 0$, then agent j is a neighbour of agent i: $w_{ij} = 1$ means that agent i approves, trusts, or follows agent j, while $w_{ij} = -1$ means that agent i disapproves, mistrusts, or antagonises agent j. Signed edges have been interpreted in the opinion formation literature in terms of either cooperative/antagonistic interactions (Altafini 2013), trust/mistrust (Xia et al. 2016), or approval/disapproval (Cartwright & Harary 1956). In our model, if $w_{ij} = 1$ (respectively $w_{ij} = -1$), then agent i perceives the opinion of agent j as x_j (resp. $-x_j$). The set of neighbours of agent $i \in V$ is

$$\mathcal{N}_i = \Big\{ j \in V \mid w_{ij} \neq 0 \Big\}. \tag{1}$$

2.3 The agent opinions evolve in discrete time and the opinion update relies on the assumption that agents cannot determine their neighbours' opinions precisely. Instead, each agent can classify its neighbours according to how close their *perceived* opinion is to its own opinion. For instance, if agent j influences agent i, and $x_i = 0.61$ and $x_j = 0.34$, then it is unrealistic to expect agent i to know *exactly* the opinion of agent j, or to assume that agent i knows that the opinion difference is *exactly* 0.27. However, agent i can perceive that agent j agrees less than itself. On the contrary, if $x_j = 0.89$, agent i can perceive that agent j agrees more than itself.

2.4 Therefore, agent i can at most classify agent j according to an estimation of Δ_{ij}, which is the weighted difference between its opinion x_i and the opinion of agent j, x_j: $\Delta_{ij} = x_i - w_{ij}x_j \in [-2, 2]$. Let us divide the interval $[-2, 2]$ in five equal subintervals. Then, depending on the subinterval to which Δ_{ij} belongs, agent i can *perceive* that agent j: (1) agrees much more, (2) agrees more, (3) agrees comparably, (4) agrees less, or (5) agrees much less with the statement; see Figure 1. If $w_{ij} = -1$, then agent i disapproves/mistrusts/antagonises agent j, therefore the weighted opinion difference is $\Delta_{ij} = x_i - (-x_j) \in [-2, 2]$. If $w_{ij} = 1$, then agent i approves/trusts/follows agent j and the weighted opinion difference is $\Delta_{ij} = x_i - x_j \in [-2, 2]$.
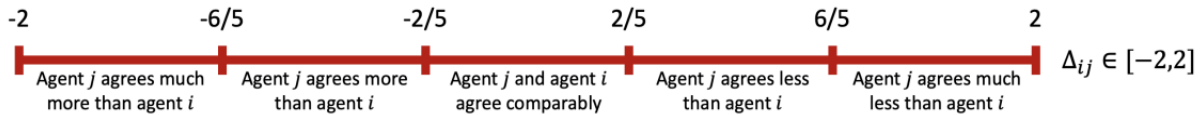

Figure 1: Partition of the interval $[-2, 2]$ in five equal subintervals. Depending on the interval to which the weighted opinion difference $\Delta_{ij} = x_i - w_{ij}x_j$ belongs, agent i will *perceive* that agent j agrees either: *much more*; or *more*; or *comparably*; or *less*; or *much less*.

2.5 The combined effect of signed edges and neighbour classification leads to a three-step process: first, agent i perceives the opinions of its neighbours; then, the opinions of neighbours that agent i disapproves, mistrusts, or antagonises have the sign reversed; finally, the neighbours are classified according to the adjusted perceived opinion distance.

2.6 The set $\mathcal{N}_i$ of all the neighbours of agent i is thus partitioned into five time-dependent subsets: $D_i^+[k]$, $D_i[k]$, $N_i[k]$, $A_i[k]$, and $A_i^+[k]$, which contain the neighbours that agree much less, less, comparably, more, and much more, respectively. Mathematically these subsets are defined as:

$$
\begin{aligned}
D_i^+[k] &= \big\{ j \in \mathcal{N}_i \mid {}^6\!/_5 \leq \Delta_{ij}[k] \leq 2 \big\} \\
D_i[k] &= \big\{ j \in \mathcal{N}_i \mid {}^2\!/_5 \leq \Delta_{ij}[k] < {}^6\!/_5 \big\} \\
N_i[k] &= \big\{ j \in \mathcal{N}_i \mid {}^{-2}\!/_5 < \Delta_{ij}[k] < {}^2\!/_5 \big\} \\
A_i[k] &= \big\{ j \in \mathcal{N}_i \mid {}^{-6}\!/_5 < \Delta_{ij}[k] \leq {}^{-2}\!/_5 \big\} \\
A_i^+[k] &= \big\{ j \in \mathcal{N}_i \mid -2 \leq \Delta_{ij}[k] \leq {}^{-6}\!/_5 \big\}
\end{aligned}
\tag{2}
$$

where $\Delta_{ij}[k] = x_i[k] - w_{ij}x_j[k]$. The cardinality of these sets has the following interpretation:

$|D_i^+[k]| =$ number of neighbours that agent i *perceives* as agreeing much less than itself at time k

$|D_i[k]| =$ number of neighbours that agent i *perceives* as agreeing less than itself at time k

$|N_i[k]| =$ number of neighbours that agent i *perceives* as agreeing the same as itself at time k

$|A_i[k]| =$ number of neighbours that agent i *perceives* as agreeing more than itself at time k

$|A_i^+[k]| =$ number of neighbours that agent i *perceives* as agreeing much more than itself at time k

2.7 The overall behaviour of each agent results from the combination of three complementary inner traits: *conformism*, leading the agent to agree with its neighbours; *radicalism*, driving the agent to reinforce its opinion; and *stubbornness*, anchoring the agent to its current opinion. The conformism, radicalism and stubbornness degree of agent i is respectively denoted by α_i, β_i and γ_i. The parameters $\psi_i = (\alpha_i, \beta_i, \gamma_i)$, quantifying the *inner traits* of agent i, satisfy $\alpha_i, \beta_i, \gamma_i \in [0, 1]$ and $\alpha_i + \beta_i + \gamma_i = 1$ for all i. We call *inner traits assignation* the collection of inner traits of all agents, $\psi := (\psi_i)_{i \in V}$. The model features are summarised in Figure 2.

2.8 The opinion change $\Delta x_i[k]$ of agent i at time k is thus the convex combination of the behaviour of a purely conformist, purely radical, and purely stubborn agent,

$$\Delta x_i[k] = \alpha_i f_i^{\text{con}} + \beta_i f_i^{\text{rad}} + \gamma_i f_i^{\text{stb}}, \tag{3}$$

with f_i^{con}, f_i^{rad}, and f_i^{stb} taken as:

$$f_i^{\text{con}} = \frac{\lambda}{|\mathcal{N}_i|}\Big(\xi|A_i^+| + |A_i| - |D_i| - \xi|D_i^+|\Big), \qquad f_i^{\text{rad}} = \frac{\lambda}{|\mathcal{N}_i|}\mu|N_i|x_i[k], \qquad f_i^{\text{stb}} = 0, \tag{4}$$

where λ, ξ, and μ are positive parameters: λ weighs the overall opinion change magnitude, ξ weighs the increased influence that neighbours with distant opinions have over conformist traits, and μ weighs the influence of the agent's own opinion in radical traits. We call these *opinion evolution parameters*: $\Omega = (\lambda, \xi, \mu)$; see Figure 2.

2.9 To better understand Equations (4) and choose reasonable values for the parameters, one can think of how an extreme agent ($\alpha_i = 1$, or $\beta_i = 1$, or $\gamma_i = 1$) behaves.

- A purely conformist agent ($\alpha_i = 1$, $\beta_i = 0$, $\gamma_i = 0$) evolves towards an opinion comparable to that of its neighbours. For instance, if $N_i = \mathcal{N}_i$ (all the neighbours of agent i agree comparably), then agent i does not change its opinion. If $A_i = \mathcal{N}_i$ (all the neighbours of agent i agree more), agent i increases its opinion x_i by λ; given that all the neighbours of agent i are in the set A_i, a value $\lambda = 0.4$ guarantees that, if all the neighbour opinions remain unchanged, then at the next time step all the neighbours of agent i will be in the set N_i, hence perceived as having a comparable opinion. Instead, if $A_i^+ = \mathcal{N}_i$, then the opinion of agent i needs to increase $0.8 = 2\lambda$ in order to be perceived as comparable to its neighbours' at the next time step, and therefore a natural choice is $\xi = 2$. The same reasoning can be applied to the sets D_i and D_i^+.

- A purely radical agent ($\alpha_i = 0$, $\beta_i = 1$, $\gamma_i = 0$) ignores neighbours with a different opinion and only cares about agents that think comparably to itself, hence it reinforces its current opinion $x_i[k]$ depending on the magnitude of its own opinion and on the fraction of its neighbours in the set N_i. To make sure that radical traits can affect the opinion change more strongly than conformist traits, we need $\mu > 1$. In fact, if $\mu = 1$, then $|f_i^{\text{rad}}| < |f_i^{\text{con}}|$ in general: the opinion change caused by the radical trait (which is proportional to $x_i[k]$, and $|x_i[k]| \leq 1$) is smaller in magnitude than the one caused by the conformist trait. In our simulations, we set $\mu = 5$. The effect of different values of μ can be seen in Table 5.

- A purely stubborn agent ($\alpha_i = 0$, $\beta_i = 0$, $\gamma_i = 1$) does not change its opinion under any circumstance.

2.10 The new opinion of agent i at time $k+1$ is the sum of the previous opinion $x_i[k]$ and the opinion change $\Delta x_i[k]$, modulated by the saturation function σ

$$\sigma(x) = \begin{cases} x & \text{if} \quad |x| \leq 1 \\ \text{sign}(x) & \text{if} \quad |x| > 1 \end{cases} \tag{5}$$

so as to guarantee that the opinions remain in the interval $[-1, 1]$. The complete opinion update law is therefore:

$$x_i[k+1] = \sigma\left(x_i[k] + \frac{\lambda}{|\mathcal{N}_i|}\Big(\alpha_i\xi\big(|A_i^+| - |D_i^+|\big) + \alpha_i\big(|A_i| - |D_i|\big) + \beta_i\mu|N_i|x_i[k]\Big)\right), \qquad \forall i \in V. \tag{6}$$

2.11 The opinions evolve, with simultaneous periodic updates at fixed discrete-time instants, according to Equation (6), which is fully deterministic: at each time step, every agent updates its opinion according to Equation (6), relying on the neighbour classification obtained through Equation (2). The parameter values and the signed digraph are constant and, once they are assigned, the model evolution is completely determined by the initial opinions.

2.12 The mathematical model formulation relies on four main assumptions: 1) opinions can be represented by real numbers in the [-1, 1] interval and agents update their opinions simultaneously over discrete time instants (standard assumptions for opinion formation models); 2) agents can either trust or mistrust the opinions they perceive from their neighbours, as reflected by the edge signs in the digraph (Altafini 2013; Xia et al. 2016); 3) agents cannot measure the opinion of their neighbours with infinite accuracy, but they can classify them according to their perceived opinions, through the classification mechanism in Equation (2); 4) the behaviour of each agent is the result of a combination of three psychological traits: conformism, radicalism, and stubbornness. These four assumptions, together with their mathematical representation in Equation (6), are the foundations for the proposed model.

Model parameters

2.13 The Classification-based (CB) model has three types of parameters: the signed digraph weights w_{ij}; the inner traits assignation $\psi_i = (\alpha_i, \beta_i, \gamma_i)$; and the opinion evolution parameters $\Omega = (\lambda, \xi, \mu) = (0.4, 2, 5)$ whose values are fixed, and chosen based on the model interpretation. Later, a parameter sensitivity analysis explores how the model evolution is affected by changes in opinion evolution parameters.

2.14 If the model has n agents, then:

- The signed digraph has weight matrix $W \in \mathcal{W}_n$. In general, $\mathcal{W}_n = \{-1, 0, 1\}^{n \times n}$, but we can focus for instance on small-world, or strongly connected, networks.

- The inner traits assignation is $\psi \in \mathcal{A}_n$, where:

$$\mathcal{A}_n = \left\{ \psi = (\psi_i)_{i \in V} = \left((\alpha_i, \beta_i, \gamma_i) \right)_{i=1}^{n} \mid \alpha_i, \beta_i, \gamma_i \in [0, 1] \quad \text{and} \quad \alpha_i + \beta_i + \gamma_i = 1, \quad \forall i \in V \right\}. \quad (7)$$

2.15 We omit the subscript n from the sets $\mathcal{W}$ and $\mathcal{A}$ for simplicity. Given n agents, a signed digraph $W \in \mathcal{W}$, an inner traits assignation $\psi \in \mathcal{A}$, and a vector of initial opinions $x[0]$, the opinion formation model evolves according to Equation (6). The vector $x[K]$ of opinions after K iterations can be explicitly represented as a function of W, ψ, and $x[0]$ by the map $\mathcal{F}_\Omega$ ($x[K]$ also depends on Ω, whose value, given by the model interpretation, is fixed) as:

$$x[K] = \mathcal{F}_\Omega(x[0], W, \psi, K) \qquad (8)$$

2.16 The value of K depends on the type of statements and the prediction horizon. For statements related to core values or beliefs, opinions are not expected to change very fast and one could consider roughly 10 changes per year. Therefore, if the model is used to predict the opinions after 5 years, $K = 50$. On the other hand, the opinions on more superficial topics could change faster and, over the same 5-year timespan, it could be $K = 500$. See Figure 2 for a summary of the model parameters and features.
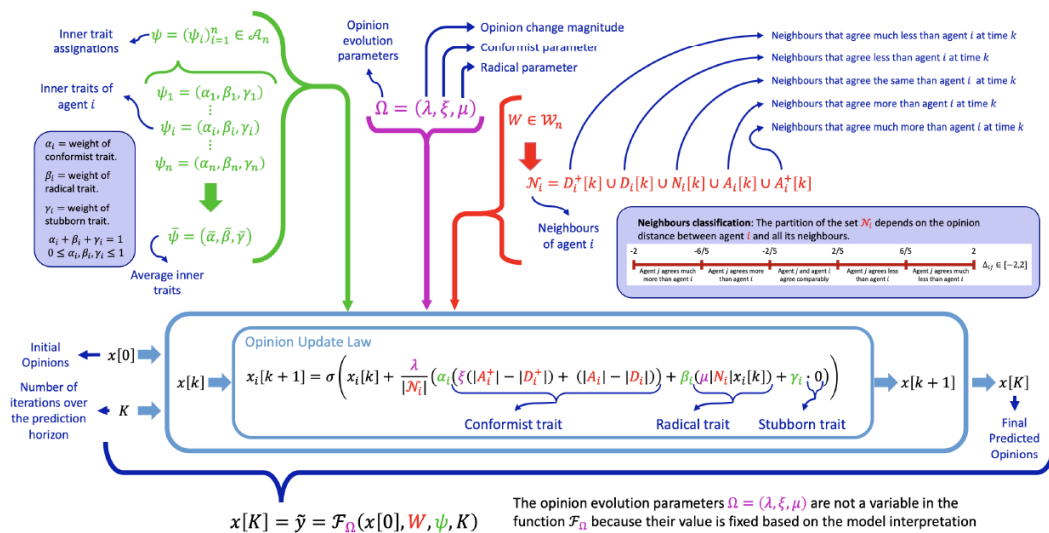
Figure 2: Visualisation of the model features and parameters. The model has three parameter types: inner traits assignation ψ (in green), opinion evolution parameters Ω (in magenta), and signed digraph weights W (in red). These parameters appear in the Opinion Update Law of Equation (6), which is a convex combination of contributions by conformist, radical, and stubborn traits, with the opinions of neighbouring agents evaluated through a classification-based approach. Given the initial conditions $x[0]$ and the number K of iterations over the prediction horizon, the Opinion Update Law produces the final predicted opinions $\tilde{y} = x[K]$. The opinion evolution parameters Ω can be fixed based on the model interpretation. Then, the final predicted opinions in each particular case are a function $x[K] = \mathcal{F}_{\Omega}(x[0], W, \psi, K)$ of the chosen initial opinions, signed digraph weights, inner traits assignation, and number of iterations in the prediction horizon.

2.17 To validate the model – namely, assess its potential to closely reproduce the evolution of opinions in real life with suitably chosen parameters – we consider *real* initial and final opinions, denoted by x and y respectively, taken from survey data. Assuming that y are the real opinions K iterations after the real initial opinions x, these data can be used to find values of the model parameters (edge weights W and inner traits ψ) that produce opinions that match as closely as possible the real opinion evolution, through the minimisation problem:

$$(\widehat{W}, \widehat{\psi}) = \underset{\substack{W \in \mathcal{W} \\ \psi \in \mathcal{A}}}{\arg\min} J(y, \tilde{y}) \qquad \text{such that} \qquad \tilde{y} = \mathcal{F}_{\Omega}(x, W, \psi, K), \qquad (9)$$

where the cost function $J(y, \tilde{y})$, described and explained in Equation (15), quantifies the mismatch between opinion vectors y and $\tilde{y}$.

2.18 If the same population is asked to quantify the agreement with Q different statements, the signed digraph cannot change. However, the inner traits assignation can vary depending on the statement, since each individual may have different attitudes towards different topics. Therefore, if $\psi^{(l)}$ represents the inner traits assignation associated with statement l, values for the parameters W and $(\psi^{(l)})_{l=1}^{Q}$ that produce predicted opinions as similar as possible to the real ones can be found through the *free optimisation problem*

$$\left(\widehat{W}, \widehat{(\psi^{(l)})}_{l=1}^{Q}\right) = \underset{\substack{W \in \mathcal{W} \\ \psi^{(l)} \in \mathcal{A}}}{\arg\min} \sum_{l=1}^{Q} J(y_l, \tilde{y}_l) \qquad \tilde{y}_l = \mathcal{F}_{\Omega}(x_l, W, \psi^{(l)}, K) \qquad (10)$$

where x_l and y_l are the known initial and final opinions related to statement l.

2.19 If instead all the inner traits assignations are constrained to be the same for every question, we consider the *constrained optimisation problem*:

$$(\widehat{W}, \widehat{\psi}) = \underset{\substack{W \in \mathcal{W} \\ \psi \in \mathcal{A}}}{\arg\min} \sum_{l=1}^{Q} J(y_l, \tilde{y}_l) \qquad \tilde{y}_l = \mathcal{F}_{\Omega}(x_l, W, \psi, K) \qquad (11)$$

2.20 The free optimisation problem, where the inner assignations can change, allows for a more thorough study of the behaviour of a population, while the constrained optimisation problem allows for a more rigorous testing

of the prediction capabilities of the model in the form of cross-validation: the answers to some questions can be used as training datasets to choose the model parameters, and the model performance can then be tested on the remaining questions.

Simulation Results

3.1 To gain insight into the Classification-based (CB) model, this section presents four different types of simulation results: 1) **Simulations in Simple Cases** evolve the model in simple, special cases to gain intuition into its behaviour; 2) **Parameter Sensitivity Analysis** studies how changes in each of the model parameters (inner traits assignation, signed digraph, opinion evolution parameters) affect the model behaviour; 3) **Characterisation of Model Outcomes** uses distributional measures, including the recently proposed *Bias*, *Diversity*, and *Fragmentation* (Lorenz et al. 2021), to characterise the variety of opinion vectors that the Classification-based model can produce; 4) **Model Validation with Real Data** leverages real data from the WVS to show that the CB model has the potential to reproduce the time evolution of real opinions in society (with parameters chosen through the *free* and the *constrained* optimisation problems of Equations (10) and (11) respectively).

Due to the deterministic nature of the model, running it with the same initial opinions, parameters, and interconnection network always produces the same results. Given a parameter choice and a network, the model evolution can only change due to different initial conditions (see the repeated model runs in Figure 5).

3.2 To facilitate the interpretation of simulation results, we introduce some definitions. Given the inner traits assignation $\psi = (\psi_i)_{i \in V} = \left((\alpha_i, \beta_i, \gamma_i) \right)_{i \in V}$, the associated *average inner traits*

$$\bar{\psi} = (\bar{\alpha}, \bar{\beta}, \bar{\gamma}) \quad \text{where} \quad \bar{\alpha} = \frac{1}{n} \sum_{i \in V} \alpha_i \qquad \bar{\beta} = \frac{1}{n} \sum_{i \in V} \beta_i \qquad \bar{\gamma} = \frac{1}{n} \sum_{i \in V} \gamma_i, \tag{12}$$

represent the traits of an average agent in the considered society or population with $n = |V|$ agents. Inner traits assignation ψ and the corresponding average inner traits $\bar{\psi}$ can be plotted in a ternary diagram as shown in Figure 3a (the ternary plots were made using modified scripts from Sandrock (2012)). Figure 3b explains how to interpret a point in the ternary diagram.
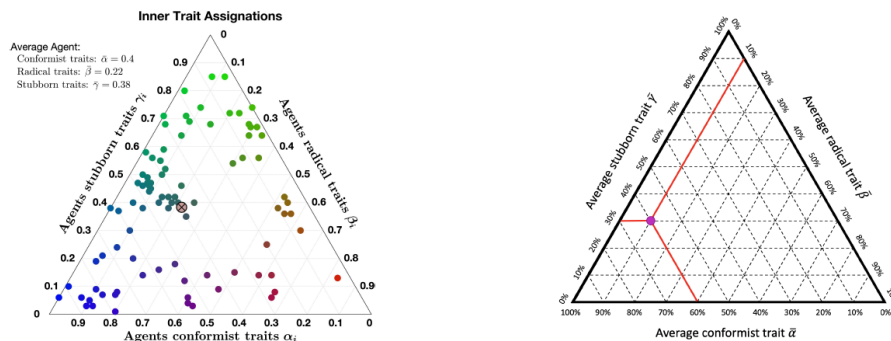
(a) Each dot represents the inner traits of an agent; its RGB colour reflects the weight of each trait (blue: conformist; red: radical; green: stubborn). The crossed dot represents the average inner traits.

(b) Example of average inner traits in the ternary diagram: 60% conformist, 10% radical, 30% stubborn.

Figure 3: Ternary diagrams visualising inner traits assignations ψ and average inner traits $\bar{\psi}$. Panel (a) shows the whole inner traits assignation, with each dot corresponding to the traits of a singe agent, along with the average inner traits (crossed dot). Panel (b) only shows the average inner traits of a complete population (magenta dot).

