Understanding Polarisation: Bridging Opinion Dynamics with Empirical Distributions and Groups

Duncan Cassells

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Complex Networks, Department of Computer Science,
Sorbonne Université,
France
October, 2025

Abstract

The concern of this thesis is the modelling of opinion polarisation. There is increasing worry that the political debate of today has levels of polarisation that result in hostility between groups and the obstruction of collective decision making. One research approach to studying polarisation is opinion dynamics; it consists of the formal mathematical modelling and simulation of the opinions of a population to understand how they might reach consensus, polarisation, or somewhere in between. Simulating opinions and understanding population dynamics through scenarios, as well as identifying tipping points between consensus and polarisation, avoids the difficulty and complexity of observing a population's opinions change over time. However, a drawback is the limited empirical validation of current opinion dynamics research, which raises questions as to its relevancy in real-world application. A further shortcoming of the current literature is that the impacts of social identity and attitudes towards groups are rarely considered, despite the importance placed on these concepts in the social science literature. These two research gaps – empirical validation and group identification – are the subjects addressed in the following work in order to place opinion dynamics in an improved real-world context.

To answer this research aim, two contributions in the form of two modelling extensions are proposed. First, the perception of groups is incorporated into an existing opinion dynamics model such that the convergence or divergence of opinions is directly impacted by an individual's understanding of others as either in-group or out-group. The second extension seeks to enable the empirical validation of models by providing a framework that finds plausible model parameters for simulation. This is achieved by a mean-field approximation of the same model and subsequent numerical simulation, which allows for the modelling of opinion distributions of populations rather than the agent-based approach which focuses on an individual's opinion change. A set of criteria is then established in order to find simulated distributions that match behaviour displayed by empirical distributions and so may be considered plausible.

The findings from the group identification extension reveal that treatment of out-group is a central part of understanding the eventual polarisation of a population, while the influence of in-group interactions can temper extreme

opinion shifts or, conversely, fragment groups from within. While the results of the mean-field extension identify a set of model parameters for which a given model can plausibly simulate an opinion distribution, which is a step towards closing the gap between theoretical opinion dynamics and empirical validation. Together, these findings contribute to ongoing debates surrounding the development of polarisation in public and private spheres, as well as enhancing the relevance of opinion dynamics through connection to social science theory and empirical validation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The interest and disquiet caused by the levels of perceived political polarisation today have seen responses on multiple fronts. Polarisation has increasingly become a subject for social scientists in Europe and America (Wagner 2021; Finkel et al. 2020), launched large collaborative studies to test interventions intended to reduce hostility between political groups and antidemocratic attitudes (Voelkel et al. 2024), and been named as the "word of the year" in 2024 (Merriam-Webster 2024). The response in this thesis will be to assess the macroscopic emergence of polarisation from the microscopic opinion dynamics of large social simulations.

Models of opinion change and their study – known as opinion dynamics – is a research approach to understanding the polarisation of political views held by individuals. It is a tool that allows for the testing of theories (Geschke et al. 2019) given that opinion surveys are costly and complex, while also providing analogies to reflect upon the properties of opinion updating systems (Olsson and Galesic 2024). This formal mathematical modelling approach to understanding opinion change has applied statistical physics methods to social systems, and pointed to possible explanations for agreement and polarisation of a population's opinions by artificially reproducing salient features and trends (Castellano et al. 2009; Jusup et al. 2022). However, doubts remain on the ability of models to connect to empirical data (Flache et al. 2017) and their relevance to social science (Jensen 2019), which leads to the work presented in this manuscript.

Polarisation is a multi-faceted phenomenon in itself, so an understanding must be developed. There are multiple types of polarisation: 1) ideological stances can become distant and clustered (e.g., bimodal) among the public

(ideological polarisation, or attitude polarisation if around a specific issue; Abramowitz and Saunders 2008; Lord et al. 1979), 2) political or social groups may develop animosity between them (affective polarisation; Druckman et al. 2021), and 3) individual views on issues may be constrained by views on other issues leading to high issue alignment (partisan alignment; Jost et al. 2022). The nature of these three types of polarisation varies. Ideological and attitude polarisation are observations of a state, while affective polarisation and partisan alignment are states as well as mechanisms that can create polarisation. Polarisation, as a state, may be considered as a property of an individual (e.g., for attitude polarisation, if the position of the person moves away from centrist positions on an issue) or of a population (if the distribution of the population moves away from the center). Polarisation may also be considered as a process through which individual or collective states increase in time (DiMaggio et al. 1996). Finally, there are many ways to measure polarisation depending on data and desired insight which adds further complexity to the word (Bramson et al. 2017). Therefore, the term polarisation refers either to states or dynamical processes, of individual or collective scope, and of different types (mainly ideological, attitude, affective, or alignment).

One shortcoming in the current opinion dynamics literature is the lack of investigation into whether model simulations can produce outcomes (such as the distribution of a population's opinions) that resemble empirical data (Gestefeld and Lorenz 2023). The limited amount of research work addressing why models do not reproduce the kinds of opinion distributions seen in reality – as measured in surveys (Hetherington 2009), or inferred from social media traces (Ramaciotti et al. 2022) – will be named as *the empirical gap*. The missing connection with data can either be considered at the microscopic level of finding experimental evidence that upholds or disproves existing opinion update rules of the model (Banisch and Shamon 2022); or, at the macroscopic level of what models can produce plausible opinion distributions that match data (Carpentras 2023a) – with opinion update mechanisms that are motivated by social theory. The search for plausible simulated distributions is the approach taken up later in this thesis.

A further challenge for opinion dynamics is that there is a disconnect between the emphasis placed on affective polarisation – characterised by negative emotion towards other group identities – by social scientists (Iyengar et al. 2019; Robison and Moskowitz 2019; Turner-Zwinkels et al. 2025), and the individual-level interactions that occur in the agent-based models

of opinion dynamics (Proskurnikov and Tempo 2017). There are a variety of regimes for interaction between agents that mimic social interactions between individuals. Examples, given in Starnini et al. (2025), are one-way communication of opinion (passive communication) and two individuals discussing and adapting opinions (pairwise interactions). The common point in these communication regimes is that they treat each agent, or individual, as identical atoms without the context or group association that pervades social behaviour (Tajfel 1974).

Addressing both a lack of group identification and the empirical gap aims to improve the connection between opinion dynamics and reality. Concerning group identification, relevant research questions are: what additional insight on existing opinion dynamics research are gained by including group identification?, is the inclusion of groups then important for understanding polarisation?, and, does understanding dynamics in terms of groups, rather than individuals, offer a useful lens to analyse results? While questions related to the empirical gap are: can simulated distributions approximate empirical distributions?, and, does that give rise to models that are plausible, under certain parameters and initial conditions, for working with empirical data? This facilitates the setting of the following research objectives.

Objective One. Obtaining fine-grained temporal data of opinion change to test and calibrate models is difficult, however it remains possible to test macro-properties of opinion dynamics models, such as the distribution of opinions produced. Testing for macro-properties of models necessitates simulating models under multiple parameter combinations, to find plausible simulations, which becomes computationally intractable. A mean-field approximation is introduced to arrive at a simpler model which can be simulated across the parameter space. With extensive simulation now possible, a framework for the potential falsification of models is put forward. The hypothesis is that the model should be able to produce a simulated distribution that resembles an empirical distribution under some set of parameters. If simulations are not consistent with the empirical distribution, then it is proposed that the model is false for this combination of parameters and distribution. The result is a narrowing of the gap between empirical observations and opinion dynamics.

Objective Two. It has been identified that treating agents as identical in the individual-level interactions of models does not allow for the modelling of social group dynamics. Group formation and influence are perceived as important in social sciences due to their impact on opinion formation and

the prevalence of affective polarisation, which relies on groups. As such, perception of groups and identification with them will be introduced into an opinion dynamics model. Thereafter, updating of agents' opinions will be influenced by identification of other agents as belonging to the same group, or not. It is then possible to uncover how the treatment of others owing to group identification, rather than identical treatment as individuals, can explain the polarisation of opinions. This more nuanced understanding of how individuals perceive and treat each other connects opinion dynamics with the significance granted to affective polarisation and group identity in social science literature.