3.3 The *general agreement* of an opinion vector $x = (x_i)_{i=1}^n$, quantified by the pair (θ_+, θ_-) where

$$\theta_- = \sum_{x_i < 0} x_i \qquad \text{and} \qquad \theta_+ = \sum_{x_i > 0} x_i, \tag{13}$$

and can be interpreted as the overall level of agreement and disagreement in the whole society.

3.4 All the simulations involve a population of $n = 100$ agents. All the digraphs used in both Parameter Sensitivity Analysis and Model Validation with Real Data have a small-world network topology, with an assigned

probability for positive and negative edges, and are strongly connected. We consider small-world networks because they have a high clustering coefficient (neighbours of neighbours of agent i are likely also neighbours of agent i) and low diameter (maximum distance between two agents of the network), which are believed to be characteristics of real-life social networks (Elgazzar 2003; Watts & Strogatz 1998). The directed small-world networks were built based on the Watts-Strogatz algorithm. Appendix D describes the computation of network metrics. The signed digraphs are not restricted to be structurally balanced, to account for the fact that also non-structurally-balanced networks have been considered in the literature when modelling social dynamics (Estrada 2019; Leinhardt 1977; Opp 1984).

3.5 In all the considered simulations, the initial opinions, traits and networks are assigned independently. A different approach – which is left for future work – could be to assign them in some correlated way: e.g., initial opinions and network could be correlated by assigning the initial opinions such that two vertices connected by an edge have a very similar (or very distant) initial opinion; traits and network could be correlated by assigning the agent parameters with a probability that depends on the corresponding vertex characteristics, for example assuming that vertices with higher out-degree have a higher probability of being completely conformist, or radical. Correlations between initial opinions, traits, and network characteristics can reproduce different types of societies present in real life (for instance, in a society that values tradition, highly stubborn agents may be more influential than others, and hence the corresponding vertices may have a higher out-degree).

Simulations in simple cases

3.6 To better understand the model behaviour, we simulate the model evolution in special simple cases. First, for the same digraph with a lattice topology, we vary the inner traits assignations (Figure 4). Then, for the same inner traits assignation, we consider different digraph topologies (complete, lattice, ring, small-world) and different, randomly chosen, initial opinions (Figure 5).

Different inner traits assignations

3.7 We consider a signed lattice digraph, where each agent has 4 in-neighbours and the edges are positive with probability 0.77. All the agents have the same inner traits, combining only two inner traits: stubbornness and radicalism; radicalism and conformism; conformism and stubbornness. Starting from the same initial opinions (which are discrete, and hence the plot with purely stubborn agents shows evenly distributed horizontal lines), Figure 4 shows the opinion evolution over 30 time steps.
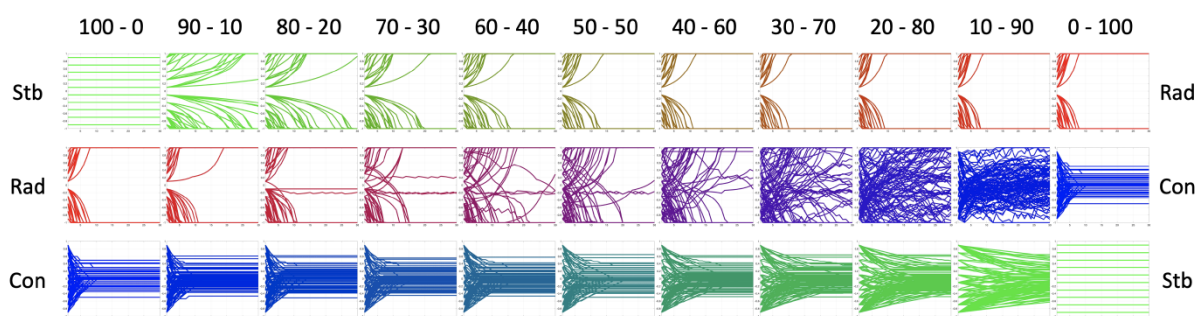
Figure 4: Evolution of the CB model, starting from the same initial opinions, over a signed lattice digraph where the edges are positive with probability 0.77. In each simulation, all 100 agents have the same inner traits that are a combination of only two of the possible traits: stubbornness (Stb), radicalism (Rad), and conformism (Con). The top labels show the proportion of each trait: the upper left (respectively, right) graph corresponds to a simulation where all agents are purely stubborn (resp. radical). The colour of the lines is the RGB representation of the inner traits assignations (blue: conformist; red: radical; green: stubborn).

3.8 Figure 4 shows that Radicalism tends to form polarisation by driving the agents to extreme opposite views. Conformism tends to create consensus; however, because of the classification approach, the agents do not converge to the very same opinion (close enough agents are unable to perceive their opinion difference, because opinions are assessed with finite resolution). Stubbornness slows down the effect of the other two traits; only in a fully stubborn population everyone keeps its initial opinion. Among the three traits, radicalism appears to

have the greatest effect: even a small amount of radicalism can prevent conformism from forming consensus (as seen in the 10-90 Rad-Con plot), and can yield polarisation in a very stubborn society (as seen in the 90-10 Stb-Rad plot).

Different digraph topologies

3.9 Additional intuition on the model behaviour can be gained by studying the effect of different initial opinions and different digraph topologies with fixed inner trait parameters. Figure 5 shows 10 simulations starting from various, randomly chosen, initial opinions and evolved over four signed digraphs with Complete, Lattice, Ring, and Small-World topologies. The inner traits assignation for all these simulations is kept constant and is shown in Figure 3a.

3.10 Both the digraph topology (dictating how the agents communicate among them) and the initial opinions (providing the starting point of the evolution) have a significant effect on the opinion evolution and the final predicted opinions. For Lattice and Ring digraphs, there is a clear tendency towards consensus at one extreme opinion (completely agree or completely disagree), even when, as in this case, the average radical trait is relatively low. A possible explanation is that in both these topologies agents have less in-neighbours, so the radical trait can have a stronger effect. Another possible explanation is that both these types of networks have a larger average path length, and diameter, than Complete and Small-World networks, and therefore the 'consensus effect' takes more time to act than in more connected networks.

3.11 Indeed, since they share common features, the Complete and Small-World digraphs (small diameter), as well as the Lattice and Ring digraphs (large diameter), showcase similar behaviours and similar final opinion distributions, across all the chosen initial opinion distributions. We have observed that this tendency is recurrent for several different choices of inner traits assignations.
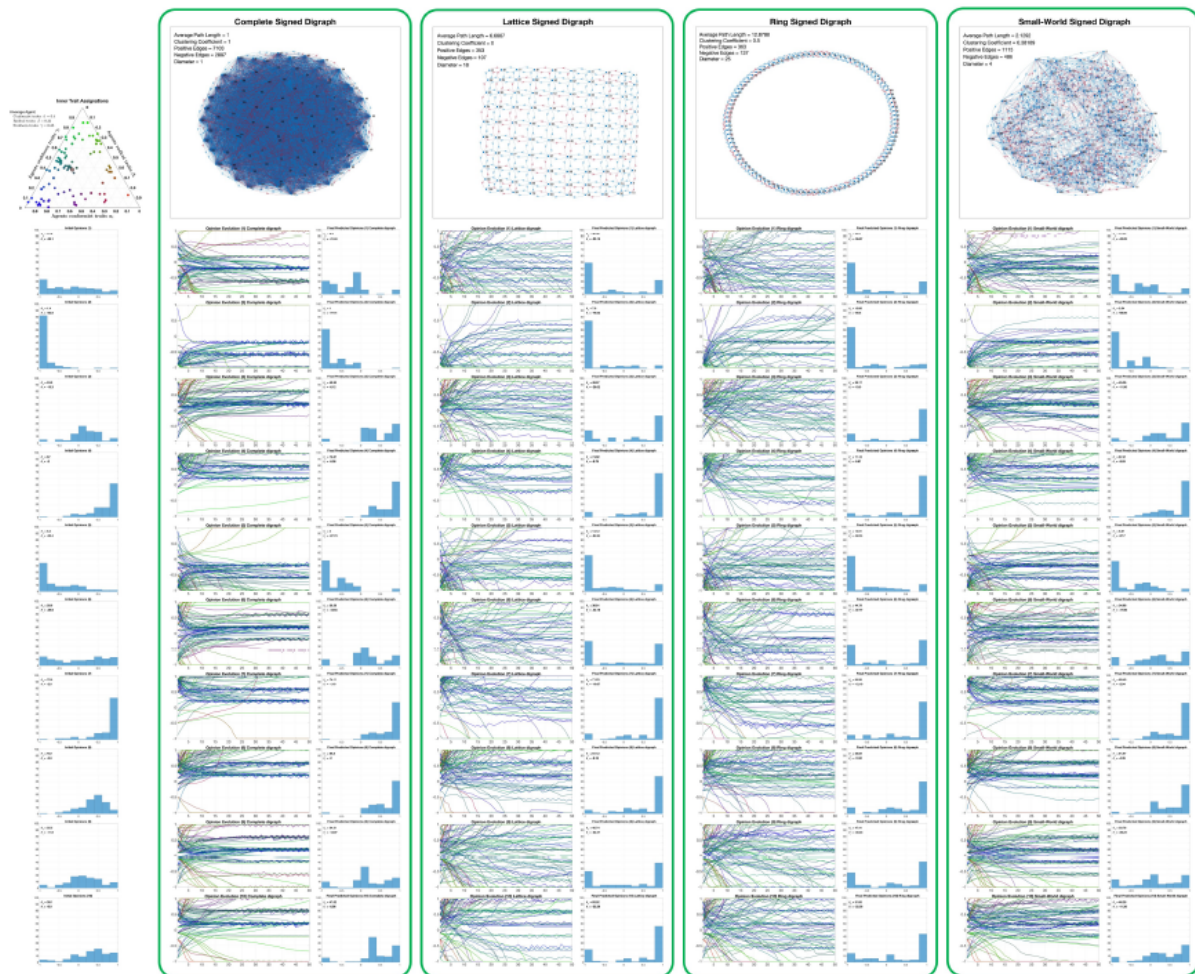
Figure 5: Evolution of the CB model, with 100 agents, over four different digraph topologies and starting from ten different, randomly chosen, initial opinions, shown on the left-most column. Each of the four columns framed in green shows the opinion evolution and the final opinion distribution for each of the initial opinions, for a different choice of the digraph (Complete, Lattice, Ring, Small-World), shown at the top of the column along with its metrics. The inner traits assignation is presented at the top left and has average traits: $\overline{\alpha} = 0.4$ (conformist), $\overline{\beta} = 0.22$ (radical), and $\overline{\gamma} = 0.38$ (stubborn). The ternary diagram of these inner traits can also be seen in Figure 3a.

3.12 As is apparent from Figure 4, the inner traits assignation has a tremendous effect on the opinion evolution. The simulations of Figure 5 reinforce this idea by showing that, although the initial opinions and digraph topology do have an impact, keeping the same inner traits assignation restricts the final opinion distributions to some characteristic patterns.

Parameter sensitivity analysis

3.13 We select a set of *nominal parameters* (which, for given initial conditions, produce *nominal simulation results*) as a baseline with which other parameter choices can be compared. We choose a nominal inner traits assignation that leads to model outcomes that closely reproduce real data from the World Values Survey (in fact, it is close to some of the inner traits assignations resulting from the *Free* optimisation problem (10), see Figure 19a), and therefore has the potential to represent a realistic society; moreover, it allows us to showcase the wide range of different opinion evolutions that the model can produce. Then, we vary inner traits assignations, signed digraph and opinion evolution parameters, one by one, and study their effect on the simulated behaviour.

Nominal parameters and nominal results

3.14 We consider the initial opinions shown in Figure 6a, which evolve according to the model with the nominal parameters: $\lambda = 0.4, \xi = 2, \mu = 5$, inner traits assignations in Figure 6b, and signed digraph in Figure 6c.

| (a) Initial Opinions Histogram | (b) Nominal Inner Traits Assignation | (c) Nominal Signed Digraph |
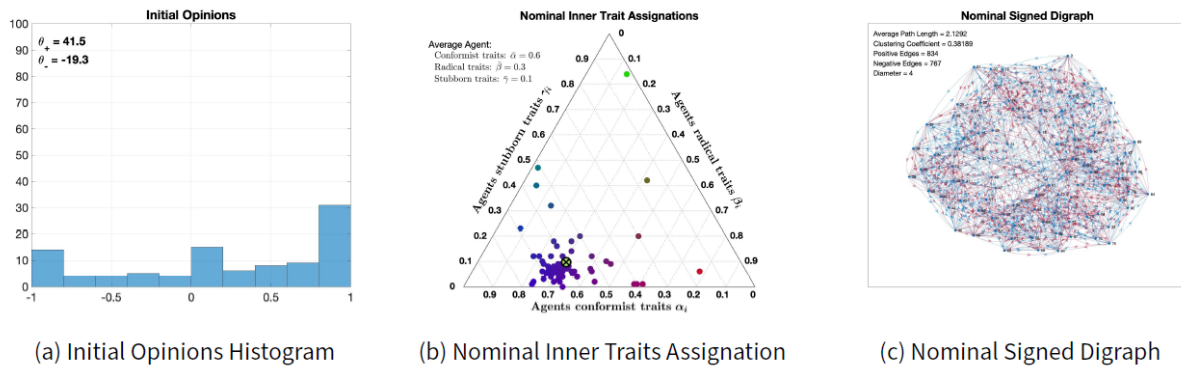
Figure 6: Initial opinions and nominal parameters.

3.15 The initial opinions shown in Figure 6a have $\theta_- = -19.3$ and $\theta_+ = 41.5$, indicating a strong general agreement since $\theta_+ > -\theta_-$. Figure 6b shows that most agents have very strong conformist traits, with a notable percentage of radicalism, resulting in an average agent (crossed dot) with 60% conformist traits, 30% radical traits, and 10% stubborn traits. The nominal signed digraph in Figure 6c is highly connected, with average path length 2.12, clustering coefficient 0.38, diameter 4. It has 834 positive edges and 767 negative edges.

3.16 The nominal results are shown in Figure 7. Figure 7a shows the opinion evolution of every agent. The line colour represents the percentage of conformist, radical, and stubborn agent traits (blue for conformist, red for radical, and green for stubborn). The purple colour of most lines corresponds to a combination of conformist and radical traits. The discontinuity in the opinion change is due to the classification process leading to a discontinuous opinion update law. The opinion evolution of the various agents shows a great variability in opinion changes, without a clear global tendency.

| (a) Nominal Opinion Evolution | (b) Nominal Final Opinion Histogram |
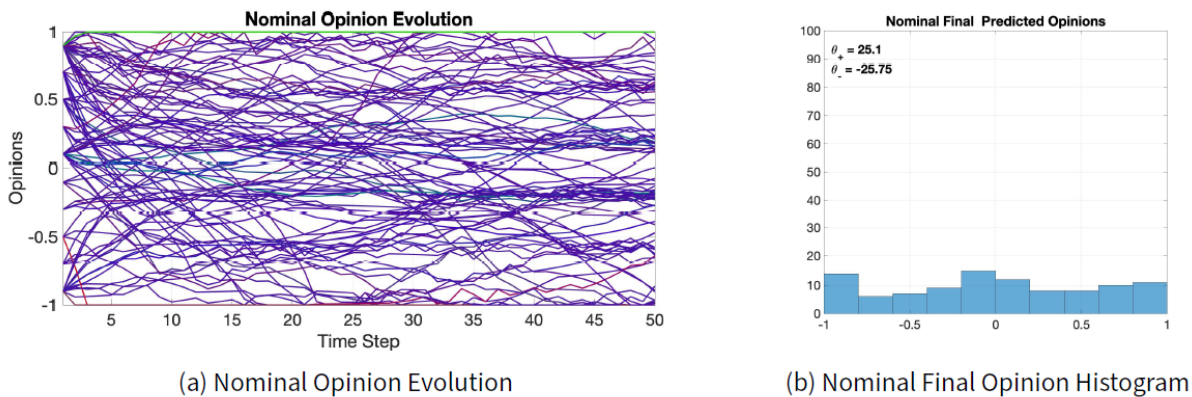
Figure 7: Simulation results with the nominal parameter values (evolving 100 agents).

3.17 Figure 7b shows the histogram of the nominal final opinions predicted by the model after 50 time steps. Compared with the initial opinions, the final opinions appear to have a more uniform distribution: in fact, for the nominal final opinions, $\theta_- = -25.75$ and $\theta_+ = 25.1$, hence $\theta_+ \approx -\theta_-$. The behaviour of the opinion evolution and the distribution of the final opinions is explained by the presence of two opposing forces that drive the opinion of all the agents: on one hand, the tendency to achieve consensus, due to the conformist traits, drives the agents towards the centre; on the other hand, the radical traits move the opinions towards extreme values.

Varying the inner traits assignations ψ

3.18 To evaluate the effect of different inner traits assignations, we change the nominal inner traits assignations of Figure 6b and simulate the opinion evolution, keeping all the other parameters unchanged. The two new inner traits assignations, shown in Figures 8a and 8d, are simply rotations of the nominal inner traits assignations. The corresponding opinion evolutions are shown in Figures 8b and 8e, while the final opinion histograms are presented in Figures 8c and 8f.

(a) Inner traits assignations 1

(b) Opinion evolution inner traits assignations 1

(c) Final opinion histogram

(d) Inner traits assignations 2

(e) Opinion evolution inner traits assignations 2

(f) Final opinion histogram
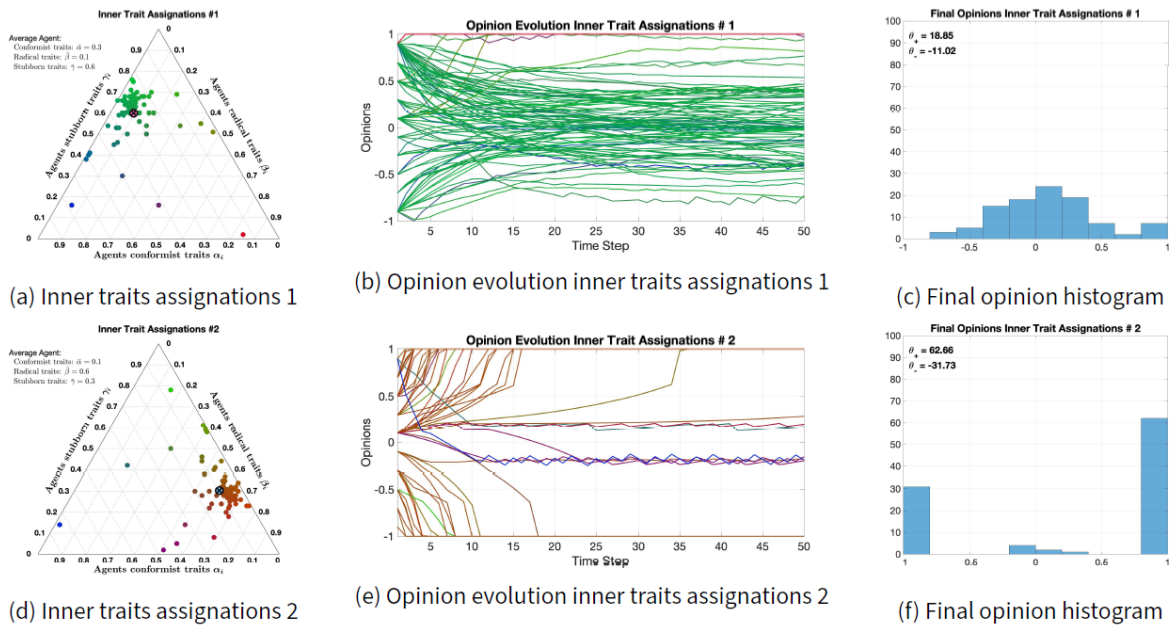
Figure 8: Effect of changing the inner traits assignations (evolving 100 agents).

3.19 Comparing the opinion evolutions of Figures 7a, 8b, and 8e and the final opinion histograms of Figures 7b, 8c, and 8f reveals the profound effect of different inner traits assignations on the opinion evolution. In the inner traits assignation of Figure 8a, the agents are mostly stubborn and conformist. This results in a very slow convergence towards the mean, driven by conformist traits and slowed down by stubborn traits. Because of the neighbour classification mechanism, even completely conformist agents would not reach perfect consensus, but would rather converge to an opinion subinterval where all the agents perceive that the others have a comparable opinion. This tendency towards the mean can be seen in the final opinion histogram of Figure 8c, where both $\theta_- = -11.02$ and $\theta_+ = 18.85$ are much closer to 0 than the final opinions in Figure 7b.

3.20 On the other hand, the inner traits assignation of Figure 8d gives agents pronounced radical traits. Both the opinion evolution in Figure 8e and the final opinion histogram in Figure 8f show that agents lean towards extreme opinions. A bunch of agents keeps its opinion closer to zero. The line colours (closer to blue and green) show that these agents do not have very strong radical traits, and instead they are more conformist and stubborn: such traits allow these agents to avoid extreme opinions.

Varying the signed digraph W

3.21 To study the effect of changing the signs of the weights of the signed digraph, the nominal signed digraph of Figure 6c is modified into the signed digraphs shown in Figures 9a and 9d. The topology is unchanged, but the number of positive and negative edges is changed. The resulting opinion evolution and final opinion histograms are shown in Figures 9b and 9e, and in Figures 9c and 9f respectively.