From these objectives, the following contributions have been made to the scientific community. The research that answers to Objective One is currently a pre-print article to be submitted for review to a journal. While the results for Objective Two appear in an initial article (Cassells et al. 2024b), as well as a more complete exploration of group identification in agent-based models that is currently under review (Cassells et al. 2025).

Further contributions with colleagues completed as part of the past three years of study, but outside of the scope of this thesis, include: an article assessing the dimensionality of the American political space through social network data (Ramaciotti et al. 2024); an article presenting a dataset of political attitudes of X/Twitter users (Vendeville et al. 2025); and an exploration of the relationship between international food insecurity and migration as part of a complex systems workshop (Cassells et al. 2024a).

The structure of the following chapters in this manuscript is as follows. Chapter 2 is a review of the current literature. The chapter begins by providing a foundational background, first, on the understanding of polarisation, and then, on the development of opinion dynamics. The following sections present state-of-the-art research that addresses the two areas of research that have been identified to better connect opinion dynamics with reality. Namely, the inclusion of group identification within models and the closing of the empirical gap between simulations and observed distributions.

Two chapters laying out modelling theory are next; each one will handle a different strand of the research to draw opinion dynamics closer to reality. Introducing group dynamics to an agent-based model is in Chapter 3. This is done by bestowing previously atomised agents with a recognition of group structure and ensuing differential treatment of other agents dependent on whether they are perceived as belonging to the same group, *in-group*, or belonging to another group, *out-group*.

The second theory chapter focuses on the modelling of distributions rather than individual agents to close the empirical gap. Chapter 4 therefore enables the modelling of large populations that can only be represented by distributions, and the comparison of simulated distributions with empirical distributions. The transition from modelling agents to modelling distributions is achieved by a mean-field approximation of the model and implementation with the finite volume method.

The applications of the modelling methods from Chapters 3 and 4 can be found in Chapter 5. The first half responds to the question of what effect introducing group identification has on the polarisation outcomes of the agents' opinions. The remainder of the chapter is dedicated to providing a framework that can be used to falsify opinion dynamics models for given opinion distributions and model parameters.

Chapter 2

Literature Review

This chapter will discuss the existing knowledge and research relating to the issue of polarisation of populations and the modelling thereof, into which the work of following chapters will situate itself. This review will develop by first presenting the topic of polarisation, after which the development of opinion dynamics as a tool to study opinion change and resulting polarisation will be discussed. Following this, current research gaps and challenges for the application of opinion dynamics will be highlighted – namely, social identity and the missing link between models and empirical data.

2.1 Polarisation Definitions and Measures

A broad definition of political polarisation is an increase in discord between the opinions of a population, be that an ideological lack of consensus or an emotional antipathy to stances. Such disagreement has, of course, existed and fluctuated over time. Studies have appraised the ebb and flow of polarisation (Poole and Rosenthal 1991), as well as secondary effects on the acceptance of science (Rekker 2021). Today, the predominant worry is the intensification of polarisation manifesting as a dislike for others rather than pure issue disagreement (Garzia et al. 2023).

The role and effect of polarisation is a much discussed subject for political and social scientists (Fiorina and Abrams 2008; Finkel et al. 2020; Druckman et al. 2021), as well as computer scientists and complex systems specialist alike (Conover et al. 2011; Bakshy et al. 2015; Waller and Anderson 2021; Falkenberg et al. 2022; Peralta et al. 2024). While there is not an ex-

pectation that populations should systematically display agreement on most issues – a level of polarisation may be positive for democracy to challenge the status quo, or addressing inequalities (Kreiss and McGregor 2024) – high levels of polarisation are a worry due to their destabilising impacts on public life, such as eroding national unity (McCarty et al. 2016) and obstructing collective decision-making, as with recent examples regarding public health (Gollwitzer et al. 2020). Because polarisation is such an extensive research topic, the presentation given here does not intend to cover the history of the research in this area, nor to describe the conceptual richness attached to this body of knowledge. Instead, the presentation provided in this section aims at providing an introduction to the topic and, over all, a point of contact between the main concepts relevant for this thesis and the operationalisation of terms needed in simulations presented in subsequent chapter. The rest of this section on polarisation will first provide an overview of the different types of polarisation, and then discuss how the phenomenon may be measured, to provide an understanding of polarisation prior to any modelling and simulation work.

2.1.1 Types

Given the complex nature of polarisation, as well as its wide usage, there are many categorisations to consider. A first distinction is whether polarisation is considered as a property of the state of a system or as a system process (DiMaggio et al. 1996) – either one can study the polarisation of a distribution (Fiorina and Abrams 2008), or one can study the dynamics of polarisation (Dandekar et al. 2013; Levin et al. 2021). Both interpretations are important, a measurement will be understood to address the state of polarisation and how measurements change over time will address the process of polarisation. For the purposes of this review, polarisation will be considered as occurring in an opinion space that can be multi-dimensional, where each dimension represents an issue and individuals are spread along the dimension according to their issue stance; that is, ranging from the most negative to the most positive attitudes towards a policy (e.g., liberalise immigration policy, or prioritise environmental protection over economic growth). A further high-level political science consideration is the debate as to the effective scope of individuals that may be polarised. There is argument between the view that polarisation should be considered as a population-wide phenomenon (Abramowitz and Saunders 2008) and the view that polarisation is limited

to political elites (Fiorina et al. 2008). While this is an important distinction, it will not feature in the discussion here because a difference between opinion leaders and lay people does not feature in later modelling, so the assumption is that it is a mass phenomenon.

Polarisation may manifest in different ways and as such have different meanings. A recent survey of mechanisms found in the psychology literature identifies three principal types of polarisation: ideological polarisation, partisan alignment, and affective polarisation (Jost et al. 2022). Referring to, respectively: opinion change towards the extremes of the main dimension of the opinion space, the constraint of the opinion space leading to high alignment between opinion dimensions (how few dimensions the space can be represented by, or does an opinion for one topic effectively predict opinions on other topics), and positive or negative attitudes held by members of social groups about the groups themselves. Furthermore, the survey highlights cognitive-motivational mechanisms that push individuals to defend individual beliefs ("ego justification"), in-group beliefs ("group justification"), and conservative attitudes that are in favour of the status-quo ("system justification"). A partition of the population into two modes is required in most considerations of the Jost et al. (2022) article, which reflects its US context. In a European context, opinion distributions are often multimodal (Gestefeld et al. 2022), and more multidimensional than in the US, which adds further complexity to understanding polarisation.

Different types of polarisation do not stand alone but may interact to add complexity. An example of interaction between polarisation types is partisan sorting; it is as an alignment process acting on a population to produce groups and thus has knock-on effects on ideological and affective polarisation (Mason 2015). The knock-on polarising effects may be increased affective polarisation due to a stronger identification of in-group and out-group; or ideological polarisation may increase due to individuals vacating the opinion space between groups for more extreme opinions that align with the groups. Further complexity can arise with paradoxical situations such as the experience of attitude homogeneity at an individual level and the existence of attitude heterogeneity in the social networks of the collection of individuals (Baldassarri and Bearman 2007), meaning that individuals can perceive that those attitudes which they observe in their social network are similar despite the existence of a range of attitudes within the population.

Affective polarisation has become much discussed and typically considered as particularly important to managing polarisation in a population

(Iyengar et al. 2019; Kubin and Von Sikorski 2021; Yarchi et al. 2024). It is concerned with "in-group love" or "out-group hate" (Brewer 1999) and provides the motivation for implementing group identification into models for opinion change in Chapters 3 and 5. The increase in ideologically driven group identification pushing attitudes to extremes (Mason 2018) and the impact of affective polarisation on democratic systems (Reiljan et al. 2024) make this type of polarisation the most compelling for research in current circumstance; motivating the integration with opinion dynamics discussed in this thesis.