_

(a) Signed digraph 1   (b) Opinion evolution signed digraph 1   (c) Final opinion histogram

(d) Signed digraph 2   (e) Opinion evolution signed digraph 2   (f) Final opinion histogram
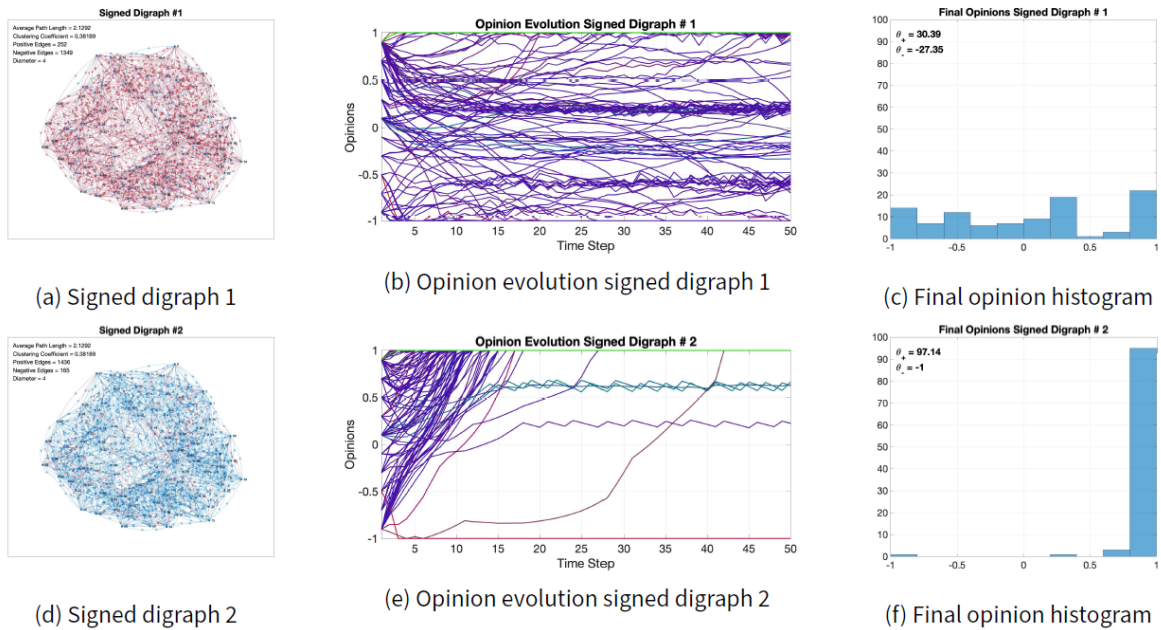
Figure 9: Effect of changing the signed digraph (evolving 100 agents).

3.22   Compared with the nominal results in Figures 7a and 7b, the most different outcome occurs when most edges are positive (digraph in Figure 9d). In this case, the end result is almost perfect consensus for the $+1$ opinion, because the initial opinion, with $\theta_- = -19.3$ and $\theta_+ = 41.5$, is more skewed towards $+1$. The presence of negative edges is crucial to avoid trivial consensus outcomes even when the agents are not completely conformist. The opinion evolution in Figure 9e shows that, initially, conformist traits pull the opinions towards positive values, and then radical traits make them increase in value until they reach $+1$. Purely radical agents would have produced polarisation instead of consensus.

3.23   When increasing the number of negative edges (digraph in Figure 9a), the final opinions in Figure 9c are different from the nominal ones, but the qualitative behaviour is comparable.

Varying the opinion evolution parameters Ω

3.24   We study the sensitivity with respect to the opinion evolution parameters $\Omega = (\lambda, \xi, \mu)$, where: λ is the overall opinion change magnitude, and can also be thought of as a time scaling parameter; ξ gives more weight to distant opinions for conformist traits; μ increases the opinion change for radical traits. We change these parameters one at the time, with respect to the nominal parameters, and compare the results with the nominal results in Figure 7.

3.25   Figure 10 shows the opinion evolution and final histogram for $\lambda = 0.2$ and $\lambda = 0.8$. The final histograms in Figures 10b and 10d do not change much with respect to the nominal. The most significant change can be noticed in Figures 10a and 10c, showing that indeed a higher value of λ produces larger changes in the opinions. Overall, however, the effect of varying λ is very limited.
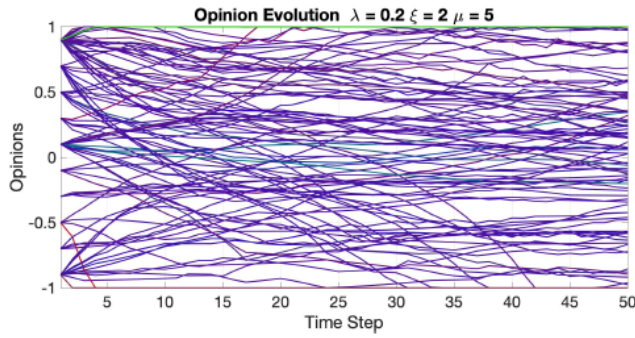
(a) Opinion Evolution
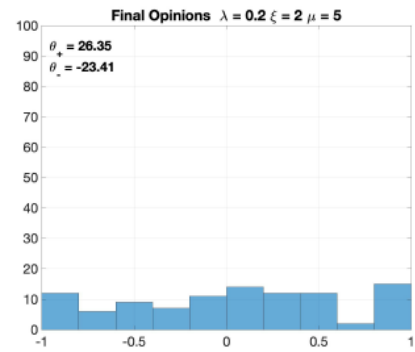
(b) Final Opinion Histogram
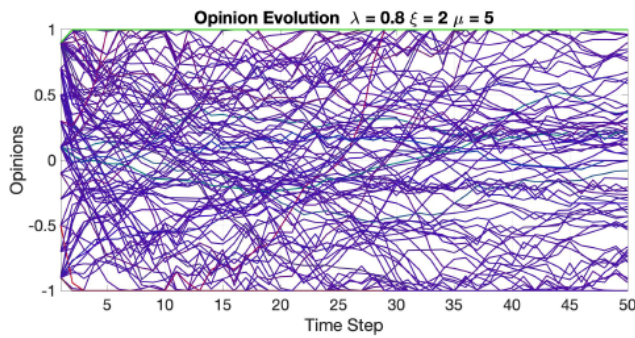
(c) Opinion Evolution

(d) Final Opinion Histogram

Figure 10: Effect of changing λ from the nominal value $\lambda = 0.4$ to $\lambda = 0.2$ (Figures 10a and 10b) and $\lambda = 0.8$ (Figures 10c and 10d) evolving 100 agents.

3.26 The effect of varying ξ is shown in Figure 11. The changes in both the opinion evolution and the final opinion histogram are quite noticeable. A value of $\xi = 1$ means that distant opinions have the same attracting power as closer opinions for the conformist traits, hence in general the conformist trait has less influence over the whole opinion change, which is instead dominated by the radical traits. The result is visible in the opinion evolution in Figure 11a and the final opinion histogram in Figure 11b. On the contrary, increasing the value to $\xi = 4$ yields a stronger conformist tendency towards consensus, evident when comparing the nominal final opinions in Figure 7b with the final opinions with $\xi = 4$ in Figure 11d, and the respective θ_- and θ_+.

(a) Opinion Evolution

(b) Final Opinion Histogram
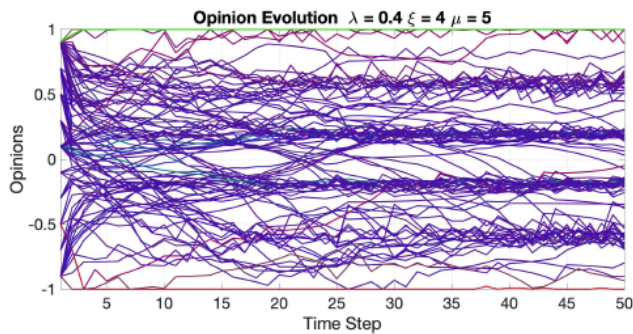
(c) Opinion Evolution
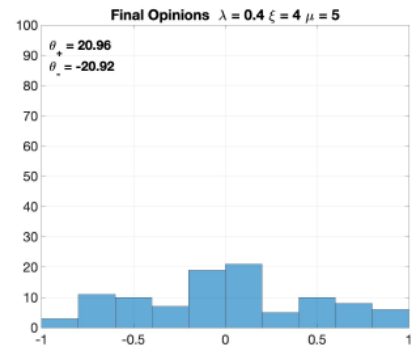
(d) Final Opinion Histogram

Figure 11: Effect of changing ξ from the nominal value $\xi = 2$ to $\xi = 1$ (Figures 11a and 11b) and $\xi = 0.8$ (Figures 11c and 11d) evolving 100 agents.

3.27 Parameter μ modulates the effect of radical traits on the opinion evolution. Comparing Figure 12b with Figure 12d shows that a larger μ increases radicalism in the population, which leads to polarisation for the given initial opinions. A similar effect is achieved by varying ξ: in fact, both ξ and μ affect the balance between the conformist tendency towards consensus and the radical tendency towards polarisation. Although both ξ and μ play a role in the conformist-radical balance, they are not completely complementary: an increase in ξ is not equivalent to a decrease in μ. This can be seen by comparing Figures 11d and 12b: increasing ξ produces final opinions that are more evenly distributed than those obtained by decreasing μ. Moreover, increasing radicalism does not always lead to polarisation: this happens only when the opinions have both positive and negative values. If the opinions have only positive values or only negative values, then radicalism will move all of them to a single extreme, resulting in consensus. Therefore, it is not possible to generalise the idea that more radicalism always leads to polarisation, regardless of the initial opinions.
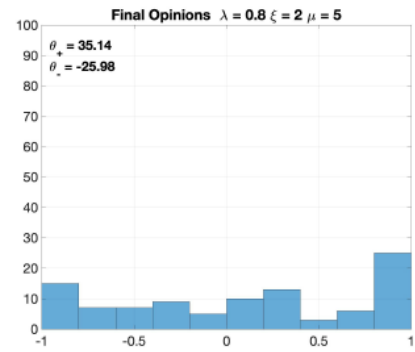
(a) Opinion Evolution

(b) Final Opinion Histogram
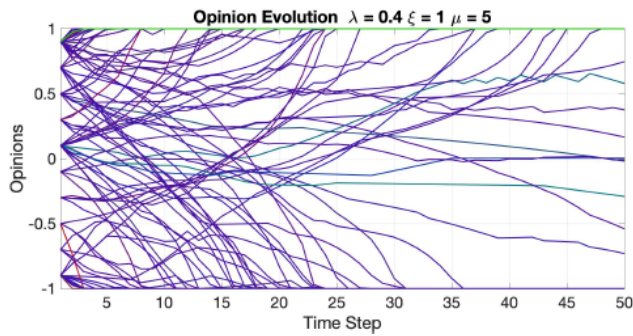

(c) Opinion Evolution

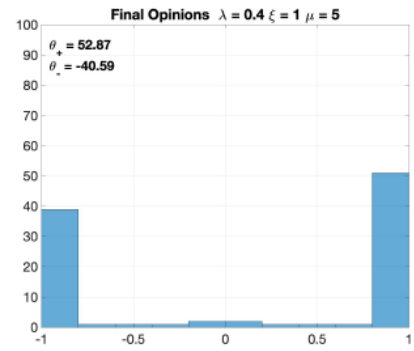(d) Final Opinion Histogram
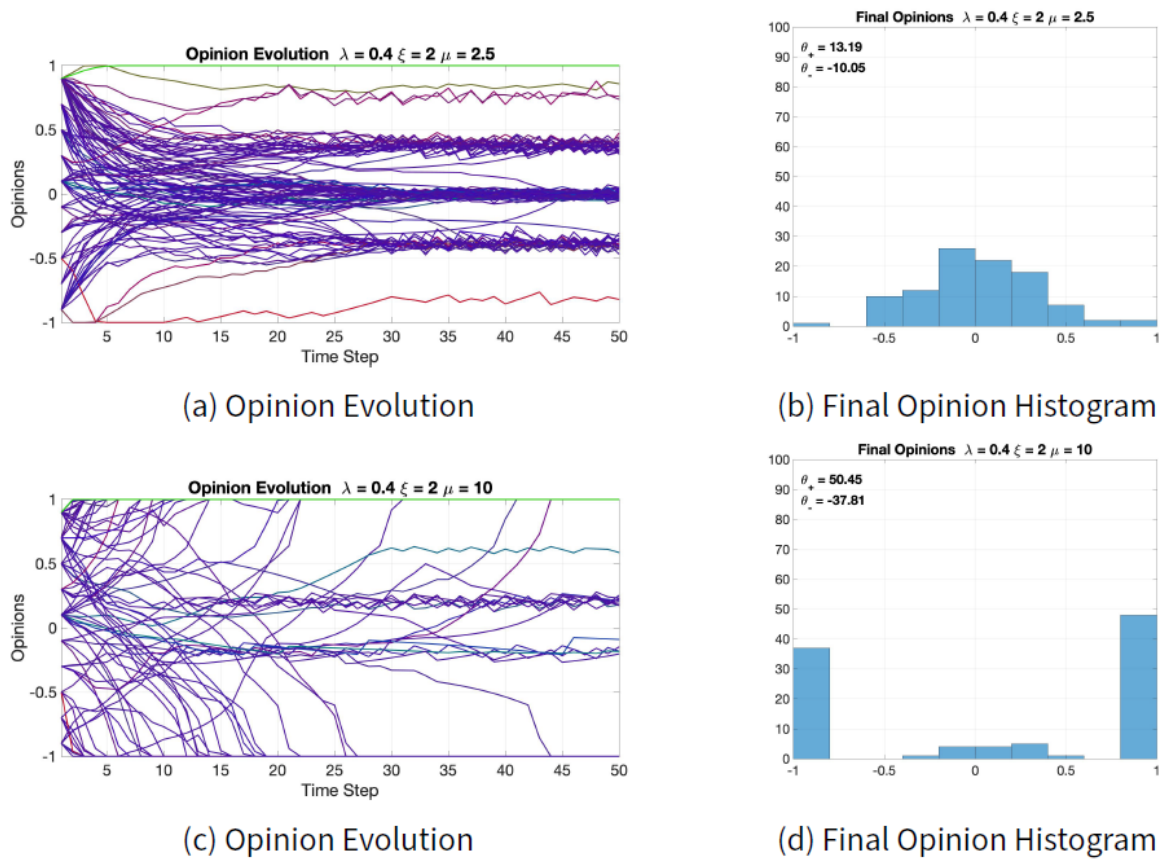
Figure 12: Effect of changing μ from the nominal value $\mu = 5$ to $\mu = 2.5$ (Figures 12a and 12b) and $\mu = 10$ (Figures 12c and 12d) evolving 100 agents.

## Characterisation of model outcomes

3.28 We characterise the model outcomes using different distributional measures: first, *Bias*, *Diversity*, and *Fragmentation* of the opinion distributions, recently proposed by Lorenz et al. (2021), to relate the inner trait assignations with qualitative properties of the model; then, plots of the *mean* and of the *mean of the absolute values* of the final opinion distributions that the model can generate.

Bias, *Diversity*, and *Fragmentation* analysis

3.29 Recently, Lorenz et al. (2021) have proposed *Bias*, *Diversity*, and *Fragmentation* as measures to analyse opinion formation models: *Bias* is the deviation of the mean opinion from the neutral opinion (indifference), *Diversity* is the normalised standard deviation of opinions, and *Fragmentation* measures how uneven the histogram of a given opinion distribution is. Their values range from 0 to 1; details are in Appendix F.

3.30 Figures 13a and 13e show the histograms of two different initial opinion distributions and their *Bias*, *Diversity*, *Fragmentation* values. In both cases, *Bias* is very low because the mean of opinions is almost zero. Figure 13a has a higher *Fragmentation* than Figure 13e, because Figure 13a shows two subgroups, whereas Figure 13e only one. Figure 13a has a higher *Diversity* than Figure 13e, because the two subgroups in Figure 13a are located at opposite sides.

3.31 Given a signed digraph W and inner trait assignations ψ, we evolve the model starting from the chosen initial opinions, and compute *Bias*, *Diversity*, and *Fragmentation* for the resulting final opinion distribution. To investigate how the inner traits affect *Bias*, *Diversity*, and *Fragmentation*, we evolve each of the initial opinions in Figure 13a and Figure 13e for 50 time steps, over the same constant signed digraph W (with Small-World network topology, see Figure 16a), with each of the inner trait assignations ψ in the set Υ, defined as:

$$\Upsilon = \left\{ \psi = (\psi_i)_{i \in V} = \left((\alpha, \beta, \gamma) \right)_{i=1}^{n} \mid \alpha, \beta, \gamma \in \{0, 0.05, 0.1, \dots, 0.9, 0.95, 1\} \quad \text{and} \quad \alpha + \beta + \gamma = 1 \right\}. \quad (14)$$

3.32 The set Υ contains 231 different inner trait assignations where all the agents have the same inner traits. Evolving the initial opinions for each of the inner trait assignations in Υ for 50 time steps results in 231 final opinions. For all these final opinions, *Bias*, *Diversity*, and *Fragmentation* are computed and shown in ternary diagrams as heat maps. Figures 13b, 13c, 13d (respectively, 13f, 13g, 13h) show the resulting ternary diagrams for the initial opinions in Figure 13a (resp. 13e).

(a) Initial opinion histogram.   (b) *Bias* plot.   (c) *Diversity* plot.   (d) *Fragmentation* plot.

(e) Initial opinion histogram.   (f) *Bias* plot.   (g) *Diversity* plot.   (h) *Fragmentation* plot.

Figure 13: *Bias*, *Diversity* and *Fragmentation* ternary plots for the initial opinions shown to the left. All opinions evolve for 50 time steps over the signed digraph shown in Figure 16a. All simulations are for 100 agents.

3.33 Interestingly, although the initial opinions are different, the resulting ternary diagrams are relatively similar. *Bias* is very low, except when the average radical trait is in the range 0.1-0.3 and the average stubborn trait is below 0.7; in fact, when the radical trait is high, agents tend to move to extremes and, if there is a comparable number of agents with positive and negative opinions (as in this case, for both initial opinions), the *Bias* will be near zero. On the other hand, when the conformist trait dominates, it moves opinions towards the mean, which in this case is also near zero. High *Bias* is achieved only when there is an appropriate combination of conformist and radical traits that moves agents from one half of the opinion interval to the other. *Diversity* and *Fragmentation* are almost 1 (the highest possible value) for medium to high levels of radicalism, because agents move to extremes.

3.34 Figure 14 shows simulation results analogous to the ones presented in Figure 13, but for an initial opinion distribution having higher *Bias*. In this case, because more agents have a positive opinion, the radical trait moves more agents towards the extreme opinion 1, resulting in turn in a higher *Bias* and lower *Diversity* for the final opinions (which, in many cases, are 1 for almost all agents). *Fragmentation* is still very high, since there is significant concentration of agents around opinion 1.

(a) Initial opinion histogram.   (b) *Bias* plot.   (c) *Diversity* plot.   (d) *Fragmentation* plot.
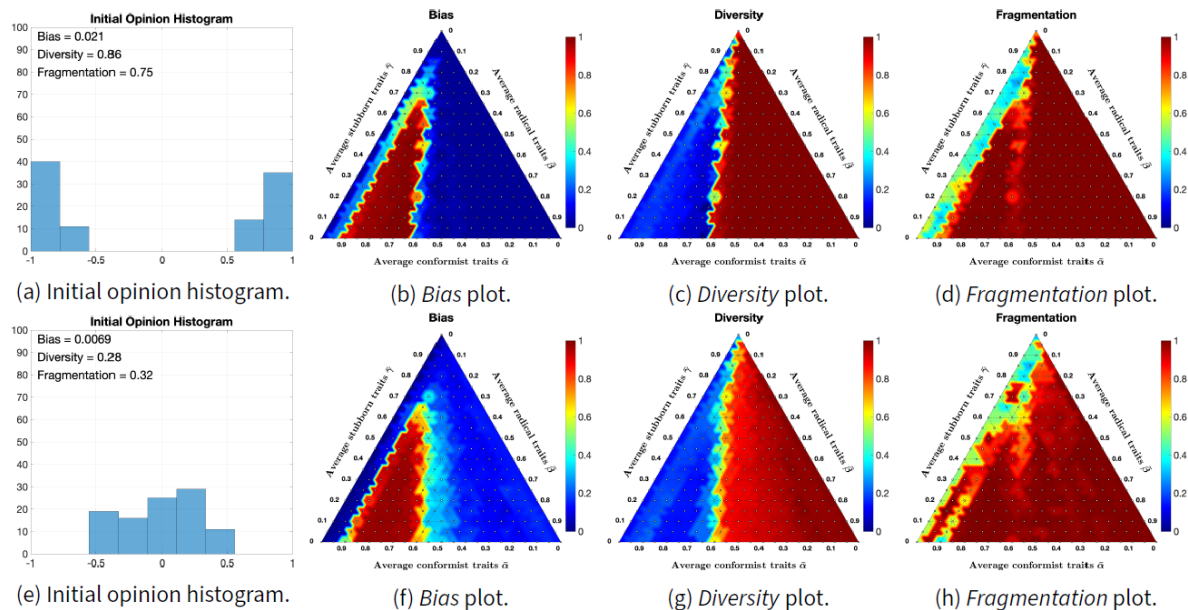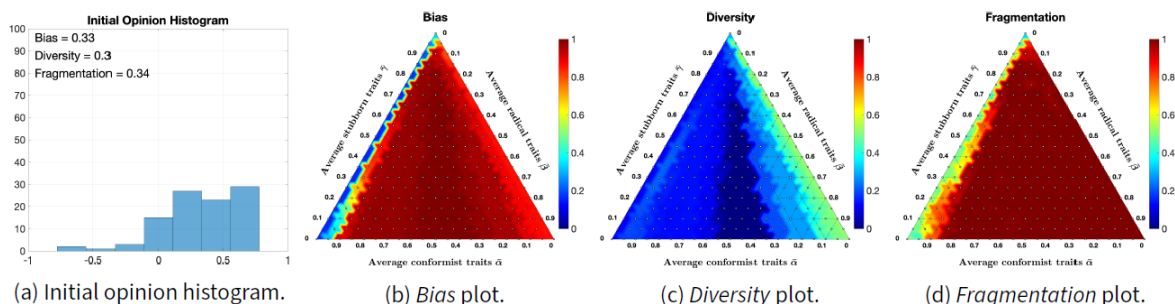
Figure 14: *Bias*, *Diversity* and *Fragmentation* ternary plots for the initial opinions shown to the left. All opinions evolve for 50 time steps over the signed digraph shown in Figure 16a. All simulations are for 100 agents.

3.35 The results in Figure 14b can also be visualised by plotting in a Cartesian plane the initial *Bias* on the x-axis and the final *Bias* on the y-axis for the 231 simulations. The result is a plot with 231 points of different colour. The colour of each point in the Cartesian plane is the RGB representation of the inner traits assignations (blue: conformist; red: radical; green: stubborn) associated with the corresponding simulation. Since all the simulations start from the same initial opinions (for instance, those shown in Figure 14a), all the points have the same x-coordinate. This allows for several *Bias* ternary plots to be represented in the same figure, providing an overall visualisation of how the final *Bias* relates to the initial *Bias* for different inner trait assignments. Such a plot is shown in Figure 15a for 231 different initial opinions. Analogous plots for *Diversity* and *Fragmentation* are reported in Figures 15b and 15c respectively.