2.1.2 Measurement

There are therefore many meanings to political polarisation, not to mention the wide range of uses of the term in further disciplines. For an overview of this multitude of measures, Bauer (2019) presents a summary encompassing income polarisation, political polarisation, and cultural polarisation, amongst others. The following discussion of polarisation measurement will centre on the political opinion context, detailing different approaches and ending with a presentation of the polarisation measure to be used in later work.

In the scope of this thesis, polarisation will be discussed as a state of social systems, foregoing the operationalisation of the notion of polarisation as a process. A first approach to measuring the state of polarisation may be metrics that provide insight into the population's distribution of attitudes towards a topic; for example, the spread of responses (perhaps measured by the standard deviation) to a question asking individuals to self-position on a scale measuring attitudes towards income redistribution (Waldrop 2021). This would be a simple measure of ideological polarisation, although this misses any idea of the groups that affective polarisation entails (Sieber and Ziegler 2019). Along this line of thought, DiMaggio et al. (1996) considers "within-group" polarisation and "between-group" polarisation – operationalised as dispersion or bimodality and spread between group means or peakedness of groups, respectively. While this provides a step forward for understanding a high polarisation scenario as having minimal within-group spread and substantial differences between groups, a drawback of this approach is that two measures are needed to assess the state of polarisation which makes how to compare two distributions unclear. For example, if standard deviation is used to measure spread and kurtosis is used to measure peakedness, in the case where a distribution is high in the first measure

but low in the second measure while another distribution is contrarily low in the first measure but high in the second measure, then it is unclear which distribution is more polarised. Furthermore, changes in the two measures may be more or less indicative of polarisation but this is unclear; a change of 0.1 in measure one is unlikely to have the same meaning as a change of 0.1 in measure two.

The most commonly discussed methods to measure polarisation are quantifying the dispersion, or spread, of a distribution and description of distribution modes, or peaks, (Bramson et al. 2017). Measures of spread used in the literature – such as *mean absolute deviation*, *standard deviation*, and *disagreement index* – have been found to correlate strongly (Gestefeld et al. 2022), so they can be used interchangeably with little difference in results.

Measures of modes, a measure of group identities as represented by a clustering of opinions in a local maximum of the opinion distribution, offer a wider variety of approaches. To assess the "peakedness", or bimodality, of an distribution DiMaggio et al. (1996) suggests kurtosis. Howver, kurtosis only measures how extreme the tails of a distribution are (Westfall 2014), and is therefore not an effective measure of multimodal distributions, at least when used without another measure (for example, variance) for further context (Downey and Huffman 2001). An alternative is to use a statistical test for uni-modality, such as Hartigan's dip test, to identify the emergence of a new opinion group, as in (Falkenberg et al. 2022), or the bimodality coefficent (Freeman and Dale 2013). However, test statistics for uni-modality provide no insight into multi-modal distributions with no indication of shape or total count of modes; that is, when there are more than two modes.

Both dispersion and modality provide measurement options that can assess ideological and affective polarisation in the terms defined by Jost et al. (2022). However, there still remains the third type of polarisation identified by the authors: partisan sorting. While this type of polarisation is less relevant to European contexts, insight into the dimensionality of the opinion space and any change to the number of highly pertinent dimensions is valuable information to understand polarisation. DiMaggio et al. (1996) suggests Cronbach's alpha to measure the constraint of the opinion space.

Further to these measures, the topic of measurement of polarisation is not settled. There continues to be research that produces novel measurements (Hohmann et al. 2023), and the understanding of the link between perception and measurement of polarisation continues to develop (Steiglechner et al. 2025). With no standard practice, apart from the use of a dispersion measure

such as standard deviation which ignores notions of groups, works continue to use a variety of measures that fit the authors' criteria.

The measures of polarisation reviewed here are relevant to opinion spaces, however it is important to note that researchers have also looked to quantify polarisation in other types of data. An understanding of these different types of measures is presented by Musco et al. (2021), defining "statistical measures" that are functions of an opinion vector and "local measures" which measure properties of network structure to identify polarisation. Further to these definitions, the authors also find that group-based measures better align with perceptions of polarisation in reality, which adds further motivation to using a measure that account for notions of group. Local network measures often require labelling of the individuals as groups within the population, followed by measuring modularity to assess network segregation (Newman 2006). If labels do not exist within the dataset then they may be assigned by community detection or partitioning algorithms, using algorithms such as label propagation (Raghavan et al. 2007) or METIS (Karypis and Kumar 1995) to assign clusters, followed by calculating modularity (Conover et al. 2011; Garimella et al. 2018). Further measures, other than modularity, for assessing polarisation can be found in Garimella et al. (2018). Other types of data may also be used to quantify polarisation, although this typically requires a numerical opinion value to be calculated on the data – for text data, sentiment analysis may be used and then difference in sentiment score analysed (Gurukar et al. 2020).

For this thesis, the measure used will be the Duclos-Esteban-Ray (DER) measure of polarisation (Duclos et al. 2006), which returns a single value assessment of polarisation as a property of a distribution of individuals on an opinion scale or dimension. It maintains the two aspects of polarisation as "within-group" and "between-group" set out by DiMaggio et al., instead termed as an "identity-alienation" framework, and so incorporates the importance of groups within polarisation. The measure is derived as an axiomatic theory that accounts for both local concentration (representing within-group/identity notions of polarisation) and distribution spread (representing between-group/alienation notions of polarisation) to give,

$$P_\alpha(f) \equiv \int \int f(x)^{1+\alpha} f(x') \|x' - x\| dx' dx.$$

where x and x' are points in the opinion space, $f(x)$ is the density at x, and α is a parameter between 0.25 and 1 (further details of the calculation and

Measure	Definition	Data	Use
Standard deviation	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$	Vector	Opinion spread.
Mean absolute deviation	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i-\bar{x}\rvert$	Vector	Opinion spread.
Disagreement index	$\frac{1}{n(n-1)}\sum_{i\neq j}\lvert x_i-x_j\rvert$	Vector	Opinion spread.
Kurtosis	$\frac{\frac{1}{n}\sum(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum(x_i-\bar{x})^2\right)^2}$	Vector	Peakedness.
DER	$\int\int f(x)^{1+\alpha}f(x')\lVert x'-x\rVert dx'dx$	Vector	Identity and alienation.
Hartigan's dip test	$\inf_{G\in\mathcal{U}}\sup_x\lvert F(x)-G(x)\rvert$	Vector	Multimodal test.
Bimodality coefficient	$\frac{\gamma^2+1}{\kappa+\frac{3(n-1)^2}{(n-2)(n-3)}}$	Vector	Bimodality indicator.
Cronbach's alpha	$\frac{m}{m-1}\left(1-\frac{\sum_{k=1}^{m}\sigma_k^2}{\sigma_X^2}\right)$	Vector	Space constraint.
Modularity, Q	$\frac{1}{4E}\sum_{i,j}\left(A_{ij}-\frac{d_id_j}{2E}\right)\delta(c_i,c_j)$	Network	Community structure.

Table 2.1: Polarisation measures discussed with definitions, data type, and an intuitive interpretation. The notation is as follows: x_i is an opinion of individual i, $\bar{x}$ is the mean opinion, n is the population size, σ^2 is the variance of opinion, γ is the skew of opinion, κ is the kurtosis of opinion, $f(x)$ is the density at x, $F(x)$ is the discrete cumulative distribution of observations, $G(x)$ is the cumulative distribution of the unimodal function that minimises the maximum difference with $F(x)$, m is the number of dimensions of the space, k is a dimension of the space, σ_k^2 is the variance of dimension k, σ_X^2 is the sum of the entries of the covariance matrix, E is the number of edges in the network, A_{ij} is the adjacency matrix of the network, d_i is the degree of node i, $\delta(c_i,c_j)$ indicates if node i and j are members of the same community. 'Vector' refers to a measure of opinion in the opinion space, while 'Network' refers to a network representation of the individuals.

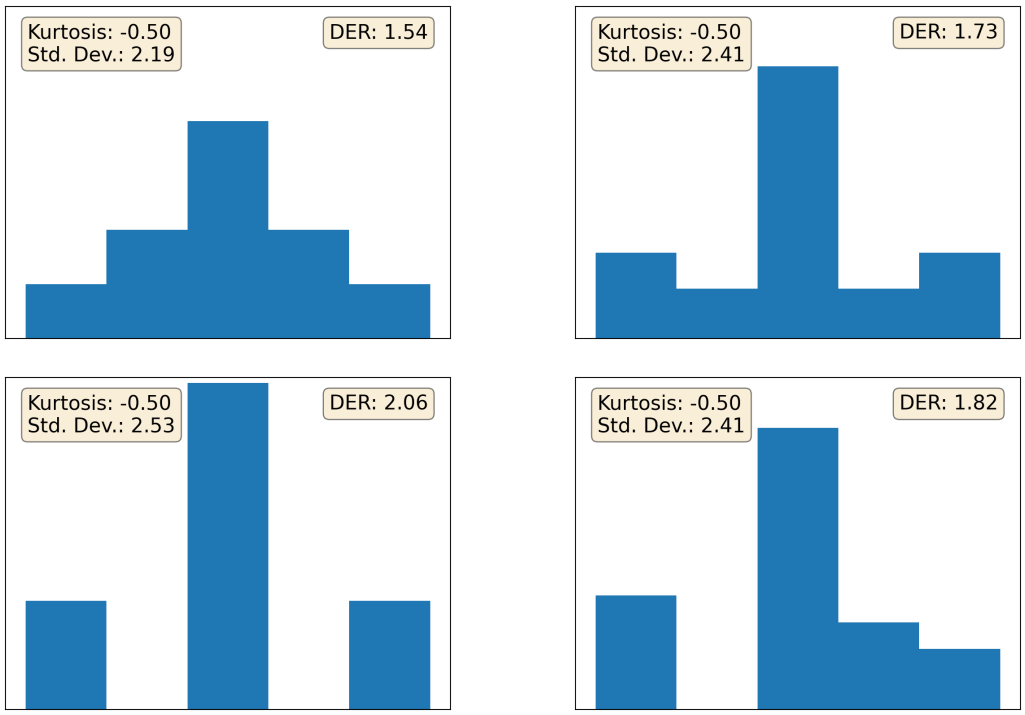


Figure 2.1: Comparison of DER as a measure of polarisation of a distribution against the combined use of kurtosis and standard deviation. Density is bucketed in intervals of size 2 centred at $x_i = 1$, 3, 5, 7, and 9. Kurtosis is the same across all panels so provides no insight, while standard deviation is the same for the two right hand panels implying that the two distributions are equally polarised if using kurtosis and standard deviation to measure polarisation. In contrast, DER values measures polarisation as different across each of the distributions.

implementation, such as approximating the density, can be found in Section 3.4). An increase in α increases the importance of identification within the identity-alienation framework. Given that this measure has rigourous axiomatic foundations, is comparable across distributions with a single-valued measure, and takes account of polarisation as both inter- and intra-group, the DER measure is determined to be the most appropriate measure when discussing polarisation of opinion distributions and thus selected for later comparison of simulated distributions as they change over time. A comparison of DER with both kurtosis and standard deviation is in Figure 2.1.

2.2 Opinion Dynamics

The study of models of opinion change developed out of the theory of statistical physics, with particular focus on the possibility of system-level – *complex* – phenomena emerging from the simple pairwise interactions of individuals (Castellano et al. 2009). The principal methods associated with the models are the study of system transitions between disorder and order (*e.g.* from a spread of opinions to a consensus), the concepts of steady states (do opinions stay in a consensus?), and agent-based modelling (computer simulations of individuals in a population under some set of rules).

Early examples of agent-based models, which were unrelated to opinion modelling and typically termed "cellular automata based models" at the time, included simplified models of self reproduction by John von Neumann and Sanislaw Ulam in the 1940s, and the game of life by John Conway in 1970 (Hegselmann 1996). A review of some principles and features of the complex behaviour emerging from a system of interacting automata/agents was then presented by Wolfram (1984). Applications of such models in the context of opinions were also developing, which Section 2.2.1 will discuss. There are comprehensive reviews of the applications of physical models in various social settings (Jusup et al. 2022), in the context of political opinions (Starnini et al. 2025), and in game-theoretic extensions (Szczepanska et al. 2022), providing extensive overview of the field.

Two primary points of value for opinion dynamics models is that they provide a method that can explain the phenomena produced given the simple and understandable interaction mechanisms that underly the model, as well as produce predictions through the lens of scenarios since system state changes can be induced by altering the model set-up. The combination of explainability with predictability is a valuable tool for understanding the world around us in computational social sciences (Hofman et al. 2021). It is important to hold in mind that the models are ultimately analogies that provide a trade-off between the "conceptual mileage" of enabling the study of system properties with new insight, and the "conceptual baggage" of the limits of modelling assumptions (Olsson and Galesic 2024).

2.2.1 Development of Models

Extensive reviews of a range of opinion dynamics models can be found in (Proskurnikov and Tempo 2017, 2018). A brief overview of the develop-

ment of models will be presented here. An initial categorisation of models is whether they concern agents with opinions that are continuous or discrete (Noorazar 2020); that is, whether opinions are a binary choice of yes-or-no agreement or if opinions exist in a space such as a scale running from zero to one with values in between representing a degree of agreement with either extreme.

The foundation of continuous opinion models is the idea of social influence (which also appears in categorical or binary opinion models), whereby an agent typically updates their opinion by averaging with the opinion of their neighbours. The precise mechanism of averaging may vary, but the starting point is a model such as the DeGroot model (DeGroot 1974) which states that for a vector of agents with one opinion each, x, at time, t, with a weight-matrix, W, representing the influence between any agent i and j, then opinions at time $t + 1$ follow the update rule $x_{t+1} = W \cdot x_t$. The closeness-to-reality and usefulness of the model was quickly questioned since it always converges to consensus if the graph of the influence network between agents is complete and aperiodic in the case of directed graphs, that is, there exists a path from one agent to another in the network and opinions to not propagate around the network in a repetitive manner.

Improvements on the DeGroot model were made in the Friedkin-Johnsen (Friedkin and Johnsen 1990) – which introduced stubborn agents who do not change their opinion – and the Hegselmann-Krause (Hegselmann and Krause 2002) model – implementing a concept of homophily through agents only updating their opinion when another agent's opinion is within a limit of difference to theirs, termed "bounded confidence". While opinion updates are synchronous in the Hegselmann-Krause model formulation, a pairwise implementation of bounded confidence is found in (Deffuant et al. 2000). Subsequent works build upon these foundations, introducing social psychology mechanisms such as biased assimilation (Dandekar et al. 2013), meaning that agents adopt a tendency to support their initial opinion when faced with inconclusive evidence. Dandekar et al. shows that if this bias is sufficiently strong then the opinions will polarise, and that this is not possible under homophily alone.

Each of the foundational DeGroot, Friedkin-Johnsen, and Hegselmann-Krause, models consider social influence as a purely positive mechanism, as a process of opinion assimilation. There is also a body of literature that considers negative social influence, also known as a *backfire effect*, which creates polarised system states (Jager and Amblard 2005; Chen et al. 2021;

Axelrod et al. 2021). A model of this type, with attractive and repulsive social influence, is the primary model considered in the following research sections of this thesis – a complete description can be found in Section 3.2.

Discrete-valued opinion models are not used in the work of this thesis but for completeness of this review they are listed here. Three model types may be considered in this context: (1) the Galam model, which treats a binary agree/disagree opinion space in which agents are shuffled into groups and each model iteration the majority group opinion is adopted by the group's agent (Galam 2008); (2) the Sznajd model, which is an adaptation of the Ising model from statistical physics (Sznajd-Weron and Sznajd 2000); (3) the voter model, for which agents chose a neighbour and copy their binary opinion (Redner 2019). A development to further these discrete models has been to add a hidden layer of argumentation for each agent which influences the opinion expressed, known as argument communication theory (Mäs et al. 2013). For the interested reader, recent applications of these model types can be found in (Banisch and Olbrich 2021; Vendeville 2025). The choice to not use a discrete-valued opinion model is to be able to connect models with opinion scales rather than actions or choices.