3.36 Figure 15a shows that the CB model cannot produce opinions with a *Bias* below 1.7 unless the initial *Bias* is below 1.7; this may happen because *Bias* is low when the opinions are either all close to 0 or partitioned in comparable agreement/disagreement groups (see Table 12 in Appendix F), and the classification algorithm prevents opinions from converging very close to 0. Figure 15b shows that the CB model can produce low *Diversity* for almost any initial opinion, primarily due to the conformist trait (prevalence of blue dots) that pushes agents towards the same opinion, thus decreasing *Diversity*. Higher *Diversity* is mostly due to the radical trait (prevalence of red dots). Figure 15c shows that the CB model can produce final opinions with a wide range of *Fragmentation* values, which are however quantised because they are computed based on histogram counts (instead of on the exact agent opinions). Red dots are present, but covered by other dots, suggesting a high concentration of dots in some areas of this plot. The green diagonal lines in Figures 15a, 15b, 15c corresponds to agents that are all completely stubborn, and therefore do not change opinion.

(a) Initial vs. final *Bias* for different initial opinion distributions.

(b) Initial vs. final *Diversity* for different initial opinion distributions.

(c) Initial vs. final *Fragmentation* for different initial opinion distributions.

(d) Initial vs. mean final *Bias* for different initial opinion distributions.

(e) Initial vs. mean final *Diversity* for different initial opinion distributions.

(f) Initial vs. mean final *Fragmentation* for different initial opinion distributions.

Figure 15: Initial vs. (mean) final *Bias*, *Diversity*, and *Fragmentation*. In the first row, the colour of each point represents the inner trait assignation of the corresponding agents. All opinions evolve for 50 time steps over the signed digraph shown in Figure 16a. All simulations are for 100 agents.

3.37 When plotting the initial *Bias* and the *mean* final *Bias* (averaged over all the 231 elements in Υ), the 231 points with the same x-coordinate become a single point, as shown in Figure 15d. The same can be done for *Diversity* and *Fragmentation*, shown in Figures 15e and 15f. Collectively, these results provide further insight into the behaviour of the CB model. Figure 15d shows that the mean final *Bias* is always higher than the initial *Bias*. Figure 15e shows that there is an almost linear relation between the initial *Diversity* and the mean final *Diversity*.

Figure 15f shows that the mean final *Fragmentation* is always near 1, most likely due to the radical traits that move opinions to the extreme even if their weight is not particularly large (as seen in Figure 4).

Mean of opinions and opinion absolute values

3.38 The mean of the opinions, $\overline{x}$, and the mean of the opinions' absolute values, $\overline{|x|}$, are other interesting metrics. The mean of the opinions has a clear meaning; the mean of the opinions' absolute values represents the average level of interest the agents have in the considered statement. If $\overline{|x|} \approx 0$, then most opinions are near 0 (indifference); if $\overline{|x|} \approx 1$, then most opinions are extreme, either complete disagreement -1 or complete agreement 1 (high level of interest in the subject). The point $p(x) = (\overline{|x|}, \overline{x})$ in the Cartesian plane is located in the triangle with vertices $(0, 0)$, $(1, -1)$, $(1, 1)$. If $p(x)$ is near the origin, then most opinions are near zero; if $p(x)$ is located along the lines $y = \pm x$, then either all agents agree or all agents disagree with the same strength; if $p(x)$ is located near the line $x = 1$, then most agents have extreme opinions; and if $p(x)$ is located near the point $(1, 0)$, then the population is highly polarised.

3.39 Starting from an initial opinion distribution x_o, keeping the signed digraph W constant, we evolve the model for different inner trait assignations. For each of the resulting opinion distributions x_f, we compute the corresponding point $p(x_f)$ and we plot all these points in the Cartesian plane, to show which opinion distributions the model can produce starting from the initial opinion distribution x_o. The point colour encodes the average inner traits, to visualise how the inner traits affect the resulting opinion distributions. This type of plot is shown in Figure 16 for three different initial opinion distributions (shown in the crossed dot) that evolved over two different Small-World signed digraphs, for all the inner trait assignations in the set $\tilde{\mathcal{A}}$ described in Appendix C. Plots along the same row evolve over the same signed digraph shown to the left. Plots along the same column have the same initial opinion distribution, represented by the crossed dot.
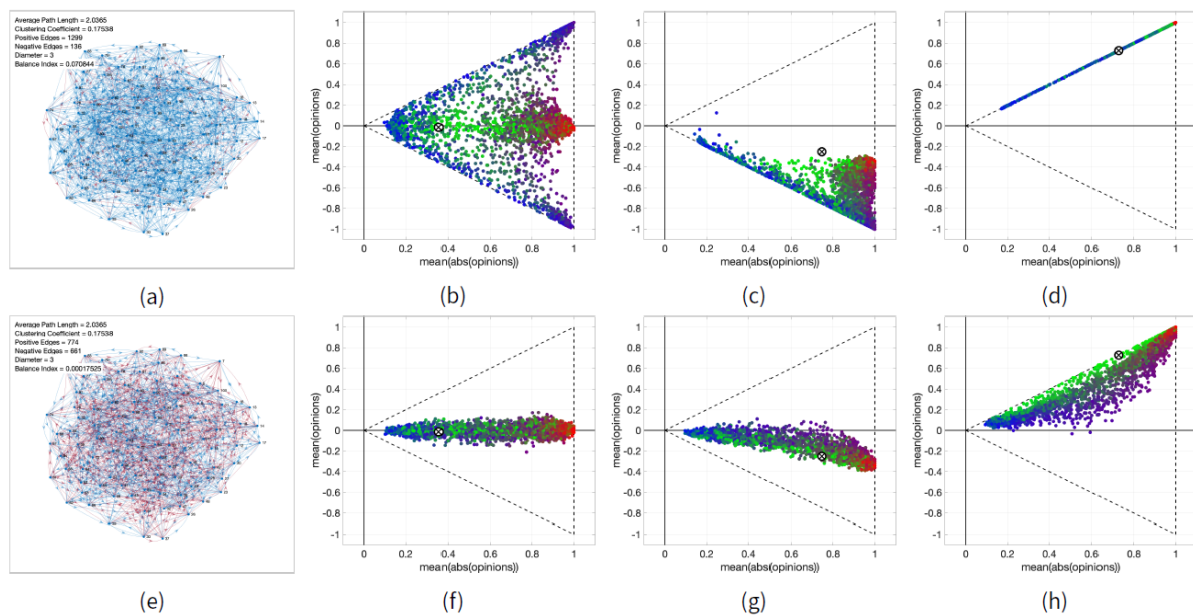
Figure 16: Plots of $\overline{x}$ and $\overline{|x|}$ starting from given initial opinion distributions (associated with the crossed dot) that are evolved, for 50 time steps, over the signed digraphs shown to the left, with inner agent parameters in the set $\tilde{\mathcal{A}}$ (described in Appendix F). The colour of each dot encodes the corresponding inner trait assignation. All simulations are for 100 agents.

3.40 The signed digraphs in Figures 16a and 16e have the same topology, but a different fraction of negative edges (9.47% in Figure 16a and 46.06% in Figure 16e). Comparing the evolution from the same initial opinions over the two different digraphs (Figures 16b vs. 16f; Figures 16c vs. 16g; Figures 16d vs. 16h) shows the significant effect the proportion of negative edges has on the system evolution. When most edges are positive, opinions reinforce themselves and the final opinions tends to move to extremes. When there are more negative edges, the digraph contains more unbalanced cycles (closed directed paths where the number of negative edges is odd) and therefore opinions tend to zero (as it happens in the Altafini model; Altafini 2013). The initial opinion

distribution also has a critical effect on the opinions that the model can produce: when the initial opinions have a significant non-zero mean, either positive or negative, the mean of the predicted opinions typically has the same sign (see Figures 16c, 16d, 16g, and 16h). Only initial opinions whose mean is near zero can produce an equal amount of final opinions with positive and negative mean (see Figures 16b and 16f).

3.41   Overall, the plots shown in Figure 16 indicate that the Classification-based model is very flexible and can produce a rich variety of final opinion distributions.

Model validation with real data

3.42   Data from the World Values Survey are used to validate the CB model, by showing that, with a suitable choice of the parameters, the model can produce opinion evolutions similar to those observed in real societies. The World Values Survey is an international organisation that conducts surveys about ethics and values in different countries around the globe. These surveys are repeated every 5 years. We considered the answers to 30 questions, shown in Table 14, in 26 countries, shown in Table 13, in Appendix G. In each question, the respondents are asked to state the extent to which they agree with a statement in a Likert-scale 10. The answers given in the surveys of wave 5 are taken as initial opinions, while the answers of wave 6 are taken as final opinions. Details on the solution of the optimisation problems are provided in Appendix E.

3.43   Two minimisation problems are stated to find model parameters that produce predicted opinions similar to the ones found in the survey answers. The ***Free* Optimisation Problem** allows the inner traits assignation to change with questions; in the ***Constrained* Optimisation Problem**, the inner traits are fixed for all questions.

3.44   Given real and model-generated opinion vectors r and y, for a population of n agents, the cost function J used in the minimisation problems (9), (10), and (11) is defined as

$$J(r,y) = \sum_{i=1}^{n} |\tilde{r}_i - \tilde{y}_i|, \tag{15}$$

where $\tilde{r} = (\tilde{r}_i)_{i=1}^{n}$ is the vector $\hat{r} = (\hat{r}_i)_{i=1}^{n}$ sorted in descending order, and $\hat{\cdot}$ is the quantisation function

$$\hat{r}_i = \arg\min_{\zeta \in \mathcal{R}}\{|\zeta - r_i|\} \qquad \forall i = 1, \ldots, n, \tag{16}$$

with $\mathcal{R}$ defined as $\mathcal{R} = \left\{ \frac{1}{2}(u_k + u_{k+1}) \mid u_k = -1 + k\frac{2}{10} \quad k = 1, \ldots, 9 \right\}$. Quantisation is needed because the World Values Survey answers we consider as real opinions use a Likert scale 10: participants could choose their opinion from 10 different options. These opinions rescaled to be between -1 and 1 produce the set $\mathcal{R}$ and, therefore, the predicted opinions also need to be quantified in the same way. Both opinion vectors (real and predicted) are sorted in descending order, so that equal opinions add a zero to the total cost.

3.45   Even for a relatively small population $n = 100$, the size of the sets $\mathcal{W}$ (underlying signed digraph structures) and $\mathcal{A}$ (inner traits assignations) is enormous. Given the tremendous size of the parameter space $\mathcal{W} \times \mathcal{A}$, performing the minimisation over all possible signed digraphs and agent inner traits would be computationally intractable. Therefore, the minimisation occurs over small subsets $\tilde{\mathcal{W}} \subset \mathcal{W}$, $\tilde{\mathcal{A}} \subset \mathcal{A}$ of the whole parameter space. As a consequence, there is no guarantee that we are estimating the *real* parameter values or making the absolute best parameter choice: with other parameter choices, not included in $\tilde{\mathcal{W}} \times \tilde{\mathcal{A}}$, the model could reproduce the data with even better accuracy. Appendix C describes in detail the sets $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{W}}$ used in our simulation results.

Free Optimisation Problem

3.46   Assuming that the agents can have different inner traits for each question, Equation (10) was used to find model parameters for each country that yield opinions similar to the real ones. Once the parameters that solve the minimisation problem (10) were found for each country, the cost associated with the prediction discrepancy for each question-country pair was computed as in Equation (15) (see Figure 17) and is shown in Table 1. Due to its complexity and the huge size of the feasibility set, the minimisation problem is solved approximately: hence, a possibly suboptimal solution is found. By solving the optimisation problem more accurately, over a longer computation time (which we could not afford, due to the very large number of question-country pairs that we consider), even smaller costs could be achieved, and hence even better fits of the real data.
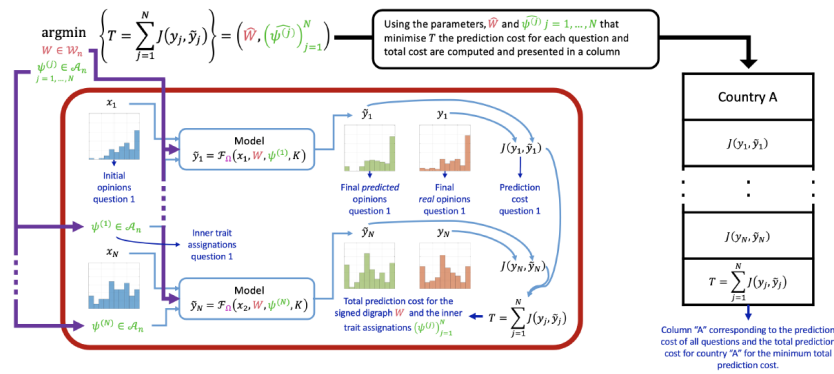
Figure 17: Visualisation of the procedure generating each column in Table 1 and of the minimisation problem in Equation (10). Assume that a survey conducted in two separate occasions in country A had Q questions. Given a signed digraph W and Q inner traits assignations $(\psi^{(l)})_{l=1}^{Q}$ for each question, the model predicts a final opinion $\tilde{y}_l$, for each question. The cost function $J(y_l, \tilde{y}_l)$ measures how close the predicted final opinion is to the real final opinion y_l. The sum of all these costs gives the total cost T. Minimising the value of T over the signed digraph W and inner traits assignations $(\psi^{(l)})_{l=1}^{Q}$ gives the parameters that best reproduce the society, $\widehat{W}$, and $(\widehat{\psi^{(l)}})_{l=1}^{Q}$. The cost for each question and the average and total cost obtained using these optimal parameters are reported in the column of Table 1 corresponding to the considered country. All simulations evolved 100 agents.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
Q1	4.2	3	3	3.6	3	2.8	2.6	2.6	2.4	2.4	1.6	3.2	3.2	2.2	2.8	2.2	3	2.4	2.8	3.6	3.2	2	3.2	3.6	2	3.6
Q2	5.2	0.8	2.4	3.6	4	4.4	4	3.6	2.2	2.8	2.2	2.2	4.2	2.6	3.4	3.4	3.6	4.4	2	2.2	2.4	2.4	2.6	2.8	6.2	3.4
Q3	4	3.8	3.4	3.2	2.4	4.6	2.4	4.8	2.6	5	2.6	4	4	6.4	4	3	3.2	3.8	2.8	2.8	2.2	2.4	2.2	2.4	4	3
Q4	5	2.4	3.2	5.2	4.6	9	6.4	6.6	2.4	4	6	5.2	6.4	6.2	2.2	2.6	4.4	6.4	4.6	22.6	3	4.8	3.4	39.6	4.4	3.8
Q5	3.4	2.8	2.6	9.6	3.8	4.4	3.2	9.8	6.4	2	4	2.6	3.6	11.4	3.4	3.6	3.6	4.6	2.6	3.8	4.6	2.6	3.8	3	2.2	3.8
Q6	4.4	2.2	2.2	2.2	2.4	2.2	4.6	6.8	3	1.8	3	2.6	2	4	3.2	1.8	1.6	3.4	4	12.4	4.8	3.4	2.6	3.2	1.6	5.2
Q7	1.4	2.4	3.2	3.2	5.2	2.4	3.4	4	2.8	4.8	2.8	2.6	3.6	4.8	3.8	2.8	2.8	2.8	2.2	2.4	1.8	2.2	4.8	2.2	3	5.8
Q8	4.2	2	2.6	3.8	4.2	6	4	5.2	2.2	3	3.2	2	3.4	3	3.8	2.2	3.8	1.8	5.4	3.2	2.4	2.6	3.2	2.4	2	3
Q9	3.6	3.4	2.8	4.2	2.4	2.6	2.8	2.6	3.8	4.6	2.8	3.6	3.4	4.2	2.2	4.4	3.6	3	2.8	2.8	3.4	2.4	3.8	2.6	3.8	1.6
Q10	3.2	2.4	2.8	3.8	3	2.4	2.4	2.2	5	3.2	2.4	3.4	2.4	4.6	3	7.8	4.2	3.8	3	3.6	3.8	2.2	2.4	2.4	3	2.4
Q11	3	2.4	5.2	2.6	4.2	3.4	3.4	2.6	6.8	3.2	4	3	4.2	4.2	3.4	4	2.8	3.6	2.6	3.6	5	6.6	3	4	2.4	3.4
Q12	4	2.6	5.4	5.8	2.6	13.4	5	3.2	3.8	3.6	2.4	4.2	3.2	1.4	2.8	3.8	2.2	2.6	2.2	2.2	2.8	2.2	2.6	3.8	2.2	2.8
Q13	6	0.8	2.2	4.6	1.8	1.8	1	5	0.8	4	1.6	1.8	4.2	2.4	4	2.4	3.2	3.2	2.8	39.8	1	2.8	0.8	3	4	7.6
Q14	0.8	1.4	2.8	3.8	1.6	1.8	4.8	1.8	2.8	2.4	2.2	3.2	2.6	1	3	3.8	1.6	1.8	2.2	4.2	2.8	2.8	2.8	4.2	1.8	2.4
Q15	1.4	2.2	1.8	3	2	1.4	1.8	3.2	3.6	1.2	2.8	1.4	1.8	2.4	2.4	1.4	2	1.4	2.6	1.8	3	1.4	2.4	1.6	1.8	
Q16	1.2	0.8	2.4	1.6	0.6	0.8	3.2	4.8	1.6	1.4	1.8	1.4	1	0.6	2.2	3	1.4	1.4	2.6	3.2	1.2	1.8	1	2.6	1.4	
Q17	3.2	4.2	4.4	3	2	0.4	1.8	1.8	1.2	3.6	2	2.6	3.2	1.8	3.8	3.2	2.2	2	2.6	2	0.8	4.2	2	2	8.4	2.6
Q18	4	1	2.2	4.6	3.2	2	2.2	4.4	1.4	3.8	2.4	2	3.4	2.8	3.4	3.6	2.4	3	3	2.4	1.4	2.8	2.2	4.2	2.8	
Q19	3.4	2.6	3.6	6	3.2	4	3.2	4.2	2.4	4.6	2.4	3.2	4	2.4	2.8	2.4	2	2.2	3.8	4.6	8.2	3.2	1.6	3.8	5	2.6
Q20	4	0.2	1.6	3.2	1.8	0.6	2.4	8.8	0.6	2.2	2.6	1.4	2.2	0.4	3.2	4.2	2	3.6	2	3	0.8	2.2	1.4	0.8	3.6	3
Q21	0.2	1.2	1	2	1	0.4	2.6	2.6	1.6	1.6	2.4	1	0.6	0.4	0.6	3.6	1.4	0.4	1.6	3.4	1.4	1.4	1.2	1.2	0.4	1
Q22	3.2	3.2	2.8	7.6	3.2	2.8	4.6	4.8	7.8	2.6	4.4	2.8	3	3	3.2	3.6	1.4	4	2.4	2.4	3.6	3.2	2.2	3.8	3	5.2
Q23	2	4.2	2.8	4	2.2	3.2	3.4	3.4	10.4	3.6	4.2	2.6	2.4	4	2.6	10	4.4	3.8	2.4	2.2	2.4	1.8	3.4	2.8	1.8	2.4
Q24	1.6	2.6	2.2	6.4	3.2	3.6	2.4	4.8	4	2.4	4	4	2	6	3.4	3.4	2.2	1.8	3.2	2.2	2	1.6	2.4	3.6	2.8	2.4
Q25	2.2	3.4	2.4	3.2	5	9	7	4	6	3	3	3	3.4	4.6	3	2.8	1	2.8	3	2.4	4.2	2.2	2	2.6	2.2	2.4
Q26	2.6	3.2	3	5.8	3	3.2	2.6	3.6	17.4	1.4	3.2	2.6	3.2	2.4	3	5.2	2.2	2.2	3.2	6.8	5.4	2.2	3.4	4.4	3.4	4.4
Q27	3	2.8	6.2	3	6.6	10.4	10.8	3.4	6.2	2.8	2	3.8	2.8	7.6	2.2	4.2	1	4.6	3	3	3.8	2.6	3.6	2	3.8	5
Q28	2.4	1.6	1.4	2.6	3.6	5.8	4.6	2.8	2.8	4	2.4	1.6	2	5.4	2.2	2.2	1.2	0.8	4.4	2.6	4.8	2.8	2.8	2.8	2.8	3.2
Q29	2.4	3.4	1.6	2.4	1.6	2.4	2.6	3	2.4	3.6	2.2	1.6	1.8	1.4	2.6	3.6	1.8	1.2	2.4	1.6	1.4	2.4	4.2	3.2	1.8	1.8
Q30	2.8	2.6	5.2	3.8	9	3.2	4.4	1.6	3.4	5	3	3.4	2	4	3.4	3.2	4.2	4.2	6.4	2.2	2.4	4.2	4.4	1.6	3.6	
Average	3.1	2.4	2.9	4	3.2	3.8	3.7	4.1	4	3.1	2.9	2.7	3	3.6	3	3.7	2.5	2.9	3	5.2	2.9	2.7	2.7	4.1	3	3.3
Total	92	71.6	88.4	121.4	96.4	114.4	109.6	122	119.8	93.6	85.6	82	89.2	107	89	109.8	76	88.2	88.8	154.6	87.2	82.4	79.8	122.2	90.8	97.8

Table 1: Results of the *Free* optimisation problem using the Classification-based model. Each column corresponds to a different country and each row to a different question. The cell values correspond to the optimal cost for all the countries and questions. The average cost along all the countries is 3.2815. The two final rows shows the column average and total. Cells with cost less than 7 are in green, the others are in red. Of the 780 possible question-country pairs, 755 have a cost less than 7 (an accuracy of 97% in total). The average cost of accurate (green) question-country pairs is 2.97.

3.47 Figure 18 shows the model predictions for some question-country pairs. The original opinion is shown in blue, the real final opinion in orange, and the predicted final opinion in green; the corresponding cost J (discrepancy) is reported. For costs less than 7, the model produces predicted final opinions that accurately represent the real final opinions. These cases correspond to green cells in Table 1, while cells with a cost higher than or equal to 7 are highlighted in red and constitute a small minority.

3.48 To carry out a thorough comparison with other opinion formation models, an analysis equivalent to the one reported in Table 1 is performed also for the Null (where the opinions do not change over time), the French-DeGroot (FG) (DeGroot 1974) and the Friedkin-Johnsen (FJ) (Friedkin 1986; Friedkin & Johnsen 1999) models. The results for the Null model are reported in Table 2; those for the FG model are analysed in Table 8 in Appendix A, and those for the FJ model in Appendix B. Comparing Table 1 with Table 2 shows that the CB model performs remarkably well, yielding a 97% accuracy in contrast to the 43% accuracy of the Null model: although there is a strong tendency towards stubbornness and opinion distributions tend to change only slightly over time, keeping the opinions exactly constant does not lead to good predictions.
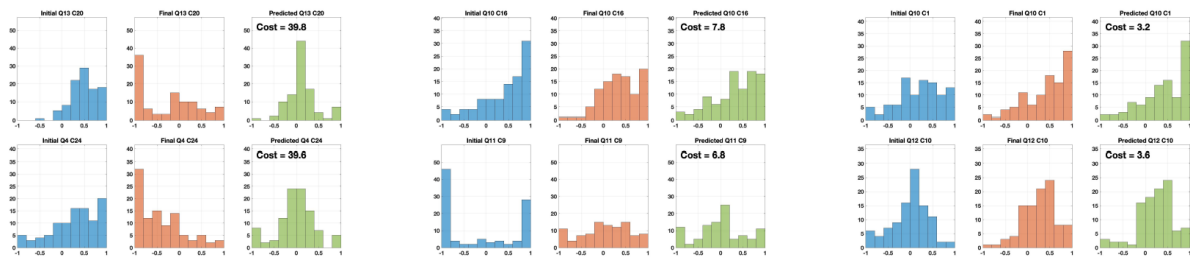
Figure 18: Examples of the opinion predictions achieved with the CB model, with varying resulting costs. For each trio of histograms, the initial opinions are in blue, the real final opinions in orange, and the predicted final opinions in green. The value of the cost J (in Equation (15)) is shown in the predicted histogram: cost values within 7 are shown to correspond to an accurate reproduction of the real opinion distribution.

Table 2: Results using the Null model, for comparison with Table 1. Out of 780 possible question-country pairs, 332 have a cost less than 7 (an accuracy of 43% in total). The average cost of accurate (green) question-country pairs is 4.17.

3.49 Plotting the average inner traits $\bar{\psi}$ for all question-country pairs for which the cost is less than 7 provides possible hints on how these societies could potentially be formed. However, because of the large parameter space and relatively small data set, we cannot make conclusive statements on actual societies just based on the optimisation results, as very different inner traits assignations may produce similarly low costs: we just propose a *possible* explanation. The resulting ternary diagram is presented in Figure 19. Figure 19a shows the position of each question-country pair. Figure 19b shows a density plot over the ternary diagram indicating the regions where most question-country pairs are found.

(a) Ternary diagram plotting the average inner traits for the question-country pairs with cost < 7 according to Table 1.

(b) Density plot of the ternary diagram: most question-country pairs have an average agent with more than 95% stubborn traits.
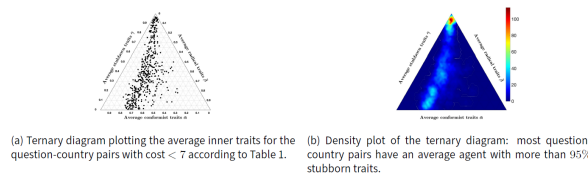
Figure 19: Analysis of the location of the average agents for all the question-country pairs with cost less than 7.

3.50 Despite the small data set and possible multiple local minima with similar low cost, fitting real data to gain an insight into the composition of actual societies reveals a clear trend: most average inner traits include a strong stubborn component, as shown by the high density in the stubbornness corner in Figure 19b. Also, the non-stubborn part can be roughly divided into 70% conformist and 30% radical, as shown by the trend in Figure 19a. This distribution is almost constant across all question-country pairs. Again, this is a possible explanation, and more data and more thorough explorations of the parameter space (extremely challenging from a computational standpoint) would be needed to make more conclusive statements. Hence, this is not conclusive evidence that most people are stubborn. There may be other explanations, for instance that not too many opinion exchange events take place in an average person's life. Graph-theoretically speaking, *isolation* due to

the lack of outgoing edges from a node (i.e., lack of interactions) is associated with the concept of *stubbornness*. However, from a mathematical model it is impossible to draw conclusions on whether the opinion of an agent remains unchanged because the agent refuses to consider the different opinions it is exposed to, or because the agent intentionally avoids exposure to different opinions, or because the agent simply lacks the opportunity to come into contact with different opinions. Furthermore, the traits themselves can be interpreted in different ways: for instance, a lower value of stubbornness can be regarded as a greater openness to change.

3.51 **Opinion Evolution Parameter Variation:** The results presented in Table 1 and Figures 18 and 19 are obtained by solving the minimisation problem (10) with *nominal* opinion evolution parameters $\lambda = 0.4, \xi = 2$, and $\mu = 5$. We now analyse the results of the minimisation problem when these parameters are changed. Tables 3 to 5 present how this variation affects the percentage of accurate question-country pairs (namely, those associated with a cost smaller than 7), the average cost of accurate question-country pairs, and the ternary diagram plot.
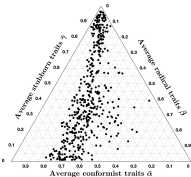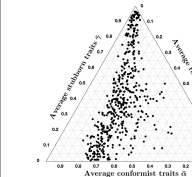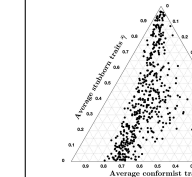
	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.8$
% of accurate country-question pairs	93.7	96.8	97.8
Average cost of accurate country-question pairs	2.79	2.97	3.02
Ternary Diagram Plot			

Table 3: Effects of varying λ while keeping the nominal values $\xi = 2$, and $\mu = 5$.
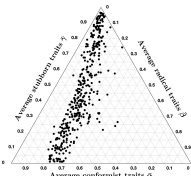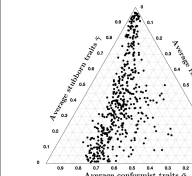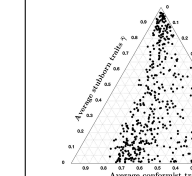
	$\xi = 1$	$\xi = 2$	$\xi = 4$
% of accurate country-question pairs	96	96.8	95.8
Average cost of accurate country-question pairs	2.84	2.97	3.45
Ternary Diagram Plot			

Table 4: Effects of varying ξ while keeping the nominal values $\mu = 5$, and $\lambda = 0.4$.
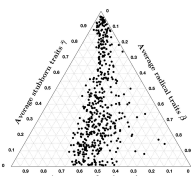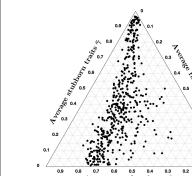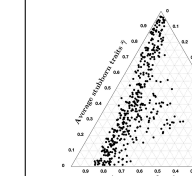
	$\mu = 2.5$	$\mu = 5$	$\mu = 10$
% of accurate country-question pairs	96.8	96.8	97.2
Average cost of accurate country-question pairs	2.84	2.97	3.15
Ternary Diagram Plot			

Table 5: Effects of varying μ while keeping the nominal values $\lambda = 0.4$, and $\xi = 2$.

3.52 Tables 3 to 5 show that, even after varying the values of λ, ξ, and μ, the percentage of accurate question-country pairs remains around 96%, and the average cost of accurate question-country pairs is between 2.79 and 3.45 which is quite remarkable since it means that the high accuracy achieved with the CB model is very robust to variations in the *opinion evolution parameters* λ, ξ, and μ (while it is not robust with respect to changes in signed digraph weights or inner trait assignments, see Figure 2).

3.53 Comparing the ternary diagrams shows the persistent tendency of question-country pairs to lie along a line where the proportion between conformist and radical traits is constant. For most simulation results, this proportion is still 70% conformist and 30% radical, as in the nominal case (Figure 19a). The proportion only changes when varying μ: for $\mu = 2.5$, we have 60% conformist and 40% radical agents, while for $\mu = 10$ we have 80% conformist and 20% radical agents. Therefore, it appears that μ can be tuned to regulate this proportion.

Constrained optimisation problem

3.54 If the agents are assumed to have the same inner traits for every question, then the model parameters can be found using the *constrained* optimisation problem in Equation (11). One advantage of using this approach is that, since each country has the same topology and inner traits assignation for all the questions, these parameters can be identified by solving the *constrained* optimisation problem (11) for a subset of all available questions (training dataset), and then tested on the remaining questions (test dataset). This was not possible previously, when assuming a different inner traits assignation associated with each question.

3.55 This procedure is commonly known as cross-validation. Generally, a subset of available data is used to train an algorithm (in this case, to identify the model parameters $\widehat{W}$ and $\widehat{\psi}$) and the remaining data is used to test the trained algorithm (in this case, the model with identified parameters $\widehat{W}$ and $\widehat{\psi}$). To eliminate result biases due to the selected training datasets and test datasets, cross-validation is performed multiple times for different partitions of the data. A common approach is to divide the data in K subsets and validate the model K times so that, at each iteration, only one subset is taken as the test dataset. This is known as K-fold cross-validation.

3.56 Table 6 shows the result of sixfold cross-validation on the available data (the questions are divided in six subsets of five questions each: $\{1, \ldots, 5\}, \{6, \ldots, 10\}, \ldots, \{26, \ldots, 30\}$). The first six rows show the mean cost for the five questions in the test dataset for each country for each cross-validation (CV1 to CV6). The last row shows the mean of the first six rows.
The simulation results summarised in Table 6 show that the model is able to accurately reproduce the final opinions for the tested data. Although the values are higher than 7, it is important to note that these predictions are done based on the assumption that the inner traits are the same for every question, while in reality the inner traits of the agents may change when considering their attitude towards different types of questions (which is taken into account by the free optimisation approach).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
CV1	7.4	8.5	12.6	7.4	13.8	21.3	18.5	13.9	24.7	9.6	10.4	8.2	6.3	10.4	11.4	11	7.8	9.4	9.1	13.6	12	7.3	11.2	9.9	5.4	9.4
CV2	6.3	8.4	9.3	11.4	6.6	16.8	15.6	19.6	14.3	12.6	9.7	7.6	8.6	15.5	5.7	10.7	7.8	7.5	6	18.5	16.9	8.6	7.5	7.1	6.3	8.6
CV3	8.5	10.7	10.5	10.6	7.7	12.5	20.1	34.6	14.7	10.4	14.5	10.3	8.8	8.1	11.9	12.6	10.9	8.3	10.2	16.3	10	8.8	12.9	6.8	10.4	12.7
CV4	10.1	10.6	11.6	5.7	10.9	19.1	10.6	19.7	20.4	13.6	13.9	9	12.8	9.4	11.5	14	12.1	7.7	7.7	26.8	14.6	11.7	9.8	18.1	7.6	10.9
CV5	9.7	6.8	9.9	6.6	20.5	7.7	10.1	13.6	7.9	8.8	8.4	8.2	8.9	7.5	11.4	13	8.1	15.2	7.5	15.2	8.2	5.3	10.7	7.8	5.8	15.2
CV6	11.2	9	14.9	13.7	20.3	11.4	11.4	22.5	14.9	16.6	18.2	8.7	11.9	21.1	11.2	9	16.2	16.2	6.8	19.8	8.6	9	12.4	20.3	10.9	9.4
Mean	8.9	9	11.5	9.2	13.3	14.8	14.4	20.7	16.1	11.9	12.5	8.7	9.5	12	10.5	11.7	10.5	10.7	7.9	18.4	11.7	8.4	10.8	11.7	7.7	11

Table 6: Results of the sixfold cross validation. Each column corresponds to a country, and each row to one of the six cross validations. The value in cell (i, j) is the average cost of the test data in cross validation i for country j. The last row represents the mean over all the rows.

3.57 Table 7 is analogous to Table 1, but now the model parameters are obtained with the *Constrained* optimisation problem (11), which yields a higher cost, as expected, since the optimised inner traits assignations can be very different when unconstrained, see Figure 19a.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
Q1	9.4	11.8	18.6	5.2	10.2	12.2	9.2	22	6	15.8	26.8	5.8	8.4	7.2	7	4.8	12.2	15.6	6.6	4.6	5.4	5.8	12.6	8	7.6	7
Q2	11.6	10.4	13	8	22.2	14.2	8	13.6	8.8	14	20.4	3.4	8.4	25	6.4	6.2	26.4	15.4	6	12.2	8	6.8	12.4	6.8	7.6	8.2
Q3	7.8	9.4	13	9.4	14.8	10.6	10	30.6	17.8	12.6	15.2	14.4	11	25.2	6.6	5.6	18.4	8	7.4	22.8	10.8	7.2	11.8	7	6.6	10.4
Q4	22.6	9.4	22.2	25.8	42.8	14.2	16.8	26	14.6	29.6	20.4	14.6	17.8	23	24.8	12.8	14.2	25.8	9.2	51.2	9.8	17.2	15	73	22.8	8.6
Q5	4.6	4.2	7.6	20	11.4	5.6	13.2	20.2	27.2	11.2	8.2	5.2	14	25.2	11.2	15.4	9.8	16.2	5	8.2	8.8	7.8	10.4	6.8	10	12.6
Q6	6.2	3.2	7.2	16	15	4.4	12.4	9.6	4.6	10	11.4	14	6.2	5.8	5.2	12.2	5	23.4	4	29.8	20	5.2	7.6	15.4	5	4.8
Q7	9.8	8.2	15	5.2	7.2	3	6	7.8	5.4	5.2	4.4	9.8	6.4	7.2	25	17.8	4.8	16.6	7	10.2	6	6.8	7.6	8.8	3.6	13.2
Q8	7.8	7.2	8	4.6	12.2	12.8	11.6	8.6	7.6	6.8	7	3.8	8.6	5.2	16.4	11.4	7.4	14.6	7.2	18.6	1.4	2.8	7.8	7.4	3.2	6.6
Q9	8.4	6.2	5.4	4.6	34.2	8.6	14.6	21.8	8.8	14.8	6.2	7.8	16.2	8	5.6	12.8	12.8	15.2	10.8	6	7.4	6	15	2.4	9.6	21.8
Q10	16.2	9.4	14	2.6	33.8	9.8	5.8	20.4	13	17.2	7	13	5.6	7.2	11.4	4.6	10.6	10.6	6.4	8.4	11.6	6	5.8	15.4	5.2	29.6
Q11	17.4	6.6	18.6	6.6	20.6	13	12.4	40.8	23.8	6.4	23	5.2	23.4	12	22.4	14.2	18.8	13.4	7.8	7	23.4	27.4	7.2	15.6	8.8	4.4
Q12	4.4	9.4	16.2	5.8	9.8	49.2	15	18.2	10.2	25.2	12.4	10.2	8.8	11	7.8	10	11.8	6.6	14.2	13	6	5.6	6.2	47	3	16.8
Q13	17	5.2	13.4	7.4	1.8	11.2	7.6	14	20.6	5.6	7	8.2	17.4	3.2	6.4	12.8	8.4	8	5.6	79.4	8	6	8.6	6.6	8.2	13.6
Q14	6.4	8.8	6.2	4.4	8.4	7.8	13.6	8.2	23.2	8.6	17.8	6.2	11.2	8.8	6.8	19.4	11.6	5.4	4.2	19.6	16.6	12.6	14	14.6	8.8	7.8
Q15	5.4	23.2	3.4	4.2	14	14.4	4.2	17.2	24.2	12.4	9.2	15.4	3	11.8	14	13.8	9.8	5.2	6.6	14.8	19	7	13.2	6.8	9	11.8
Q16	1.2	12	4.8	6.4	6.8	13.4	22.8	28.4	18	8.4	18.4	15.6	7.2	8.4	13.8	16.8	11	6.4	8.8	21	11	7.6	13.6	8.6	8.2	10.8
Q17	16.8	10	8.4	8.4	9.2	9	20.6	32.4	21.8	14	6	6	13.6	3.8	15	10.4	16.6	13	14	13.2	8.8	13.2	12.8	6.8	17.8	16
Q18	10	7.8	9.4	13	7	20.6	20.6	29.6	10.2	9	20	8.4	6.4	9	13.4	11.2	6	6	8	18.6	6.6	5.6	11	7.8	9	13.8
Q19	5.2	17.2	14.8	16.2	9.4	8.4	15.2	48.4	8.4	6.4	12.6	11	10.4	7.4	7.6	8.2	10	6.2	7.8	9	11.2	10	11.6	5.4	9.4	8.2
Q20	9.4	6.6	15	8.8	6.2	11.2	21.2	34.2	15	14	15.6	10.4	6.4	11.8	9.6	16.6	10.8	10	9.8	18.6	12.4	7.4	15.4	6.4	7.4	14.6
Q21	8	6.2	9.6	6.4	8	9.4	12	4.8	15.8	9.2	10	9	6.6	4	4.6	13.4	13.6	5.8	9.8	18.6	11.2	9.2	13.2	7.8	11.2	13.2
Q22	4.8	11	12	8.2	8.8	17.4	26.8	21.4	6.6	17.6	11.4	10.2	13.6	16	7.2	5.2	6	22.8	10.4	10.2	6.8	9.2	24.8	32.4	10.4	11.6
Q23	7.6	12.6	11.8	19	4.2	21.4	10	26.4	25.8	18.2	11	6.2	10.8	19.2	7	22.8	10.4	10.2	6.8	9.2	17.6	12.6	5	10.6	4.2	6.8
Q24	7.4	3.2	8.8	20.4	3.8	10.8	5.8	19.8	17	8	3.4	8.4	4.2	17	3.2	4.2	5.4	7.4	5.4	17.2	3.6	4.6	5.2	5.4	6	4.4
Q25	3.6	9	4.4	3.2	8.4	25	23.4	25.8	6.4	9.8	12.6	4.4	7.8	17.2	6.6	7.8	3.6	7.2	5.8	22.8	19.6	6	5.6	5.4	5	7
Q26	6.2	7.2	7.8	21.6	8.2	6	9.4	7.4	34.6	10	6.6	10.4	6.4	2.4	3.4	25.4	8.8	14	6.8	23.8	20.8	14.8	12.6	16.6	5.6	5
Q27	5.6	6.8	16.4	3	18.6	43.2	32.6	22	30.2	7.6	8.8	12	6.4	16	8.8	5.8	3.6	14	10.2	16.6	10.4	4.6	15.2	3.8	6.2	18
Q28	6.6	3.2	7.2	2.6	7.2	23.8	11	6	23.2	12	7.8	6.4	6.8	8.4	7.4	7	5.8	7	11	3.2	12.4	3.4	9.2	4.8	12	9.6
Q29	4.2	8.4	6	2.4	1.8	10.6	11.2	14.4	15.2	6.4	12.2	3	2.8	4.2	6.8	10.8	2.6	6	2.6	7.2	5.6	6.2	10.2	12.4	2	4.2
Q30	14.4	17	25.6	7.6	33.4	22.8	28.4	19.6	20.4	11.8	16.8	9.4	9	20	30.8	6.2	18.2	6	14.8	17.4	10.6	7.4	8.8	12	9.6	13.4
Average	8.9	9	11.5	9.2	13.3	14.8	14.4	20.7	16.1	11.9	12.5	8.7	9.5	12	10.5	11.7	10.5	10.7	7.9	18.4	11.7	8.4	10.8	11.7	7.7	11
Total	266	270.8	343.8	277	399.4	444	431.4	619.6	484.4	357.4	375.6	260.4	286.4	359.6	315.4	351.6	314.4	322	236.6	551.4	350.8	253	322.6	350	231.8	330.8

Table 7: Results of the *Constrained* optimisation problem using the Classification-based model. The average cost along all the countries is 11.6746. Out of 780 possible question-country pairs, 220 have a cost less than 7 (an accuracy of 28% in total). The average cost of accurate (green) question-country pairs is 5.16.

● Summary and Conclusions

4.1 We have proposed a novel agent-based opinion formation model that has two fundamental distinctive features. First, the model drops the unrealistic assumption that agents can measure the opinion of their neighbours with infinite precision, which drastically affects the opinion evolution, and introduces a novel classification-based approach that more realistically replicates the way individuals assess and evaluate the opinions of their neighbours, by classifying them as agreeing much less, less, comparably, more or much more. Second, the model captures the complexity of the behaviour of individuals by introducing three different internal traits, associated with conformism, radicalism, and stubbornness. Instead of considering agents of different types, the model allows all these tendencies to coexist in each agent, thus representing multifaceted psychological and socio-logical phenomena in action within each individual.

4.2 In addition to the agent parameters and the underlying digraph, the model simply relies on three parameters, λ, ξ, and μ, having a natural interpretation. Based on a deterministic classification mechanism, the opinions evolve over discrete time steps according to the deterministic Equation 6. The signed underlying digraph is time-invariant. Despite its simplicity, the model can recreate opinion evolutions seen in real-life and produce a rich and wide variety of collective behaviours, without the need of introducing bounded confidence, random-ness, or more complex mechanics.

4.3 Four types of simulation analyses were carried out: (i) simulations over simple digraph and agent parameters to gain insight into the model behaviour; (ii) simulations with varying model parameters to perform a param-eter sensitivity analysis; (iii) model outcome capabilities, studied using distributional measures such as the recently proposed *Bias*, *Diversity*, and *Fragmentation* (Lorenz et al. 2021); (iv) simulations with parameters cho-sen through the approximate solution of two optimisation problems to assess the model's potential to recreate opinions similar to those seen in real life.

4.4 We used real data from the World Values Survey to assess the capability of our Classification-based model to mimic actual opinion evolutions seen in real life: building a link between theoretical opinion dynamics mod-elling and empirical data in social research is a strong focus of this work.

4.5 Our results can be relevant in future research on opinion formation models in several ways. The proposed CB model offers a flexible general framework that can be easily adapted to combine other psychological traits, change the mathematical formulation of the current traits, or include agents with more accurate opinion per-ception (by increasing the number of sets an agent can classify its neighbours in); additional analyses can be performed to assess how these variations would affect the model behaviour and characteristics.

4.6 The link between survey results and model outcomes, with parameters chosen via suitable optimisation prob-lems, is also of value to future research on opinion formation models: the field has great potential to grow by systematically connecting empirical and theoretical findings. Our proposed methodology can be tailored to the available data and the focus of the study: possible changes include embedding constraints that correlate the agent opinions, their parameters, and their location in the network, and a combination of the 'free' and 'constrained' problems.

4.7 Our analyses and simulations have highlighted the model properties and behaviour, thus answering the follow-ing questions.