2.2.2 Current Challenges

The existing collection of opinion dynamics models still have unanswered challenges of how to better reflect the *observed* state of opinions in society around us (Sobkowicz 2020). That is, there is a disconnect between real-world opinion distributions that neither resemble consensus nor complete polarisation but a point between the two extremes, while theoretical models tend towards these two extremes. What combination of changes to the modelling process will result in a narrowing of the empirical gap between theory and observation is an open question, since the need for understanding of collective behaviour remains highly important for policymakers and regulators that wish to manage social media, recommendation systems, or content moderation (Bak-Coleman et al. 2021).

Two principal challenges in the opinion dynamics literature (bestowing agents a group identity and connecting models to empirical data) are the subject of discussion in Sections 2.3 and 2.4, because they motivate the work of later chapters, while a brief overview of challenges raised in the literature is given here.

Some have questioned the fundamental relevance of physics-inspired mod-

els of opinions and other social phenomena to real social systems has been questioned, particularly given that individuals are atomised as agents and considered identitical despite this being far from reality (Jensen 2019). For instance, age, gender, or education, might determine how individuals process an interaction with the opinions of others but these are lacking in a model of identical agents. Further to these individual characteristics, it is known that indiviudals perceive social systems and incorporate resultant groupings into judgements of others. One response to this criticism is to bestow agents with a notion of group identification, which defines how agents relate to the wider population and influences the interactions with one another. Current computational implementation and social science grounding of identity is detailed in Section 2.3.

There is also criticism that many works focus on theoretical variations and further study of existing models, while connecting opinion dynamics models with empirical data remains uncommon (Flache et al. 2017). This is perhaps due to differing goals of theoretical agent-based modelling and empirical research (Carpentras 2023b), but the additional value of empirical validation when presenting results to a wider audience is clear. An additional complication of connecting to empirical datasets is that the rules of agent-based models often rely on pairwise interactions, which entails computational time growing quadratically with population size. Given that social systems of interest are usually large and need multiple model parameter combinations to be tested, it quickly becomes infeasible to explore agent-based models for large populations. A solution to this problem is to use a mean-field approximation to reduce the computational cost of the model, this will be explored later in the thesis in Chapter 4. Further discussion of existing literature that connects models to data can be found in Section 2.4.

The functioning of models in a multi-dimensional opinion setting is another frontier of opinion dynamics, although not one that will be treated in this thesis. Novel scenarios can be tested in the multi-dimensional case; for example, if an individual is anti-immgration on an immigration opinion dimension and pro-environmental protection on an environmental opinion dimension and this individual encounters another with anti-immigration and anti-environmental stances, then it is unclear how they interact. The individuals could update their opinions dimension-wise or simultaneously across both, and the knock-on effect on partisan sorting (alignment) of the opinion space is a key area to understand – with particular relevancy in multi-dimensional European systems.

2.3 Group Dynamics

The domain of this thesis is computer science, however research into group identification requires a foray into interdisciplinarity – more precisely, into social sciences – in order to encourage relevancy of model applications. In this section, social science grounding of group identification is presented, followed by connections to computational models and group/community detection concepts from computer science.

2.3.1 Social Identity and Social Boundaries

In the realm of social psychology, a seminal text on social identity theory is Tajfel (1974), explaining how social class, religion, and nationality, define group memberships that are used by individuals to categorise themselves and others. The resulting categorisation can spawn in-group favouritism and prejudice for out-groups. From the emergence of the theory, group identity has spawned many works and thus making it a central consideration to intra- and inter-group processes (Hornsey 2008). For example, research has considered the potentially dangerous implications of depicting an in-group as "virtuous" while an out-group is considered as a "threat" (Reicher et al. 2008). The authors identify a cycle whereby inhumane acts to others are accepted after a cohesive in-group is formed and then an out-group is excluded and considered as dangerous to the uniquely virtuous in-group, perversely giving moral strength to those that punish others. Furthermore, identification with a group may introduce a set of norms that are expected to be shared within the group, such as values or behaviours, that influence individuals' own perception of those values, thus having relevant consequence for how agents might perceive the opinions of others (Masson et al. 2016).

More broadly in social sciences, groupings also have an important role in the literature. They may be formalised as a symbolic boundaries used by actors to conceptually distinguish between objects, people, and practices, (Lamont and Molnár 2002). For example, structures defining the similarites and differences between male and female genders constrain and form their behaviours and attitudes (Gerson and Peiss 1985). Social boundaries can exist across many issues, characteristics, and lifestyle choices creating potentially complex social identities, but there is evidence that mechanisms such as homophily can reinforce alignment between lifestyle choices and political ideology to result in well-recognised identities, such as the infamous "latte

liberal" displaying a stereotyped affinity between hot beverage choice and ideology in the US context (DellaPosta et al. 2015), creating simplified identity structures from the convergence of multiple groups (Roccas and Brewer 2002). Furthermore, cleavage theory defines divides within a population – such as GAL (green, alternative, libertarian) on one side and TAN (traditionalist, authoritarian, nationalist) on the other – along which political opinions show clear divergence (Marks et al. 2020), as socially structured opinion groups underpinned by a functionalist logic. Because perceptions of groups hinge on social cues related to social divisions, identity is not limited to lifestyle choices and physical manifestations, it may also be expressed in virtual contexts (Santagiustina et al. 2025) and therefore has potential impact on social media debate. This connection with digital environments makes grouping relevant to one of the most promising empirical domains for agent-based simulations: social media.

Ultimately, identification with a group is a precondition for the emergence of affective polarisation of political opinions, in the multiparty systems of Europe (Wagner 2021) as well as the well-known US case (Iyengar et al. 2012). Evidence further points to the influence of prior attitudes and beliefs (formed by a social identity) influencing the evaluation of new information (Taber and Lodge 2006), the modelling of which provides further impetus to include a notion of group identification in opinion dynamics.

2.3.2 Computational Approaches to Groups

The importance of group identification in opinion formation and communication, have led to research interest into how to develop opinion dynamics models to operationalise the concept. One approach has been to extend pairwise interactions between two agents to higher-order interactions by considering a *hyper-edge* that connects multiple agents at once (Battiston et al. 2020). This has been implemented, and labelled as group interactions, in an opinion dynamics context (Iacopini et al. 2022; Pérez-Martínez et al. 2025), however the result is that agents consider dynamics at a widened local level rather than identification with a group that might not be topologically bounded (in principle, a individual may identify another one as in-group, regardless of the shortest path length between them).

Existing computer science techniques such as clustering of data points (Rodriguez et al. 2019) and community detection within networks (Fortunato 2010; Schaub et al. 2017) developed without a social science motivation

but the goal of identifying groupings, or patterns, within a dataset/population operationalises a similar notion on either geometrical grounds for an opinion space, or topological grounds for a network structure. Given the differing motivation, however, caution should be taken to ensure that this analogy is appropriate in each social setting. This will be the object of careful consideration in Chapter 3.

Using a computational group recognition method to define groups within a population and then apply this identity for an opinion dynamics model is yet to be widely investigated, and this avenue of research will be taken up in Chapter 3 to extend an existing opinion dynamics model in the research that follows. An alternative to clustering/community detection is presented by Salzarulo (2006), who uses the difference between the average opinion of close-proximity neighbours against those neighbours that are outside of close-proximity to define group identities, and Yang et al. (2021) follows a similar line of thought; both methods require agents to reference a prototypical average opinion of an evolving group. Modelling evolution of social identity through pairwise agent interactions is another option (Törnberg et al. 2021), but does not rely on perception of the population.