- What does the model actually do? Quantitatively, the model evolves by iterating Equation 6. Qualitatively, the combination of the classification mechanism and psychological traits results in agents having two main behaviours: mostly conformist agents aim for consensus with their neighbours, while mostly radical agents move toward extreme opinions. The stubborn trait slows the opinion evolution, without affecting its trend. The collective population dynamics is a combination of these effects, which makes the model particularly flexible and enables a wide range of opinion outcomes: the model can generate opinion distributions ranging from extreme polarisation to consensus, depending on the agent parameters. Due to the imperfect opinion perception, the agents rarely converge to the exact same non-extreme opinion, even if they all are completely conformist.

- What does the model teach us about human behaviour? The model provides some interesting indications. The ability of the Classification-based model to recreate opinion transitions seen in real populations suggests that radicalism is an essential trait in the opinion formation mechanism, when modelling large-scale opinion evolution. Additionally, the model also shows that radicalism alone does not necessarily lead to polarisation: polarisation only occurs when the population is mostly radical *and* the initial opinions are mostly divided (comparably the same number of agents agree and disagree).

- What kind of dynamics does the model produce and why does it produce these dynamics? Since the stubbornness trait mainly slows the opinion evolution, let us focus on the other two traits. Completely conformist societies move towards the mean of the initial opinions (without reaching consensus on a single opinion, but asymptotically driving all opinions in the neighbourhood of a given opinion). As radicalism increases, some agents move to extreme opinions and influence other agents, resulting in a more diverse or partitioned set of opinions (possibly forming clusters). If radicalism keeps increasing, at some point all the agents move to extreme opinions, resulting in either polarisation, or consensus at either complete agreement or complete disagreement. The sign of the edges of the signed digraph has a more subtle effect. Increasing the fraction of negative edges makes the dynamics less obvious to predict, especially when most agents are conformist: if the digraph is sufficiently unbalanced, the opinions move towards 0 (as it happens with a structurally unbalanced network in the model by Altafini (2013)); if the signs are suitably arranged, opinions may also converge to values that are not near the initial opinion mean.

- Are any of the results surprising in some way? The significant effect of radicalism on the opinion evolution in a large population is remarkable. The model shows that radicalism is more impactful than conformism (see for instance the simulations with only radical and conformist traits in Figure 4); this effect may be due to the mathematical implementation of the radical trait. It is also surprising that the sign of the digraph edges has such a subtle effect, while for instance in the Altafini model its impact on the opinion evolution is much more noticeable.

4.8 A particularly interesting direction for future work is the formulation and solution of the optimisation problems that include correlation constraints between initial opinions, agent parameters, and location in the network. An optimisation problem that is located between the 'constrained' and 'free' problems would also be interesting to define: the agent parameters for each question may be allowed to be different, but within a maximum possible difference, to represent the case in which the agent traits can change depending on the question, but cannot be completely different because each agent preserves some main personality traits across questions. This 'hybrid' optimisation problem could then be solved using heuristic approaches, such as genetic algorithms. This could be a possible way to advance in the solution of the inverse problem for opinion formation models, and as such would be a considerable progress in the direction of studying opinion dynamics with empirical data. Additional research directions are to study of the effects on the opinion evolution of different network topologies, opinion-dependent agent parameters, and the effect of opinion evolution parameters Ω that are agent-dependent.

Appendix A: Comparison with the French-DeGroot model

Here, an optimisation problem analogous to the *Constrained optimisation problem* for our CB model is solved for the classical French-DeGroot (FG) model. It is important to note that the two models aim at reproducing opinion evolution dynamics in completely different contexts: the FG model was developed for small-group interactions over a relatively short time interval, usually leading to consensus; the CB model aims to recreate opinion changes in large-scale societies over long time intervals, as indicated by the use of country-wide survey results with approximately 5 year separation. Still, it is interesting to see how the results differ.

To make Tables 1 and 8 comparable, for the FG model the digraphs used in each country are selected following the same minimisation problem as the one solved for the CB model. Since the FG model does not involve agent parameters, we only minimise over the set of digraphs $\mathcal{W}_{\text{FG}}$. Both the set of digraphs for the CB model, $\mathcal{W}$, and for the FG model, $\mathcal{W}_{\text{FG}}$, have the same number of elements and there is a one-to-one topology correspondence; the digraphs in $\mathcal{W}$ are signed and unweighted, while those in $\mathcal{W}_{\text{FG}}$ are unsigned and row-stochastic (as required by the different nature of the two models). As expected, the FG model, focused on small-group consensus processes, does not lead to accurate predictions of opinion changes in large societies for long time intervals.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
Q1	28.6	41.2	32.4	32	35	37.6	32.6	33.6	34.6	31.6	26.8	32	33.6	36	36.2	33	28.8	27.4	30.6	36.8	32	31.8	33	39.2	25.8	34.6
Q2	32.2	34.8	28.6	30.6	34	37.6	39.2	38.4	33.2	31.2	28	32.4	31.8	41.2	31.8	36.6	26.6	23.4	32.8	32.8	35	36.4	32.4	40.4	30.2	28.8
Q3	38.6	45	32.6	33.4	39.2	36.6	43	41.2	38.8	37.2	33.6	42.8	39.2	46	37.6	39.2	33	38.2	32.2	36.2	40.8	36.4	32.6	41.2	41.2	39.6
Q4	49.4	62.4	51.4	55	53.2	48	43	57.2	38	35.8	42.6	58.6	48.4	64.4	43.6	43.8	43.6	45.8	36.2	58	46.2	44.4	47.4	78	42.8	48.6
Q5	38.2	56.8	40.6	46	40.8	48.8	52.4	51.2	49.4	24.8	39.6	54.2	49.8	61	52.6	45.2	31.4	33.8	35	46.2	46.4	35.2	45	50.2	38.6	40.4
Q6	47.4	54.8	43.8	46	47.6	34.4	50.2	53.2	46	37.6	41.8	57	46.8	60.4	41.8	47.2	37	40	36.2	52.2	56.6	44.6	42.8	48.2	48	50.2
Q7	35.8	52.8	41.8	31.6	45.8	35.4	34.6	39.8	37	15.4	40.8	49.4	46	46	36	48.6	31.4	32.6	30.6	43.6	41.6	34.6	42.6	43.8	33.6	47.6
Q8	40.6	60.8	40.6	40.6	52	46.8	36.8	45.4	51.4	39.6	41.8	48.8	48.6	48.2	42.6	49.6	38	34.2	42	42.8	41.2	43	44.4	51	41	47.4
Q9	41.4	48	40.2	25.4	51.8	35.4	29	33.4	37.2	30	36.4	44	45.4	42.2	38.6	33.4	35.2	37.4	32.4	37.8	46.8	34.4	32.8	40.6	30.4	43.6
Q10	41.6	46	40.4	26.8	53.6	33.4	28.4	35	35.4	31.2	36.6	46.6	33.2	39.8	37.2	37.2	30.2	37	38.8	34.2	36.6	35.6	35	34.4	30.4	46
Q11	52.8	51.2	40.8	43.8	50.4	43.4	39.8	40.8	45.6	38	39.6	49	51.8	51.6	51.6	33.2	40	49.6	34.2	40.6	45.8	41	44.4	43	47	46.4
Q12	37.8	52.2	40.4	23.2	37	48.8	38.6	38.4	40.8	33.2	39.2	54.6	39.6	44	39	33.6	33.4	36	29	37.4	41.6	34	27.2	46.4	34.8	39.8
Q13	68.2	8.8	33.4	43.4	30.2	22.8	7.4	33.8	6.6	45	36.8	25.4	38.8	26.4	50.6	34	50.4	54	50	61.6	7.6	41.2	23.4	46.2	47	54.2
Q14	25.8	42.4	50.2	40.8	24.2	22.2	22.4	45.4	39.6	26	47	58	33.8	27.6	36	59.2	26.2	27.2	32.4	52.6	38.2	38.6	25.6	37.2	29.4	33.4
Q15	22.6	56.8	21.4	25.6	20.6	21.4	12.2	39.2	22.6	7.6	40.4	32.2	29	32.6	22	54.2	21.8	26	21.8	36.6	21.8	25.4	7.2	36.2	25.8	21.2
Q16	22.6	23.4	24	24.6	22.2	4.2	23.8	38.2	10	6.4	41.4	32.8	11.8	6.6	22.4	56.8	23.6	25	20	35.8	22.6	14.2	6	25.6	24.6	24.4
Q17	58.6	59.8	49	28.4	51.4	5.6	22	39.6	5.2	52	39.6	56	49.6	34.2	69	53.6	47.6	47.2	43.6	39	26.6	50	26.6	37	57	58.4
Q18	49.6	33.4	43.6	35.4	53.4	37	23.6	40.2	18.4	18.6	38.6	38.4	50.6	43	47.4	54	53	50.4	42.4	40.4	23.8	31	38.4	32.2	48.2	58
Q19	43	59.6	52.8	46	52.2	41.8	42	53	39	44.4	47.2	59.4	41.6	56	47	50.6	44.6	32.4	41.6	44.4	37.6	47.6	42.8	41	55	
Q20	46.6	24.8	26	31.8	32.6	5.6	22.8	39	12.6	38.6	37.4	44.4	32	7.2	55.6	56	42.2	53.8	37.4	20.6	23.4	30.6	9	27	43	42.6
Q21	4.4	22	5.2	28.8	26.2	3.4	29.6	38.6	11.2	22.4	37.4	28.8	7.2	23.8	11.8	53.8	5	6.4	21.2	20.4	20	22	8.8	23.8	8.6	22.8
Q22	46.2	58.8	47	39.2	60.6	52.8	56.6	49.6	41.2	37	47.2	60.8	45.8	58.6	49.2	44.2	36.2	44.8	36.8	54.8	57	38	44.6	46.6	42.8	52.8
Q23	38.6	54	47.8	47.4	42.6	40.6	47.6	52.2	51	37.2	49.8	52.8	45	58.4	34.6	41.8	48.4	28.2	34.6	49.8	47	32.8	52.2	44.8	38.8	44.6
Q24	29.4	43.2	28.2	37.4	34.2	31.8	28.6	41.4	40.6	35.4	44.4	49.2	28.4	35	28.6	39.4	29	14.2	35.4	43.2	41.6	26	32.4	32.4	34	30.2
Q25	48.8	48.6	41.6	29.4	44.8	44.6	51.4	45	46.2	39	46	57.8	42.8	40.8	43.2	44.6	30.6	42.8	35.6	38	52.6	33.2	35.4	46.2	44.2	46.4
Q26	45.6	60	45.8	49.2	47.8	47	42.6	58.2	53.2	26.6	47.6	62	46.4	64.8	41.6	47.6	40.8	40	41.6	53.2	63.8	36	59	50	49.8	48
Q27	44.6	49.6	44.2	26.8	48.8	49	51.2	46	51	40.4	43.6	50.6	31.8	33.8	37.4	42.2	32	31.6	39.2	49.2	41.6	31	39.4	35.6	41.6	41.8
Q28	28.6	44.6	30.6	25.2	36	36.2	37	42	47.8	41	42.2	44	30.8	32.6	26.8	40	27.2	42.8	35.6	38	52.6	33.2	36.4	29	33	32
Q29	25.6	42.6	35.4	24.6	24.4	29.6	25	37.4	32.2	31	30.8	31.8	27.4	29.2	35.2	37.8	25.4	15.2	25.4	35	28.8	25.2	24.8	36.4	28.6	24.2
Q30	36	46	39.2	31	46	34.8	34.4	40.8	33.2	39.2	32	48	44.4	35	35.2	28.2	32.1	54.2	38.2	34.4	29	33	36.4	39.4	40.4	36.8
Average	39	46.1	38	35	41.3	33.9	34.9	42.7	35.2	33.4	39.8	47.2	37.8	41.5	39.3	44	34.3	33.5	34.5	41.7	38.4	34.5	33.5	41	37.4	41.3
Total	1170.8	1384.4	1139	1049.4	1238.6	1016.6	1048.2	1280.8	1056.2	1001.2	1194.2	1416.6	1134.6	1243.8	1178.4	1320.4	1029.4	1005.4	1034	1251.4	1153.4	1036.4	1005.6	1229.4	1121.2	1237.8

Table 8: Results of the optimisation problem using the French-DeGroot model, for comparison with Table 1. The average cost along all the countries is 38.4323. Of the 780 possible question-country pairs, 13 have a cost less than 7 (an accuracy of 2% in total). The average cost of accurate (green) question-country pairs is 5.43.

Appendix B: Comparison with the Friedkin-Johnsen Model

The Classification-based model can be seen as an extension of the Friedkin-Johnsen (FJ) model (Friedkin 1986; Friedkin & Johnsen 1999), in the sense that both models include in the agents' behaviour inner traits described by tuples. In the FJ model, each agent i is characterised by two parameters: *susceptibility* $a_i \in [0,1]$, determining how strongly the agent is affected by its neighbours' opinions and forgets the initial opinion (Friedkin & Johnsen 1999), and *prejudice* $b_i = 1 - a_i \in [0,1]$. A value of $a_i = 1$ means that the agent has complete susceptibility to interpersonal influence (similar to complete conformism), while a value of $a_i = 0$ means that the opinion remains the same for all times (similar to complete stubbornness). In the CB model, each agent i is associated with three parameters: *conformism* $\alpha_i \in [0,1]$, *radicalism* $\beta_i \in [0,1]$, and *stubbornness* $\gamma_i \in [0,1]$, such that $\alpha_i + \beta_i + \gamma_i = 1$ for all $i \in V$. Therefore, a FJ model where all agents have a susceptibility of a is similar to a CB model where all agents have inner traits $\alpha = a$ and $\gamma = 1 - a$ (hence $\beta = 0$). Still, the interpretations of stubbornness and prejudice are slightly different: prejudice in the FJ model means that agents tend to keep their *initial opinion*, while stubbornness in the CB model means that agents tend to keep their *current opinion*, which leads to the same outcome only when all the agents are completely stubborn. Apart from the outlined similarity, the FJ and CB models are different: crucially, the FJ model is linear, while the CB model is highly non-linear, which severely limits the applicability of closed-form analysis tools.

Figure 20 shows the evolution of the same initial opinions according to the two models, for different values of a, α, and γ. The digraphs have the same topology; randomly generated weights are considered for the FJ model, while for the CB model all the edge signs are taken positive to match the absence of antagonism in the FJ model. The FJ model exhibits a slower change as a increases, while with the CB model, as soon as conformism is introduced, the opinions converge to an interval where all the agents perceive that their neighbours' opinion is similar enough to theirs. This difference is caused by the two different interpretations and implementations of stubbornness and prejudice. Another important difference is that, as the susceptibility value increases, the final opinions of the FJ model tend to converge to a single opinion, and yield perfect consensus when $a = 1$. Conversely, the final opinions of the CB model never converge to perfect consensus, even when $\alpha = 1$ and $\gamma = 0$, as a consequence of the classification-based approach.
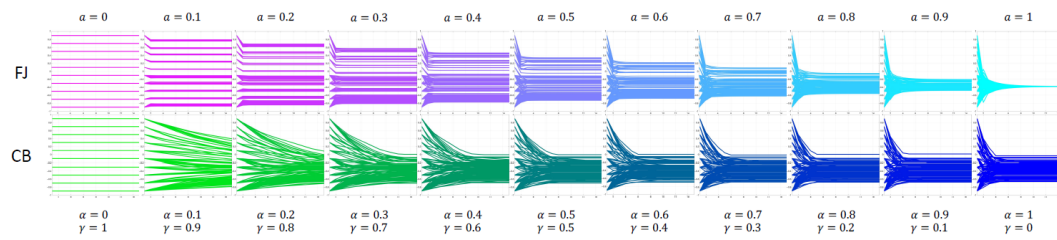
Figure 20: Comparison between the Friedkin-Johnsen (FJ) model, for different values of susceptibility a, and the Classification-based (CB) model, for corresponding values of conformist (α) and stubborn (γ) weights. All the 100 agents have the same values of a (FJ) and of α and γ (CB). The simulations start from the same initial opinions and evolve over digraphs with the same topology. The degree of susceptibility, prejudice, conformism, and stubbornness is represented by the colours cyan, magenta, blue, and green respectively.

The differences between CB and FJ model help visualise the strong implications of the classification-based mechanism for assessing the opinion of others, which captures the fact that opinions cannot be perceived with perfect resolution and accuracy, and hence changes the model behaviour significantly: it grants the model new properties, such as the existence of multiple equilibria that can span the complete spectrum of opinions. For instance, in Figure 20, the CB model with $\alpha = 1$ and $\gamma = 0$ generates equilibrium opinions that span almost 40% of the opinion interval $[-1, 1]$ (a wider span can be achieved with different topologies), while the FJ model with $a = 1$ leads to identical equilibrium opinions. The non-linearity introduced by the classification-based assessment of the opinion of others can completely change the resulting dynamics and lead to the emergence of peculiar features, which would not emerge from models where the agents have perfect access to the opinion of others, as the comparison with the Friedkin-Johnsen model highlights.

The results of an analysis for the FJ model, equivalent to the one reported in Tables 1 and 7 for the CB model, are reported in Tables 9 and 10. To make the results comparable, when solving the optimisation problems (10) and (11) for the FJ model the sets $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{W}}$ are modified as follows: the topology of each network in the set $\tilde{\mathcal{W}}$ is kept the same, but the weights are changed so that they are all positive and the associated adjacency matrix is row-stochastic. The inner traits assignations in $\tilde{\mathcal{A}}$ are transformed into parameters of the FJ model using the mapping $a_i = \alpha_i / (\alpha_i + \gamma_i)$; if $\alpha_i + \gamma_i = 0$, then $a_i = 0.5$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
Q1	1.8	8	2.4	1.2	1.8	2.2	2.4	6	7.2	6.2	2.6	1	1.4	4.8	5.4	9.4	4	2.6	1.6	10.6	4.2	1.8	1.2	5.6	2.4	5.4
Q2	5	4.2	1	1.4	8.8	7.4	2.6	14	4	3.6	3.8	3.6	1	13.8	2.4	8.6	9	1.4	2.8	6.4	3.2	3	1.6	3	4	3.8
Q3	3	6	2.2	2.2	7	3.2	3.2	13.4	11	5	4.8	2	10	18	2	1.4	9.4	1.8	1.4	9.4	2	2.2	2.8	4.4	5.4	8.6
Q4	15.6	12.2	18.6	19	29.4	9.8	10.8	25	3	15.6	9.6	9.8	10.8	24.4	17.8	5.6	6.2	17.2	3	38	2.4	15.8	5.4	52	14.2	8
Q5	4	6.6	8.4	13.2	4.6	4.8	8.2	10.4	18.4	3.2	9.2	3.8	7.6	23.6	1.6	11.2	0.8	8.8	0.4	12	1.6	4	3.6	5.6	6.6	5
Q6	5	7.8	6.4	5.8	13.8	4	4	14.6	7.6	4	8	11.8	3.8	10.2	3.6	13	1.8	14.8	0.4	23.8	11.8	3.2	1.8	15.8	4.6	7.4
Q7	7.2	7	11.6	1.6	4.8	2.2	3.8	3.2	8.4	5.2	11.8	5.4	1.8	11.8	7.8	25.2	5	4	1.6	6.2	0.6	6.8	2.8	3.8	4	8
Q8	10.6	8.6	1.8	1	11.2	5.4	7.2	5.2	4.6	5.2	12.6	1	3	9.2	4	21.8	5.2	3.6	5	14.2	0.4	3.2	1.6	8	5	4.6
Q9	13	6.6	3.2	2.2	20.6	1.8	9	6.6	13.4	15.2	11.2	3.6	15.4	8.8	8.4	3.8	7	15.6	6.6	10	5.6	4.2	9.8	7.2	6.6	14
Q10	17	9.8	7.2	1	20	4.4	5.2	8.4	11.8	11.8	10	5.2	9.2	7.4	6.8	5	5.4	11.2	3.2	9.4	3	3.2	10.8	8.6	1.8	18
Q11	18.2	4.2	6.2	2.8	12.8	9.6	3.2	10.4	12	13.6	10.4	1.2	19.8	10.2	17.6	2.6	14.2	12.2	1	8.4	13.8	17.4	5.8	7.4	4.6	3.8
Q12	13.4	9.6	6	2.4	2	31.8	6.4	11.6	3.8	19.8	16.2	6.4	13.8	9.6	4.8	0.6	8.6	9.2	11.6	8.4	5	2.8	1	32.2	6	13
Q13	13.8	3	6.8	1.6	1.2	1.4	1.2	8	4.2	2.6	15	1.6	12	3.4	5.2	17.4	3.6	4	4.2	2.2	64.8	1	3.6	2.8	3.2	6.2
Q14	1.8	7.4	7.4	3.8	6.8	0.2	6.4	8	2.2	1	23.6	7	6.4	3.6	16	35.2	3.6	5.4	2.6	25	12	10	2.8	10.2	1	5.4
Q15	2.8	18.4	2.6	4.2	4.4	0.8	12.2	17.4	0.6	2	17.6	6	0.4	3	7.6	26	2.2	4.4	0.8	17	8.4	2.6	1.6	8.6	1.8	3.4
Q16	0.6	4.8	3.8	4.4	2	0.4	5	14.2	5.6	0.6	15	4.4	2.6	1.2	6	35.4	3.2	1.6	0.4	16.4	2.2	3	0	4.6	0.6	4.8
Q17	13.8	10.2	5.8	6.6	3.2	1.2	2	17.6	2.2	7.6	9.2	1.8	8.6	3	7.8	22.4	7.4	3.4	9.2	10	4	10.6	3.8	1	14.4	12.6
Q18	1.6	3.6	3.8	16.2	2	10	4	21.2	10.4	2.6	14	3.6	1.6	4.8	11.8	22.4	2.4	6.4	1.8	8.8	1.6	1.2	2.2	5.6	5.2	11.2
Q19	4.6	14.4	9.2	16.8	4.2	2.4	4	26.6	13.8	4.6	11.6	7.2	5.2	6.6	5.2	9	6.4	4.6	2.6	14.4	7	4.4	4.4	2.8	7.2	7.4
Q20	5.4	2.4	4.6	10.8	2.4	1.8	4.8	20.6	6	8.6	12.2	3.6	3	2.2	5	32.2	3.2	5.2	6.6	9.4	4	2.4	1.8	4.2	5.8	10
Q21	1.4	1.8	0	9	1.8	0	1.8	15.6	0.8	4.2	11	2.6	1.2	4	1.4	28.2	0.6	2.4	0.2	12.4	2.4	0.8	0.6	1.8	2.8	4.6
Q22	4.2	9.2	8.6	3.2	7.2	11	19.2	11.8	13.8	4.8	10.8	5.2	10.6	12.2	5	0.6	2.2	2.6	1	24.6	26.6	10	4.4	7.8	2	8
Q23	5.6	10.2	7.2	13	1.6	14.2	3.4	6.2	18.6	11.8	16	1	5.8	15.8	5	11.4	5	8.6	1.6	18.6	13.6	6.2	1.6	7.2	1.6	3.4
Q24	3	5.8	3.8	17.8	2.4	12.8	2.2	12	20.6	3	13.2	5	13.2	17.2	2.6	15.4	0.8	5.2	4.2	17.6	3	2.8	4.6	3	1	3.2
Q25	4.8	8.8	1.2	1.8	5.6	20.2	16.4	10.6	8	3.8	12.4	2.2	3.8	16.6	1.6	18.4	1.4	2.8	0.8	19.2	13	4.2	3.4	4.6	3.4	3
Q26	3	3.8	1.8	14.8	3	1.6	1	6	25.2	4	11.8	5.2	2.4	3.2	1.6	17.4	0.8	7.2	2.6	17.4	18.8	6	7.4	10	3	2.4
Q27	2	9	7.2	6.2	16.6	35.2	23.4	8.4	26.2	3.8	20.8	2.8	3	15.8	0.6	12.6	1.4	14	5.4	19	4	6.4	12.6	2.8	6.2	8.2
Q28	4.4	3	0.4	6.2	9.2	19.6	9.2	3.8	15.2	8.4	17.4	1.4	4.2	10.8	1.8	14.2	1.2	4.8	9.4	10.4	7.2	6.4	8.2	1.6	2.2	4
Q29	1	8	4.2	1.6	1.2	6	9.8	7.6	16.4	4.8	12.6	3	1.4	3.8	2	17.6	0.2	4	1.4	9.4	0.8	2.4	9.6	2.4	1	0.4
Q30	2.2	10.4	11.8	3	21	7.4	20	3	16.6	7.2	3.4	5.8	3.4	18	20.6	8	11.2	3	8.4	9.4	7.8	1.6	6.8	3.8	0.8	1.8
Average	6.4	7.5	5.5	6.5	7.8	7.8	7.1	11.6	10.4	6.5	11.9	4.1	5.9	9.9	6	15.1	4.4	6.4	3.3	16	6.4	5.1	4.2	8	4.4	6.8
Total	190.6	224.8	165.2	194.8	232.6	232.8	212	347.4	311.6	193.8	357.8	124.2	176.4	297	179	452.2	133.4	192.8	99.8	480.6	191	152.2	126.8	238.8	131.4	203

Table 9: Results of the *Free* optimisation problem using the Friedkin-Johnsen model. The average cost along all the countries is 7.4897. Out of 780 possible question-country pairs, 460 have a cost less than 7 (an accuracy of 59% in total). The average cost of accurate (green) question-country pairs is 3.3.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26
Q1	2.6	8.2	6.6	3	3.6	3.6	7	9.2	9	7.6	3.2	1.6	2.8	6	7.4	10.8	4.6	4	2.8	11.8	9.4	2.2	5.2	7.6	4	5.6
Q2	6.2	4.8	3.2	5.6	10.4	10.2	7.2	17.6	7.6	4.4	5.2	4.6	3.2	16	4.2	11.4	10.8	2	4.6	7.2	4.4	5	4.8	3.2	4.6	5.2
Q3	4	7.6	8.4	7.2	9.6	6	5.6	15.4	19.2	5.8	5.8	3.4	13	21.2	3.2	4	10.2	2.4	2	12	5	3.4	3.8	5.4	5.8	10.6
Q4	19.2	13	23.4	25	36.8	8.8	15	42.2	8	19.4	10.4	12.4	12.8	26.2	28.4	7.2	8.8	21.8	4.4	39	6	17.2	10.2	68.6	15.8	8.6
Q5	5.6	7.2	11.2	19.4	7.2	8.2	11.6	14	31.4	4.2	11	6.4	9.8	25.2	5.6	20.6	1.6	11.8	1.8	12.4	4.6	4.8	5.2	5.6	7	5.4
Q6	5.6	8.4	10	14	17	6	15.2	23.2	11.8	6.8	11.4	16.6	5.6	12	6.8	19.8	2.2	18.2	2.4	26	18	3.8	7.6	19.4	5.6	7.8
Q7	10.8	8.2	18.4	2.6	5.6	4.2	4.8	6.8	16.6	5.6	13.8	8.2	2.8	13.6	11	29.2	5.8	5.4	3	7	1.2	9.6	5.2	4.4	4	9.2
Q8	12.6	10	5	2.6	12.6	10	9.6	8.2	10.4	6.4	17	1.8	4	10.8	8.4	27.8	7.6	7.4	9	17.4	3	4.4	4.4	8.4	5.6	5
Q9	14	7	4.6	4.6	29.6	2.2	11.8	8.2	17.8	17.8	16.4	3.8	17.2	10.2	11.4	7	9.4	17.4	10.4	10.2	8.2	6.2	12.2	8.8	8.6	20.4
Q10	21.8	12.8	11.4	3.2	33	5.8	7.8	11	23.6	13.8	13	8.6	10.6	8.6	9.2	10.6	8.8	13	6.4	9.8	8.6	4.6	12.8	11.2	3.8	27
Q11	21.8	4.2	9.6	4.4	20.6	13	11.6	14.4	15.6	18.2	13.4	3.6	23.4	13.6	23.6	9	16.4	14.2	5.4	9	17.2	23	12.8	9.2	7.4	4.2
Q12	17	10.2	9.2	5.8	2.6	44.8	11.2	23.6	14.2	26	20.6	10.2	17	11.4	8	3.2	16.4	9.6	16.4	10.6	8	3.6	2.6	44	7.4	16.6
Q13	15.2	3.4	9.8	5.6	4	3.8	2	18.2	5.8	3.4	19.4	2	15	3.4	10	21.4	5.2	6.4	5	73.4	2	4.8	5	4.6	8.2	14.2
Q14	2.6	10.4	13.4	12.6	9.4	1.4	12.2	15.8	9.2	2	26	10	7.8	4.6	9.2	38.4	8.4	6.2	5.8	27.2	18.8	13.6	3.6	14	2.2	7
Q15	4.6	24.8	4.8	5.6	10.2	2.6	12.2	27.8	2.2	3	20.6	8.2	1.8	5.2	22.2	30	7.2	6.4	2.2	18.8	15.2	2.6	2.6	11	3.2	9.2
Q16	2.4	8.4	5.4	6.8	3.6	1.8	9.6	29	9.4	3.4	16.2	7.2	3	2.8	13	38.4	4.8	3.2	2.2	21.6	7.6	4	1.2	8.6	1.2	7.6
Q17	18.6	10.6	12.2	10.8	5.2	2.6	4.6	28.4	4.4	9	13.4	4	11.2	4.4	11	25.4	9	5.2	13.6	10.8	5.6	15.8	4.4	3	18.8	13.4
Q18	2.8	4.8	8	18.2	4.6	12.4	8.2	32.2	14	4.6	18	4.6	4	6	15.8	28.8	4.6	7.4	3.4	11.4	3	2.6	4.6	9.8	6.2	14.4
Q19	6	16.2	14	21	7.6	5.2	9	43	16	5.2	16.8	9.2	8.4	8.4	11.2	11	8.2	5.4	3.8	15	12.4	9	8.8	5.8	9.2	8.6
Q20	6.6	4.2	7.2	13.6	3.4	2.2	7	34.4	8.8	12.4	20.6	6.2	3.6	3.8	6.4	34.6	4.4	7.8	7.8	12.6	9.4	4	2.4	7.2	7.6	13.2
Q21	3.8	5.6	1.6	13.4	2.8	12	6	25.2	2.6	5.6	16.8	4	1.4	5.4	2.6	32.8	3.8	2.8	2	18.8	10.6	2.2	1.8	4.6	3.2	9.4
Q22	4.4	9.4	12	6.4	9.6	13	26.6	20.4	16.2	5.2	12	8.8	14.4	13	8	3.2	5.8	3.2	2.8	25.8	32.6	11.8	9.8	9	2.4	8.8
Q23	8.6	10.4	7	22.4	3.2	17.4	9.2	6.6	37.6	15.8	16.4	3	11	19.4	8.6	19.2	7	9	3.8	19.2	20.8	9.4	5.4	9.8	2.2	4.4
Q24	4.2	6	6.8	26	3	15.6	4.6	18.6	23	3.6	15.8	7.8	4	18.8	5	23.6	2.8	5.2	5.6	20.4	4.4	2.8	6.4	3	2.2	4
Q25	5.6	8.8	1.8	2.2	6	23.8	25.6	19.6	21	4.4	15.6	5.2	7	20.4	3.8	20.6	4.6	1.6	22.6	17.8	5.4	4.4	6.8	5.8	4.4	
Q26	4	4.2	4	21	3.6	3.8	4.8	7.8	35.2	7	15.4	9.8	2.8	3.8	5.4	27.6	1.8	15.8	4.8	18	22	9.2	13.4	12	4.2	3.2
Q27	2.8	11.8	11.4	9.8	20.2	39.6	35.8	14	35.6	4.6	23	5.6	5.4	17.8	2.2	16.6	2.4	15.2	9.6	6.8	13.8	3.6	8.8	12		
Q28	4.8	3.4	3	7.8	12	23.4	12.6	5.4	21.4	8.6	23	2.4	6	11.8	3.4	18.4	3	4.8	13.2	10.4	8.8	8.4	9.6	2.6	3	
Q29	1.6	8.2	6.2	5	1.4	7.2	12.8	16.6	26.2	5.4	14.8	5.4	4	4	3	21.6	1.4	4.6	2.4	10.2	4	2.8	11.4	2.8	2	2.2
Q30	2.6	12.6	14.2	4.8	23.6	11.2	24	15.4	22.4	7.6	5		4.4	13.3	3.6	9.4	10	11	1.8	10.4	4.6	2.4	2	2		
Average	8.1	8.8	8.8	10.3	10.7	10.4	11.5	19.1	16.5	8.2	15	6.4	8	11.6	9.4	19.5	6.6	8.2	5.6	18	10.3	6.8	6.9	10.6	5.8	9
Total	242.4	264.8	263.8	310.4	322	311	345.2	572.2	496.2	246.8	450	191.4	239	347	283.2	584.6	198.6	246	167.6	541.4	309.2	204.8	205.8	318.6	174	269.6

Table 10: Results of the *Constrained* optimisation problem using the Friedkin-Johnsen model. The average cost along all the countries is 10.3918. Out of 780 possible question-country pairs, 330 have a cost less than 7 (an accuracy of 42% in total). The average cost of accurate (green) question-country pairs is 4.1.

Comparing Tables 1 and 9 shows that the CB model outperforms the FJ model, yielding a 97% accuracy in contrast to 59%. Also, the average cost of country-question pairs with cost less than 7 is lower for the CB model (2.97) compared with the one produced by the FJ model (3.3), indicating that not only more question-country pairs are predicted satisfactorily, but also the predictions are more accurate.

General agreement analysis

Given a single initial opinion vector x_o and a set of inner traits assignations $\mathscr{A}$ and networks $\mathscr{N}$, let $X = X(x_o, \mathscr{A}, \mathscr{N})$ be the set of all final opinion vectors produced by evolving the initial opinions x_o with inner traits assignation $\psi \in \mathscr{A}$ over a network $W \in \mathscr{N}$. The plot of the *general agreement* (θ_+, θ_-), as defined in Equation (13), of each final opinion vector in $X(x_o, \mathscr{A}, \mathscr{N})$ in the Cartesian plane gives a visual representation of the range of opinions that the model can produce starting from x_o. Figure 21 shows this plot for five different initial opinion vectors (seen in Figure 22) for the CB and the FJ models. In the plots, each dot is colour-coded: for the CB model, the colour represents the average inner traits (blue for average conformist trait $\overline{\alpha}$, red for average radical trait $\overline{\beta}$, and green for average stubborn trait $\overline{\gamma}$); for the FJ model, it represents the average susceptibility $\overline{a}$ (cyan represents complete susceptibility $\overline{a} = 1$ and magenta complete prejudice $\overline{a} = 0$). We consider the same set of traits as for the optimisation problem, i.e., $\mathscr{A} = \tilde{\mathscr{A}}$, and a set of networks $\mathscr{N}$ including networks 1 to 5 in $\tilde{W}$, selected because they represent networks with the same topology and varying ratio of negative to positive edges.
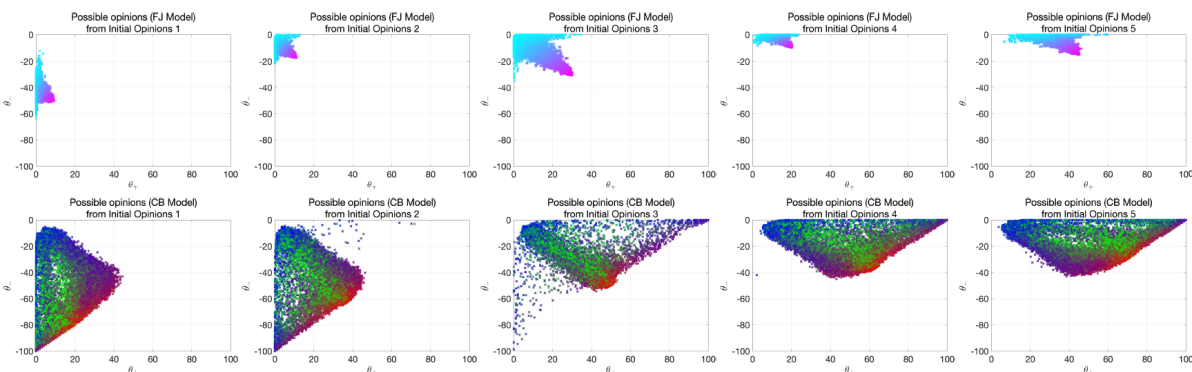

Figure 21: Row 1 (respectively, row 2) shows the *general agreement* plot of the potential opinion vectors predicted by the FJ (respectively, CB) model, starting from the initial opinion vectors shown in Figure 22. The marker colour encodes the average traits of the agents producing the final opinion: for the FJ model, cyan and magenta represent susceptibility and prejudice respectively; for the CB model, blue, red and green represent conformism, radicalism and stubbornness respectively. All simulations evolved over 50 time steps.
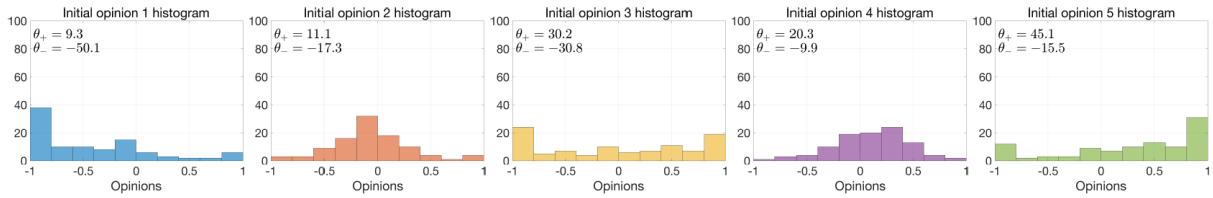
Figure 22: Initial conditions for the opinion evolutions shown in Figure 21

Figure 21 highlights that the CB model can produce a wide range of predicted opinions, primarily thanks to the interaction of the three complementary inner traits: conformism brings opinions together, yielding an effect that is similar to that of the FJ model, i.e., moving opinions along the $3\pi/4$ diagonal; radicalism allows the opinions to move in the other 3 diagonal directions ($\pi/4$, $5\pi/4$ and $7\pi/4$); and stubbornness makes opinions stay near the initial opinion. The combined effect of multiple combinations of these traits leads to possible predicted opinions that have a greater range than the ones produced by the FJ model.

Susceptibility or conformism could be considered as a trait that moves all the opinions towards a common opinion in the middle of the interval $[-1, 1]$, whereas radicalism moves all the opinions to their corresponding extremes, either -1 or 1. The combined effect moves the opinions across the interval $[-1, 1]$, and the degree of stubbornness determines the speed of the change. Interestingly, if the initial opinion is above the line $\theta_+ = -\theta_-$, then the predicted opinions also tend to stay above the line: if the overall population agrees more than it disagrees with a statement, this opinion balance is likely to be preserved, this was already noted in Figure 16. However, a change in opinion balance is not impossible: a considerable fraction of predicted opinions may approach the line $\theta_+ = -\theta_-$, and from there a change in population traits may move the predicted opinions to the other side of the line (initial opinion 2 can produce an opinion vector near initial opinion 1 and from there the next opinions can be near initial opinion 3). It is also interesting to note that, although initial opinions 1 are slightly below the $\theta_+ = -\theta_-$ line, most of the predicted opinions are above that line, an effect probably caused by the network topology or the edge signs.

Appendix C: Optimisation Approach

The subset $\tilde{\mathcal{W}}$ contains 35 different small-world signed strongly connected digraphs. Table 11 shows the values of the main metrics for the networks.

ID	1	2	3	4	5	6	7	8	9	10	11	12
APL	2.13	2.13	2.13	2.13	2.13	1.95	1.95	1.95	1.95	1.95	2.04	2.04
CC	0.38	0.38	0.38	0.38	0.38	0.18	0.18	0.18	0.18	0.18	0.16	0.16
PE	252	558	834	1115	1436	258	566	848	1145	1438	222	533
NE	1349	1043	767	486	165	1326	1018	736	439	146	1194	883
D	4	4	4	4	4	3	3	3	3	3	3	3
BI	0.00015	4.4e-05	3.8e-05	0.00013	0.042	0.00023	8.1e-05	4.8e-05	0.00027	0.049	0.00099	0.00025

ID	13	14	15	16	17	18	19	20	21	22	23	24
APL	2.04	2.04	2.04	1.75	1.75	1.75	1.75	1.75	1.68	1.68	1.68	1.68
CC	0.16	0.16	0.16	0.25	0.25	0.25	0.25	0.25	0.35	0.35	0.35	0.35
PE	746	1020	1259	362	864	1351	1813	2344	418	1079	1683	2372
NE	670	396	157	2227	1725	1238	776	245	2891	2230	1626	937
D	3	3	3	3	3	3	3	3	2	2	2	2
BI	0.00021	0.00056	0.047	2e-08	6.1e-09	4.1e-09	1e-07	0.0071	3.4e-11	5.8e-12	7.5e-13	6.7e-09

ID	25	26	27	28	29	30	31	32	33	34	35	
APL	1.68	1.68	1.68	1.68	1.68	1.68	1.62	1.62	1.62	1.62	1.62	
CC	0.35	0.32	0.32	0.32	0.32	0.32	0.39	0.39	0.39	0.39	0.39	
PE	2947	456	1063	1667	2329	2972	457	1259	1998	2717	3506	
NE	362	2823	2216	1612	950	307	3440	2638	1899	1180	391	
D	2	2	2	2	2	2	2	2	2	2	2	
BI	0.00074	4.8e-11	8.6e-12	1.3e-12	3.7e-09	0.0021	4.7e-14	3.6e-14	3.2e-14	4.3e-11	0.00033	

Table 11: Signed Digraph Information: Average Path Length (APL), Clustering Coefficent (CC), Positive Edges (PE), Negative Edges (NE), Diameter (D), and Balance Index (BI)

The subset $\tilde{\mathcal{A}}$ contains 3528 randomly generated inner traits assignations $\psi = (\psi_i)_{i=1}^n$. To avoid bias towards societies with average inner traits that are more conformist, radical, or stubborn, the set $\tilde{\mathcal{A}}$ satisfies the following property: for every inner traits assignation ψ, with corresponding average inner trait $\bar{\psi} = (\bar{\alpha}, \bar{\beta}, \bar{\gamma}) = (a_1, b_1, c_1)$, there are two inner traits assignations $\psi', \psi'' \in \tilde{\mathcal{A}}$ that satisfy $\psi' = (b_1, c_1, a_1)$, and $\psi'' = (c_1, a_1, b_1)$. In other words, the parameter space $\tilde{\mathcal{A}}$ is symmetric with respect to permutations of agent traits. Besides this property, the elements of this set were chosen at random. All the average inner traits $\bar{\psi}$ corresponding to inner traits assignations ψ in $\tilde{\mathcal{A}}$ are shown in Figure 23.
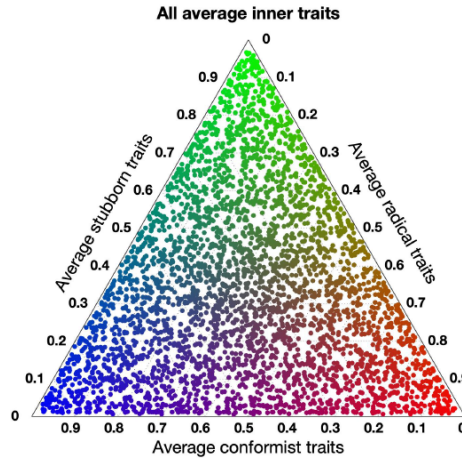

Figure 23: All the average inner traits $\bar{\psi}$ corresponding to inner traits assignations ψ in $\tilde{\mathcal{A}}$.

Due to the anonymity of the surveys, it is not possible to guarantee that the same people answered the survey in subsequent waves of the WVS. However, if the surveys are done correctly to represent society overall, the results can be anyway assumed to reflect the global opinion distribution of the general population about a given topic at a specific time, and this allows us to use the survey results in different waves in our minimisation problem, *as if* the very same people had answered. Moreover, even though all our simulations could be run for an arbitrary number of agents, we consider $n = 100$ agents (as representatives of the whole society) to keep the computational demand (especially for the optimisation problem) manageable.

Appendix D: Network Metrics

The signed digraph is represented by the weight matrix $W \in \{-1, 0, 1\}^{n \times n}$, where w_{ij} is associated with the edge going from vertex j to vertex i. We consider six network metrics: average path length (APL), clustering coefficient (CC), positive edges (PE), negative edges (NE), diameter (D), and balance index (BI). This appendix explains how these metrics are computed.