2.4 Empirically Grounded Simulations

The second challenge identified within the discussion of opinion dynamics was the perceived empirical gap: research work tends to cover simulations of theoretical models while lacking a connection to data that represents opinions. This section will first cover existing work that makes the connection between models and data, and then discuss methods to use large datasets with existing models which will provide a preliminary discussion for modelling theory found in Chapter 4.

Agent-based models typically make connections to the real world by utilising scenarios: if model parameter x increases, which represents real-world phenomenon y, then polarisation of opinions occurs. When connection to datasets is made, it is to those with small population sizes, such as surveys, given that resources needed to simulate agent-based models tends to scale badly with increasing population sizes. This unfortunately discounts the use of large online datasets of interest, and relies on expensive surveys for which it is difficult to follow on consistent population over a period of time to have accurate data about opinion changes of individuals.

The desire to add empirical weight to the theoretical models of opinion dynamics is an ongoing process, presenting difficulties that will be further discussed in this section. A wider outlook on linking data with theory in network science can be found, for instance, in the work of Peel et al. (2022), while the survey of Peralta et al. (2022) treats opinion dynamics in social networks. Calibrating existing models with experimental data is one area of focus, while providing evidence supporting the relevance of mechanisms that justify their inclusion in computational models and simulation is another. The focus on validating mechanisms provides insight into whether the model is relevant to social systems, or not.

The foundation of many opinion dynamics models is social influence, which understands individual opinion change as a result of the observation of other individuals' opinions through social interactions. There is evidence to strongly support the claim that a social influence mechanism exists and is a process by which ideas or opinions can spread within a population (Moussaïd et al. 2013; Carpentras et al. 2022). Further results confirm that similarity in opinion does induce attraction (positive opinion change towards the opinion of another) between agents, while support for negative influence (increasing distancing between opinions) at the pairwise level is inconclusive in the work of Takács et al. (2016). However, there is evidence that negative influence, or divergence, does occur in political establishments and on social media from colleagues in social sciences (Liu and Srivastava 2015; Bail et al. 2018). Given that models typically take initial inspiration from social psychology principles, it is promising confirmation that the mechanisms are indeed observed, despite some uncertainty on negative influence at the pairwise level within the opinion dynamics community. Attention is now shifted to validation of model outputs, rather than validation of the functions and mechanisms within them.

2.4.1 Calibrating and Validating Models

Validating opinion dynamics models with experimental data is not a straightforward task, with challenges arising from both the nature of models and the availability of data. On the modelling side, determining the extent to which inclusion of certain parameters increases the realism of the model is difficult. Parameters in models can take on several different roles. One parameter may represent some combination of processes: a parameter governing the magnitude of opinion change each iteration in a model represents a simplification

of a complex process. Another parameter role is as a proxy for a mechanism; such as a parameter governing how agents of different opinions are exposed to each other, which may be considered as a proxy for a simple recommendation algorithm. Finally, some parameters represent abstract concepts such as open-mindedness, which can be difficult to operationalise into a measure since it is unclear how an individual's level of tolerance or acceptance of different opinions can be measured in practice.

On the data side of the validation task, the difficulty of parameter measurement hints at the problems related to finding appropriate datasets where fundamental questions arise such as: is an opinion measurable, how can a given population be measured consistently over time, and how can empirically measured opinions be mapped into an opinion space?

General methods of incorporating data into agent-based models, non-specific to opinion dynamics, is presented by Windrum et al. (2007) and Bruch and Atwell (2015), while the measurement error of opinions in relation to opinion dynamics is discussed by Carpentras and Quayle (2022). In the case of argument-based models, Banisch and Shamon (2022) and Banisch and Shamon (2025) present a validation framework in which the authors calibrate parameters for a model grounded in argument communication theory by using survey experiments centred on arguments and debate related to electricity generation and the climate. A related example for modelling attitudes and calibrating with data is provided by Brousmiche et al. (2016). Both of these strands of research use models with foundations in psychology and so the model calibration becomes evident with a theoretical structure explaining how arguments are evaluated to result in an attitude.

However, such frameworks do not readily extend to models where opinions are represented by a continuous variable with no underlying argument appraisal process. The distinction between methods that calibrate small-scale mechanisms and those that test model predictions of opinion distributions is discussed further by Starnini et al. (2025).

Surveys are possible form of data with which to test models since individual responses are followed and at a small population size that maps well to the number of agents in an agent-based model. A promising approach to bridging the gap between discrete survey responses and continuous opinion variables is presented by Carpentras et al. (2023). In their experimental design, participants were asked to indicate agreement or disagreement with a statement, as well as their certainty on a scale from 1 to 10. This certainty score serves as a weight, transforming the binary judgement into a

graded scale that ranges from strong disagreement to strong agreement. The mapping provides responses that more closely resemble the continuous scale, opening a path to calibration when survey responses are collected under different scenarios or over time. The drawback here is that the size of survey experiment populations is usually small and that surveys are costly, and so is not appropriate for large, online, datasets. Finally, a fundamental drawback of survey-based research is that it is exceedingly difficult to link survey responses with individuals for whom there is data on exposure to each other's opinions and longitudinal assessment of opinion change.

Finally, another method to incorporate data is to ask if an opinion dynamics model can reproduce observed distributions. It is then possible to assess by goodness-of-fit measure whether the model comes close to reality, or not (Gestefeld and Lorenz 2023). The limiting factor in this approach is it requires search of a parameter space which is computationally costly for agent-based models. A solution to the computational cost is presented in this thesis, by using approximation and numerical simulation methods – an overview of which is presented next.

2.4.2 Using Large Datasets

The reliance of opinion dynamics on agent-based models means the inheritance of several limitations of agent-based simulations. Principally, the requirement to follow individual agents and each of their interactions becomes resource-intensive as population sizes grow, which poses a problem when applying the models to large socio-informational networks. The resulting simulations are slow because of the huge number of agent-agent interactions that the model calculates since the number of interactions scales quadratically with the number of agents, which makes it difficult to calibrate or validate models for a broad parameter space. This problem also arises in the field of statistical physics – for example, attempting to model individual gas particles becomes intractable as the number of particles increases – and has led to the development of tools such as mean-field approximation and numerical simulation methods to treat large systems. The resulting object of study for these methods is an opinion distribution and its evolution, rather than the evolution of an individual's opinion. Linking between a discrete number of agents and a continuous distribution representing a population is achieved by considering how the system behaves as the size of the population tends to infinity, or what is the average behaviour. The goal of being able to

model distributions is then to enable comparison to real-world populations at a macro scale (Kozitsin 2022).

Using a mean-field approximation for large populations of interacting agents has been studied for a bounded confidence model to provide analytical insight (Dubovskaya et al. 2023), the Deffuant model (Fennell et al. 2021), and to infer an interaction kernel (how individuals holding different opinions interact and impact each others' opinions) for the model (Chu et al. 2024). In fact, the work of Chu et al. (2024) finds that the error in the inferred interaction kernel decays as the dataset is enlarged, which points to the value of treating large populations. A general formulation for the mean-field limit of opinion dynamics models relying on some interaction function exists (Ayi and Duteil 2021).

Some existing research suggests that the size of a system has a role in the eventual dynamics of that system (Toral and Tessone 2006) and so by taking a large population limit these effects are ignored. More precisely, the authors take the case of a finite agent-based Galam model which has a critical value that determines transition between polarisation and consensus and show that the critical value varies with population size. The criticism is valid, however it is only relevant for the case of modelling a specific system rather than the general case that mean-field limits imply, with the criticism also standing for the arbitrary number of agents selected for an agent-based model since the population size will influence precise thresholds.

The averaging of behaviour across large populations does have the drawback of losing the detail of the complex and finite interactions that are a fact of social systems. Therefore these approaches do not replace agent-based models but they provide an approach to handle large datasets and thus provide novel results that agent-based models fail to achieve, particularly for simulating the large populations present in social media datasets. Full detail for the theory and implementation of the mean-field approximation and numerical simulation will be discussed in Chapter 4.

Chapter 3

Group Dynamics in Agent-Based Models

3.1 Motivation

Group dynamics in online polarisation has received significant attention in recent years. Identification as part of a group is integral to the concept of affective polarisation and therefore prominent in the polarisation research of political science (Iyengar et al. 2012). Similarly, the role of social identity – influenced by group identification – in opinion communication and formation is well accepted in social sciences and social psychology (Tajfel 1974; Lamont and Molnár 2002), which has a resulting impact on the opinion polarisation of a population. However, most existing work in agent-based models focuses on pairwise interactions or higher order interactions and ignores the notion of group identification at a population-level. It is therefore an open and highly relevant question as to how group identification fits into opinion dynamics. To address this problem, an existing opinion dynamics model is chosen and then extended to bestow agents with a concept of group identification that will influence how the population interacts by recognition of others as either in- or out-group.

Out of the numerous existing opinion dynamics models that were detailed in Chapter 2, the Attraction-Repulsion Model (ARM) (Axelrod et al. 2021) is selected for the work that follows. The allowance of the model for interactions that may be either positive or negative accommodates studies such as the surprising results of Bail et al. (2018) that negative interactions may

occur when cross-cutting content conveying different opinions is presented. It is therefore desirable that an agent may be drawn towards an opinion (attraction) that is observed or become further entrenched in their own opinion by distancing themselves in the opinion space (repulsion). Furthermore, the model parameters allow investigation of three behavioural aspects: *exposure* (are distant opinions frequently observed?), *tolerance* (are distant opinions attractive or repulsive?), and *responsiveness* (are opinion changes incremental, or large?); each with their own impact on polarisation. This enables investigation, in the terms of the model, into questions pertaining to existing research that suggests exposure to different content can challenge held views (Pettigrew and Tropp 2006), or even lead to compromise (Mutz 2002), alongside analysis of model simulation results.

The extension of the model to include a group identification that, in turn, influences agent interaction continues a strand of polarisation research which finds that prior attitudes or knowledge influences evaluation of new information (Taber and Lodge 2006). While the Attraction-Repulsion Model is the chosen model for this work, it should be noted that the framework introduced to bestow group identification upon agents is not exclusive to the model. The implementation adapts model parameters so that agents use either an in-group version or an out-group version, depending on how they recognise the other agent with which they are interacting. This manner of including group identification is therefore not unique to the chosen model and could be used on the parameters of alternative opinion dynamics models.

3.2 Model Description

There are three elements to the model: (1) an exposure rule determining how the opinions of agents are exposed to each other; (2) an opinion update rule on the pairwise interactions which determines the change in opinion of an agent; and (3) a group identification rule by which agents are assigned a group label. The group identification rule is the extension to the existing model by Axelrod et al. (2021) in order to create a framework to study affective elements of polarisation. When each rule is presented, it is done so from the perspective of an agent i. That is to say, the rules answer the following questions: how does agent i interact with others?, how does agent i update their opinion, and how does agent i get a group label?

Before defining the three rules that govern how the system changes over

time, the population of agents must be defined. Each agent $i \in \{1, \ldots, N\}$ has an opinion x_i that lies on a continuous scale between 0 and 1. In real world context, this may be conceptualized as an agent's position on an Authoritarian-Libertarian scale. The vector $\boldsymbol{x}$ is also defined, having opinion x_i in the i-th position. For the modelling purposes presented here, the opinion space will be one-dimensional but the model may be easily extended to a d-dimensional space $\mathbb{R}^d$ to explore multi-dimensional problems.

Unless explicitly stated otherwise, the opinion difference between agents i and j refers to the L^2-norm of x_j minus x_i, written $\|x_j - x_i\|$. The opinion difference is used to preferentially interact with agents having opinions that are closer to their own (smaller opinion difference) in the exposure rule, as well as to decide whether the interaction is attractive or repulsive and if the magnitude of opinion change in the opinion update rule.

Three parameters are included in the model by Axelrod et al. (2021) – exposure, E; tolerance, T; and responsiveness, R – that are adjusted with resulting effects on the opinion polarisation of the agent population; precise details are given below. It is on these model parameters that the impact of group identity is tested. Each parameter is taken in turn and split into two sub-parameters, with one sub-parameter reflecting the treatment of another agent that is identified as *in-group* and the other sub-parameter used when the other agent is identified as *out-group*.

By systematically splitting one of these parameters into separate values for in-groups version and out-groups, while maintaining that the other two parameters are blind to group labels, there are three parameter sets created for the model. The first set, $\{E_{\text{in}}, E_{\text{out}}, T, R\}$, considers the case when agents' exposure to in-group and out-group differs while tolerance and responsiveness remain constant. The second, $\{E, T_{\text{in}}, T_{\text{out}}, R\}$, considers when agents are more, or less, tolerant of opinion difference dependent on group recognition. The final, $\{E, T, R_{\text{in}}, R_{\text{out}}\}$, is the case when agents have stronger, or weaker, responses when faced with an opinion that is in-group or out-group.

3.2.1 The Exposure Rule

The probability of agent i is exposed to the opinion of agent j, w_{ij}, is inversely related to the opinion difference between those agents. In other words, exposure becomes more likely when x_i is closer to x_j. Probability of interaction is 1 when $\|x_j - x_i\| = 0$, the probability then decreases as opinion difference increases. This reflects the concept of homophily whereby individuals are more

likely to interact with those that they are similar to, widely acknowledged as a factor of exposure in the literature (McPherson et al. 2001). However, it is still possible for any agent to to be exposed to any other agent, albeit only a very small chance under certain model conditions, which reflects scenarios in reality such as the possibility of chance encounters or forced connection via family or workplace. The scalar value of 0.5 is kept from the original model, changing it would change the rate of decay as chance of exposure decreases with opinion difference increasing.

$$w_{ij} = 0.5^{\|x_j - x_i\|/E}. \tag{3.1}$$

The exposure parameter, E, in Equation 3.1 acts as a shape parameter on the interaction probability; the parameter controls whether an agent is relatively more or less likely to interact with other agents that have more dissimilar opinions. A large value for exposure increases the likelihood that agent i is exposed to an agent with which there is a large difference in opinion. The influence of E on interaction probability is shown in Figure 3.1 for a range of parameter values. If the opinion difference is equal to E, then agents should interact once in every two possible interactions (an interaction probability of 0.5); if opinion difference increases to be equal to $2E$, then there is an interaction probability of 0.25; in the other direction, an opinion difference equal to $E/2$ is an interaction probability of 0.71.

In the model, w_{ij} is treated as a probability with which exposure occurs rather than as an edge weight on a complete graph between the agents. The difference is that in the model simulation procedure interaction either happens or not, while under a different formulation w_{ij} could be considered as an influence weight on the outcomes of interactions between agents in a fully connected network. It is a modelling choice, and the binary outcomes forced by considering w_{ij} as a probability of interaction occurring or not at all, is the method chosen here, as well as in the original implementation, to reflect the real world nature of interactions.