A directed path is a K-tuple of vertices $(p_1, p_2, \ldots, p_i, p_{i+1}, \ldots, p_K)$ such that there is an edge from vertex p_i to vertex p_{i+1} for $i = 1, \ldots, K-1$. The length $|p|$ of a directed path p is the number of edges that it crosses. Let $P(i, j)$ be the set of all directed paths from vertex i to vertex j (if there are none, then $P(i, j) = \emptyset$). Denote by $d(i, j)$ the length of the shortest directed path from i to j, i.e., $d(i, j) := \min_{p \in P(i,j)} |p|$. Let $C(W)$ be the set of vertex pairs (i, j) such that there exists a direct path from i to j and $i \neq j$, i.e. $C(W) = \{(i, j) \mid P(i, j) \neq \emptyset \text{ and } i \neq j\}$. Then the average path length and diameter of the digraph W are:

$$APL = \frac{1}{|C(W)|} \sum_{(i,j) \in C(W)} d(i,j) \qquad \text{and} \qquad D = \max_{(i,j) \in C(W)} d(i,j) \tag{17}$$

Note that, because all the networks are strongly connected, $|C(W)| = n(n-1)$.

To compute the clustering coefficient, consider agent i, with k_i in-neighbours excluding itself: $k_i = |\tilde{\mathcal{N}}_i|$, where $\tilde{\mathcal{N}}_i := \{j \in V \mid w_{ij} \neq 0, i \neq j\}$. Then there are at most $k_i(k_i - 1)$ directed edges between these neighbours.

The fraction c_i of these edges that is actually present is the clustering coefficient of agent i. If agent i has only one in-neighbour, then its clustering coefficient is 1, and if it has no in-neighbour but itself c_i is not defined:

$$c_i = \begin{cases} \frac{|\{(j,k)|j\neq k \text{ and } i,k\in\tilde{\mathcal{N}}_i\}|}{k_i(k_i-1)} & \text{if } k_i > 1 \\ 1 & \text{if } k_i = 1 \\ nan & \text{if } k_i = 0 \end{cases} \tag{18}$$

The clustering coefficient of the network (defined by extending to digraphs the definition for undirected graphs by Watts & Strogatz (1998)) is thus the average of the clustering coefficients of all agents with at least one in-neighbour excluding themselves:

$$CC = \frac{1}{|\{i \in V \mid k_i > 0\}|} \sum_{i:k_i>0} c_i \tag{19}$$

The number of positive and negative edges are computed as

$$PE = \sum_{i,j\in V:w_{ij}>0} 1 \quad \text{and} \quad NE = \sum_{i,j\in V:w_{ij}<0} 1 \tag{20}$$

Finally, the balance index is computed as

$$BI = \frac{tr(exp(W))}{tr(exp(D))} \tag{21}$$

where $tr(\cdot)$ is the trace operator, $exp(\cdot)$ is the matrix exponential, and $D = |W|$ component-wise. This formula is a direct extension of the balance index for undirected graphs proposed by Estrada (2019); Estrada & Benzi (2014).

Appendix E: Simulation Process

The **Free optimisation problem** in Equation (10) with sets $\mathcal{W} = \tilde{\mathcal{W}}$ and $\mathcal{A} = \tilde{\mathcal{A}}$ was solved using the algorithm:

1. **Input:** survey answers for waves 5 and 6 for a given country.
2. Set $w_0 = \infty$; this will be the minimum cost across all networks
3. For network $W \in \tilde{\mathcal{W}}$

- For question $q \in \{1, 2, \ldots, 30\}$
  - Set $v_q = \infty$ to be the minimum cost for question q
  - For inner traits assignation $\psi^{(l)} \in \tilde{\mathcal{A}}$
    * Compute the predicted opinions $\tilde{y}_q$ after K iterations evolving over the network W with inner traits assignation $\psi^{(l)}$ starting with initial opinions x_q. These initial opinions are the survey results to question q in wave 5.

    $$\tilde{y}_q = \mathcal{F}_\Omega(x_q, W, \psi^{(l)}, K)$$

    * Compute the mismatch J (Equation (15)) between these predicted opinions $\tilde{y}_l$ and the real opinions y_l given by survey results of question q in wave 6.
    * if $J(\tilde{y}_q, y_q) < v_q$
      · Set $v_q = J(\tilde{y}_q, y_q)$ as the current minimum cost across all inner traits assignations.
      · Set $\widehat{\psi^{(q)}} = \psi^{(l)}$ as the inner traits assignation that gives the lowest cost for question q.
- Add all the minimum costs to obtain the minimum cost for the network W

$$J_{\text{Total}} = \sum_{q=1}^{30} v_q$$

- if $J_{\text{Total}} < w_0$

- Set $w_0 = J_{\text{Total}}$ as the current minimum cost across all networks for this country.
- Set $\widehat{W} = W$ as the network that produces the minimum cost for this country.

4. **Output:** network $\widehat{W}$ and set of inner traits assignations $(\widehat{\psi^{(l)}})_{l=1}^{30}$ that give the minimum total cost across all questions.

In the algorithm used to solve the **Constrained optimisation problem** in Equation (11), both the network and the inner traits assignations are the same for each question:

1. **Input:** survey answers for waves 5 and 6 for a given country.
2. Set $w_0 = \infty$; this will be the minimum cost across all networks and inner traits assignations
3. For network $W \in \tilde{\mathcal{W}}$

- For inner traits assignation $\psi \in \tilde{\mathcal{A}}$
  - For question $q \in \{1, 2, \dots, 30\}$
    * Compute the predicted opinions $\tilde{y}_q$ after K iterations evolving over the network W with inner traits assignation ψ starting with initial opinions x_q. These initial opinions are the survey results to question q in wave 5.

$$\tilde{y}_q = \mathcal{F}_\Omega(x_q, W, \psi, K)$$

    * Compute the mismatch J (Equation (15)) between the predicted opinions $\tilde{y}_q$ and the real opinions y_q given by survey results of question q in wave 6.
  - Add all the costs to obtain the cost for the network W and the inner traits assignation ψ.

$$J_{\text{Total}} = \sum_{q=1}^{30} v_q$$

  - if $J_{\text{Total}} < w_0$
    * Set $w_0 = J_{\text{Total}}$ as the current minimum cost across all networks and inner traits assignations for this country.
    * Set $\widehat{W} = W$ as the network that produces the minimum cost for this country.
    * Set $\hat{\psi} = \psi$ as the inner traits assignation that gives the minimum cost for this country.

4. **Output:** network $\widehat{W}$ and inner traits assignations $\hat{\psi}$ that give the minimum total cost across all questions.

The data sets used to produce the results shown in the paper can be downloaded from the following link: `https://giuliagiordano.dii.unitn.it/docs/papers/OpinionModel.zip`, together with the corresponding code and instructions on how to use the code.

● Appendix F: Bias, Diversity and Fragmentation

Table 12 explains how *Bias*, *Diversity*, and *Fragmentation* (Lorenz et al. 2021) are computed and the meaning of high and low values.

Name	Computation	High value meaning	Low value meaning
Bias	$\|\bar{x}\| = \frac{1}{\|V\|}\|\sum_{i \in V} x_i\|$ where $\|V\|$ is the cardinality of the set of vertices V	Most agents have the same opinion, which is near either complete disagreement or complete agreement (near the extremes -1 or 1).	Either most agents have a similar opinion near indifference or a comparable number of agents agree and disagree.
Diversity	$\left(\frac{1}{\|V\|}\sum_{i \in V}(x_i - \bar{x})\right)^{1/2}$ where $\bar{x}$ is the mean of the opinions	Opinions are spread out. Maximum *Diversity* is reached when exactly half of the agents have complete agreement and exactly half complete disagreement.	Agents have a similar opinion. Minimum *Diversity* is reached when all the agents have the exact same opinion.
Fragmentation	$\frac{1}{2N}\sum_{i=0}^{N+1}\|X_{i-1} - X_i\|$ where N is the number of bins in the histogram, X_i is the number of agents in bin $i \in \{1, 2, \ldots, N\}$, and $X_0 = X_{N+1} = 0$	There is a significant concentration of agents around one or multiple opinions. Maximum *Fragmentation* is reached when the histogram is such that no two adjacent bins have non-zero count, i.e., there is always at least one empty bin between non-empty bins.	The opinion distribution is evenly distributed. The minimum *Fragmentation* is reached when all the bins have the same number of agents.

Table 12: *Bias*, *Diversity*, and *Fragmentation* computation and meaning of high and low values, taken from (Lorenz et al. 2021); *Fragmentation* has been adapted to be computed from histograms.

Appendix G: Countries and Questions

We report here the list of countries (in Table 13) and the list of questions (in Table 14) that we considered, from the real data collected by the World Values Survey.

C1	Australia	C2	Brazil	C3	Chile
C4	China	C5	Cyprus	C6	Georgia
C7	Ghana	C8	India	C9	Jordan
C10	Japan	C11	Malaysia	C12	Mexico
C13	Poland	C14	Romania	C15	Slovenia
C16	South Africa	C17	Spain	C18	Sweden
C19	South Korea	C20	Thailand	C21	Trinidad
C22	Taiwan	C23	Turkey	C24	Ukraine
C25	United States	C26	Uruguay		

Table 13: Countries

Q1	Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means **no choice at all** and 10 means **a great deal of choice** to indicate how much freedom of choice and control you feel you have over the way your life turns out
Q2	All things considered, how satisfied are you with your life as a whole these days? Using this card on which 1 means you are **completely dissatisfied** and 10 means you are **completely satisfied** where would you put your satisfaction with your life as a whole?
Q3	How satisfied are you with the financial situation of your household? Please use this card again to help with your answer (1 is completely dissatisfied, 10 is completely satisfied)
Q4	How would you place your views on this scale? 1 means you completely agree with the statement **Incomes should be made more equal**; 10 means you completely agree with the statement **We need larger income differences as incentives for individual effort**. And if your views fall somewhere in between, you can choose any number in between.
Q5	How would you place your views on this scale? 1 means you completely agree with the statement **Private ownership of business and industry should be increased**; 10 means you completely agree with the statement **Government ownership of business and industry should be increased**. And if your views fall somewhere in between, you can choose any number in between.
Q6	How would you place your views on this scale? 1 means you completely agree with the statement **The government should take more responsibility to ensure that everyone is provided for**; 10 means you completely agree with the statement **People should take more responsibility to provide for themselves**. And if your views fall somewhere in between, you can choose any number in between.
Q7	How would you place your views on this scale? 1 means you completely agree with the statement **Competition is good. It stimulates people to work hard and develop new ideas**; 10 means you completely agree with the statement **Competition is harmful. It brings out the worst in people**. And if your views fall somewhere in between, you can choose any number in between.
Q8	How would you place your views on this scale? 1 means you completely agree with the statement **In the long run, hard work usually brings a better life**; 10 means you completely agree with the statement **Hard work doesn't generally bring success—it's more a matter of luck and connections**. And if your views fall somewhere in between, you can choose any number in between.
Q9	How much you agree or disagree with the statement **Science and technology are making our lives healthier, easier, and more comfortable.**. For this questions, a 1 means that you "completely disagree" and a 10 means that you "completely agree."
Q10	How much you agree or disagree with the statement **Because of science and technology, there will be more opportunities for the next generation.**. For this questions, a 1 means that you "completely disagree" and a 10 means that you "completely agree."
Q11	How much you agree or disagree with the statement **We depend too much on science and not enough on faith.**. For this questions, a 1 means that you "completely disagree" and a 10 means that you "completely agree."
Q12	All things considered, would you say that the world is better off, or worse off, because of science and technology? 1 means that "the world is a lot worse off," and 10 means that "the world is a lot better off."
Q13	How important is God in your life? 10 means "very important" and 1 means "not at all important."
Q14	Indicate if the action of **Claiming government benefits to which you are not entitled** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q15	Indicate if the action of **Cheating on taxes if you have a chance** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q16	Indicate if the action of **Someone accepting a bribe in the course of their duties** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q17	Indicate if the action of **Homosexuality** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q18	Indicate if the action of **Abortion** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q19	Indicate if the action of **Divorce** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q20	Indicate if the action of **Suicide** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q21	Indicate if the action of **For a man to beat his wife** can be never justified (1); always justified (10); or something in between in a scale from 1 to 10.
Q22	**Governments tax the rich and subsidize the poor.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q23	**Religious authorities interpret the laws.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q24	Is **People choose their leaders in free elections.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q25	Is **People receive state aid for unemployment.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q26	Is **The army takes over when government is incompetent.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q27	Is **Civil rights protect people7s liberty against oppression.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q28	Is **Women have the same rights as men.** an essential characteristic of democracy? Use this scale where 1 means "not at all an essential characteristic of democracy" and 10 means it definitely is "an essential characteristic of democracy"
Q29	How important is it for you to live in a country that is governed democratically? On this scale where 1 means it is "not at all important" and 10 means "absolutely important" what position would you choose?
Q30	And how democratically is this country being governed today? Again using a scale from 1 to 10, where 1 means that it is "not at all democratic" and 10 means that it is "completely democratic," what position would you choose?

Table 14: Questions

References

Afshar, M. & Asadpour, M. (2010). Opinion formation by informed agents. *Journal of Artificial Societies and Social Simulation*, *13*(4), 5

Altafini, C. (2013). Consensus problems on networks with antagonistic interactions. *IEEE Transactions on Automatic Control*, *58*(4), 935–946

Anderson, B. D. O. & 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

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1

Asch, S. E. (1961). Effects of group pressure upon the modification and distortion of judgments. In M. Henle (Ed.), *Documents of Gestalt Psychology*, (pp. 222–236). Berkeley, CA.: University of California Press

Banisch, S. & Olbrich, E. (2019). Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, *43*(2), 76–103

Banisch, S. & Shamon, H. (2021). Biased processing and opinion polarisation: Experimental refinement of argument communication theory in the context of the energy debate. Available at: `https://arxiv.org/abs/2212.10117`

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M. & Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, *124*(4), 048301

Cartwright, D. & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review*, *63*(5), 277–293

Ceragioli, F. & Frasca, P. (2018). Consensus and disagreement: The role of quantized behaviors in opinion dynamics. *SIAM Journal on Control and Optimization*, *56*(2), 1058–1080

Chattoe-Brown, E. (2014). Using agent based modelling to integrate data on attitude change. *Sociological Research Online*, *19*(1), 159–174

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M. & Hoyst, J. (2011). Negative emotions boost user activity at BBC forum. *Physica A: Statistical Mechanics and its Applications*, *390*(16), 2936–2944

Coaley, K. (2014). *An Introduction to Psychological Assessment and Psychometrics*. Thousand Oaks, CA: Sage

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

Deffuant, G. (2006). Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, *9*(3), 8

Deffuant, G., Amblard, F., Weisbuch, G. & Faure, T. (2002). How can extremism prevail? A study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation*, *5*(4), 1

DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121

Duggins, P. (2017). A psychologically-motivated model of opinion change with applications to American politics. *Journal of Artificial Societies and Social Simulation*, *20*(1), 13

Elgazzar, A. S. (2003). Applications of small-world networks to some socio-economic systems. *Physica A: Statistical Mechanics and its Applications*, *324*(1), 402–407

Epstein, J. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12

Estrada, E. (2019). Rethinking structural balance in signed social networks. *Discrete Applied Mathematics*, *268*, 70–90

Estrada, E. & Benzi, M. (2014). Walk-based measure of balance in signed networks: Detecting lack of balance in social networks. *Physical Review E*, *90*(4), 042802

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Palo Alto, CA: Stanford University Press

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), 2

French Jr., J. (1956). A formal theory of social power. *Psychological Review*, *63*(3), 181–194

Friedkin, N. (1986). A formal theory of social power. *Journal of Mathematical Sociology*, *12*(2), 103–126

Friedkin, N. & Johnsen, E. (1999). Social influence networks and opinion change. *Advances in Group Processes*, *16*, 1–29

Friedkin, N. & Johnsen, E. (2011). *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. Cambridge: Cambridge University Press

Fu, G. & Zhang, W. (2016). Opinion formation and bi-polarization with biased assimilation and homophily. *Physica A: Statistical Mechanics and its Applications*, *444*, 700–712

Granovetter, M. & Soong, R. (1986). Threshold models of interpersonal effects in consumer demand. *Journal of Economic Behavior & Organization*, *7*(1), 83–99

Granovetter, M. S. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420–1443

Guo, M. & Dimarogonas, D. V. (2013). Consensus with quantized relative state measurements. *Automatica*, *49*(8), 2531–2537

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E. & Puranen, B. (2010). World Values Survey: Round five - country-pooled datafile. Available at: `https://www.worldvaluessurvey.org/WVSDocumentationWV5.jsp`

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E. & Puranen, B. (2015). World Values Survey: Round six - country-pooled datafile. Available at: `https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`

Harary, F. (1959). A criterion for unanimity in French's theory of social power. In D. Cartwright (Ed.), *Studies in Social Power*, (pp. 168–182). Ann Arbor, MI: University of Michigan Press

Harary, F., Norman, R. & Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs*. Hoboken, NJ: John Wiley & Sons

Hassanibesheli, F. & Donner, R. V. (2019). Network inference from the timing of events in coupled dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *29*(8), 083125

Hegselmann, R. & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3), 2

Hegselmann, R. & Krause, U. (2006). Truth and cognitive division of labour first steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*(3), 10

Hegselmann, R. & Krause, U. (2015). Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. *Networks and Heterogeneous Media*, *10*(3), 477–509

Kacperski, K. & Holyst, J. (1999). Opinion formation model with strong leader and external impact: A mean field approach. *Physica A: Statistical Mechanics and its Applications*, *269*(2), 511–526

Kacperski, K. & Holyst, J. (2000). Phase transitions as a persistent feature of groups with leaders in models of opinion formation. *Physica A: Statistical Mechanics and its Applications*, *287*(3–4), 631–643

Kandler, A. & Powell, A. (2018). Generative inference for cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743), 20170056

Krawczyk, M., Malarz, K., Korff, R. & Kułakowski, K. (2010). Communication and trust in the bounded confidence model. Lecture Notes in Computer Science

La Rocca, C., Braunstein, L. & Vazquez, F. (2014). The influence of persuasion in opinion formation and polarization. *EPL*, *106*(4)

Larsen, K. (1974). Conformity in the asch experiment. *Journal of Social Psychology*, *94*(2), 303–304. doi:10.1080/00224545.1974.9923224

Leinhardt, S. (1977). *Social Networks: A Developing Paradigm*. New York, NY: Academic Press

Liu, Q., Zhao, J. & Wang, X. (2015). Multi-agent model of group polarisation with biased assimilation of arguments. *IET Control Theory and Applications*, *9*(3), 485–492

Lorenz, J., Neumann, M. & Schröder, T. (2021). Individual attitude change and societal dynamics: Computational experiments with psychological theories. *Psychological Review*, *128*(4), 623

Lu, F., Maggioni, M. & Tang, S. (2021). Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *Journal of Machine Learning Research*, *22*

Martins, A. C. (2008). Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, *19*(04), 617–624

Mäs, M. & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS One*, *8*(11), e74516

Mastroeni, L., Vellucci, P. & Naldi, M. (2019). Agent-based models for opinion formation: A bibliographic survey. *IEEE Access*, *7*(8701567), 58836–58848

Masuda, N. (2015). Opinion control in complex networks. *New Journal of Physics*, *17*, 1–11

Matz, D. C. & Wood, W. (2005). Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, *88*(1), 22

Mckeown, G. & Sheehy, N. (2006). Mass media and polarisation processes in the bounded confidence model of opinion dynamics. *Journal of Artificial Societies and Social Simulation*, *9*(1), 11

Opp, K.-D. (1984). Balance theory: Progress and stagnation of a social psychological theory. *Philosophy of the Social Sciences*, *14*(1), 27–49

Pinasco, J. P., Semeshenko, V. & Balenzuela, P. (2017). Modeling opinion dynamics: Theoretical analysis and continuous approximation. *Chaos, Solitons and Fractals*, *98*, 210–215

Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, *9*(1), 13

Sandrock, C. (2012). alchemyst/ternplot. Available at: `https://www.mathworks.com/matlabcentral/file exchange/2299-alchemyst-ternplot?s_tid=srchtitle`

Schweighofer, S., Schweitzer, F. & Garcia, D. (2020). A weighted balance model of opinion hyperpolarization. *Journal of Artificial Societies and Social Simulation*, *23*(3), 5

Shang, Y. (2021). Resilient consensus for expressed and private opinions. *IEEE Transactions on Cybernetics*, *51*(1), 318–331

Sobkowicz, P. (2009). Studies of opinion stability for small dynamic networks with opportunistic agents. *International Journal of Modern Physics C*, *20*(10), 1645–1662

Sobkowicz, P. (2018). Opinion dynamics model based on cognitive biases of complex agents. *Journal of Artificial Societies and Social Simulation*, *21*(4), 8

Sobkowicz, P. & Sobkowicz, A. (2010). Dynamics of hate based internet user networks. *European Physical Journal B*, *73*(4), 633–643

Su, J., Liu, B., Li, Q. & Ma, H. (2014). Coevolution of opinions and directed adaptive networks in a social group. *Journal of Artificial Societies and Social Simulation*, *17*(2), 4

Thompson, N. & Derr, P. (2009). Contra Epstein, good explanations predict. *Journal of Artificial Societies and Social Simulation*, *12*(1), 9

Troitzsch, K. (2009). Not all explanations predict satisfactorily, and not all good predictions explain. *Journal of Artificial Societies and Social Simulation*, *12*(1), 10

Urbig, D., Lorenz, J. & Herzberg, H. (2008). Opinion dynamics: The effect of the number of peers met at once. *Journal of Artificial Societies and Social Simulation*, *11*(2), 4

Urbig, D. et al. (2003). Attitude dynamics with limited verbalisation capabilities. *Journal of Artificial Societies and Social Simulation*, *6*(1), 2

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442

Xia, W., Cao, M. & Johansson, K. (2016). Structural balance and opinion separation in trust-mistrust social networks. *IEEE Transactions on Control of Network Systems*, *3*(1), 46–56

Ye, M., Qin, Y., Govaert, A., Anderson, B. D. O. & Cao, M. (2019). An influence network model to study discrepancies in expressed and private opinions. *Automatica*, *107*, 371–381

Yin, X., Wang, H., Yin, P. & Zhu, H. (2019). Agent-based opinion formation modeling in social network: A perspective of social psychology. *Physica A: Statistical Mechanics and its Applications*, *532*, 121786