3.2.2 The Opinion Update Rule

Once an interaction occurs, the opinion of agent i is updated by the function $\phi(x_i, x_j)$, where both R and T are model parameters and the opinion difference between agents i and j is again used. The opinion update may be attractive or repulsive: either the agent's opinion will become more similar

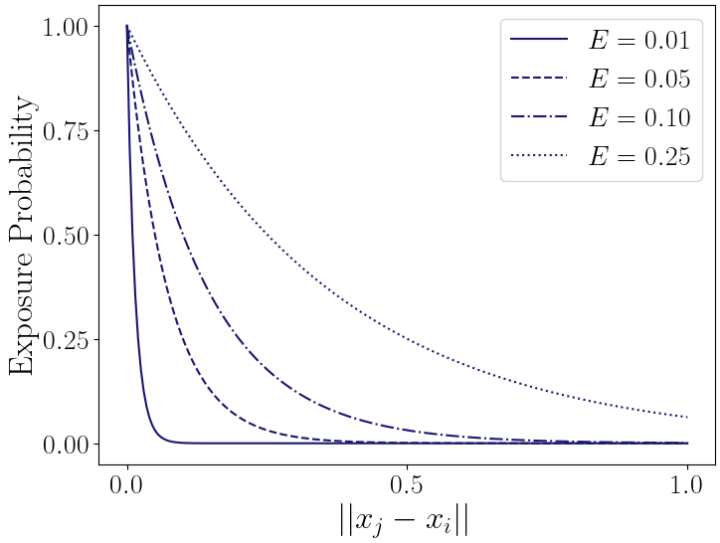


Figure 3.1: The probability that agents i and j interact as a function of the opinion difference between them, under different values of the exposure parameter E.

to the opinion x_j observed (reducing opinion difference) or less similar (increasing opinion difference). This allows for the possibility that agents may experience a 'backfire' effect, as well as the traditional averaging to consensus effect.

$$\phi(x_i, x_j) = \begin{cases} R(x_j - x_i), & \text{for } \|x_j - x_i\| \leq T, \\ -R(x_j - x_i), & \text{for } \|x_j - x_i\| > T. \end{cases} \tag{3.2}$$

Examples of the two interactions are presented in Figure 3.2.

Tolerance, T, is the parameter that controls whether the interaction is attractive or repulsive. Opinion differences below the tolerance threshold result in opinion x_i moving closer to opinion x_j, while the opposite occurs when the opinion difference is above the threshold T. Note the order of $x_j - x_i$ in Equation 3.2 to ensure the correct direction of opinion change. If $T = 1$ then interactions will always be attractive since the difference between the extremes of the opinion space is within the threshold. As T decreases, agents become less tolerant and so the range of opinions with which they disagree, causing repulsion, increases.

The second parameter in the update rule is responsiveness, R, which controls the fraction of the opinion difference by which the updating agent

i will change opinion. Consider a case in which $\|x_j - x_i\| \leq T$, then the two extremes of behaviour for agent i are either to completely adopt x_j ($R = 1$) or to completely ignore the observed opinion despite being within the tolerance threshold ($R = 0$). In the repulsive case, this same logic is applied although it is impossible to 'adopt' opinion x_j since the opinion change is in the opposite direction. Responsiveness may be considered as the speed with which opinions change.

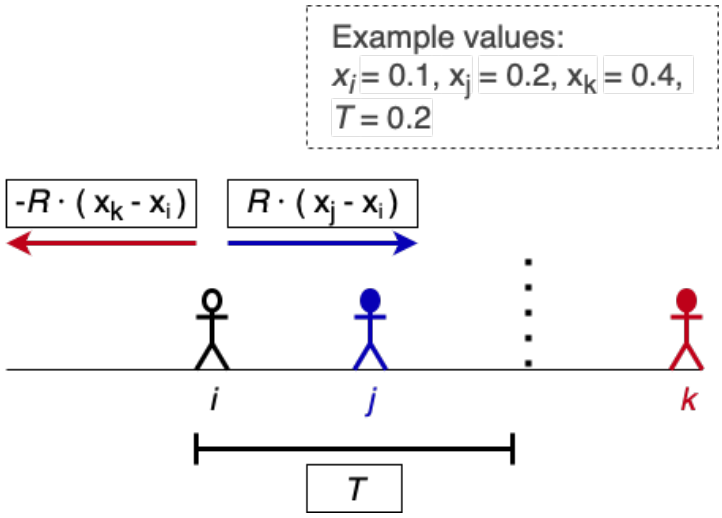


Figure 3.2: Illustrative example of the possible interactions between agents. The interaction between agents i and j is attractive since the opinion difference is 0.1 and so is within the tolerance threshold of 0.2. While the interaction between agents i and k is repulsive since the opinion difference is 0.3 and so is greater than the tolerance threshold.

3.2.3 The Group Identification Rule

Identifying groups in social systems and operationalising them in an agent-based model is a complex and novel problem. In the context of the model, agents are defined by their opinion positions and so this attribute is used to determine agents' perceptions of each other. Perception of groups on the bases of variable values (in this case, opinion positions) intuitively resembles the notion of spatial clusters. So, through the lens of a computer scientist, the problem becomes clearer in that it can be approached as a clustering

problem. Therefore the approach taken in this work is to apply an exisitng clustering method, HDBSCAN (McInnes et al. 2017), to the opinion distribution. The method is dependent on the shape and density of the distribution to which they apply, so agents positioned in, or close to, a dense part of the opinion space are identified with the same group label. A discussion of the implications of implementing group identification in this manner, and possible alternatives, may be found in Section 3.3.

Given that the group identification rule and the exposure and opinion update rules, of Equations 3.1 and 3.2, depend on the opinion distribution, the question arises as to possible interdependence between the three rules. Opinions influence the exposure of agents to other agents, and exposure then influences how opinions update, while opinions also determine group identification which influences agent interactions. All rules depend on opinions and hence there is some interdependence between the rules but they are not the same since each rule treats opinion in a different manner. As an example of the difference between the rules, consider that agents i and j have the same group label but are situated at either edge of the spread of the group's opinion, then it is possible that the opinion difference between them is above the tolerance threshold T. At the same time, another agent k may exist that has a different group label but the opinion difference with i is less than the tolerance threshold. This possible scenario occurs in the case of Figure 3.3, which shows an example clustering by HDBSCAN. Finally, the exposure rule can expose agents to others across the opinion space even if they are not in the same group. The present work takes the first step into models that consider the dependence of groups on opinion, future work may consider a further variable, in addition to opinions, upon which identification would hinge.

Identification by clustering, using HDBSCAN, centres on finding patterns in the data based on local density. It works by first finding the distance to the k-th nearest neighbour for each agent (where k is the minimum group size), named as the *core distance* for a parameter k, which is akin to transforming the space to reflect density. The next step in the method is to calculate a so-called mutual reachability distance, defined as

$$d_{\text{mreach}-k}(a, b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a, b)\},$$

where $d(a, b)$ is a metric distance between a and b (also used to calculate the cores). Once $d_{\text{mreach}-k}$ is calculated for all data points, then a minimum

spanning tree is constructed on those measures. From the connected components shown by the minimum spanning tree, a hierarchy of the connected components is built. To better imagine this, consider a dataset with a few dense clusters, points inside clusters will have small mutual reachability distances to the other data points in the cluster but at some point the cluster becomes connected and the next edge of the minimum spanning tree will be out-of-cluster, where the new out-of-cluster edge will be a significantly larger distance. If the edges of the minimum spanning tree are ordered by distance, it is possible to cut the tree into clusters by cutting at edges which have larger distance values, creating a hierarchy. Finally, the hierarchy is manipulated to return clusters. It is a flexible method that does not require the number of clusters, or groups, to be specified and the method detects clusters of different shapes and densities.

The clustering procedure is repeated throughout the model simulation at each iteration. This modelling choice ignores the persistence or "stickiness" mechanisms of groups. It would be impossible to observe phenomena such as group fragmentation or consolidation without changing group labels over time, so it is necessary to avoid static grouping. As consequence, group configuration exists in social systems in a continuum between immutability and potential reconfiguration at each temporal step. The choice presented here sides with the latter option to dispense of the modelling complexity related to group persistence in the proposition of the first model for the co-evolution of groups and opinions. The next section deals with these considerations in more detail.

Outliers, or edge cases, in the distribution are not assigned a group by HDBSCAN because they do not meet a critical level of likelihood for being part of a cluster. The agents that are not assigned a group are treated as out-groups by other agents that have a group identification but consider other outliers as in-group. Treatment of outliers is a difficult choice, on the one hand they have not been identified as a cluster so in-group identity seems illogical, while from another perspective to ignore an outlier identity and a potentially emergent group would be to disregard the solidarity of marginalised agents. In the example of Figure 3.3, two groups are successfully identified from a bimodal distribution and the agents at the edge of the modes are labelled as 'No Group'. Agents with the same label ('Group A', 'Group B', 'No Group') treat each other as in-group and any others as out-group.

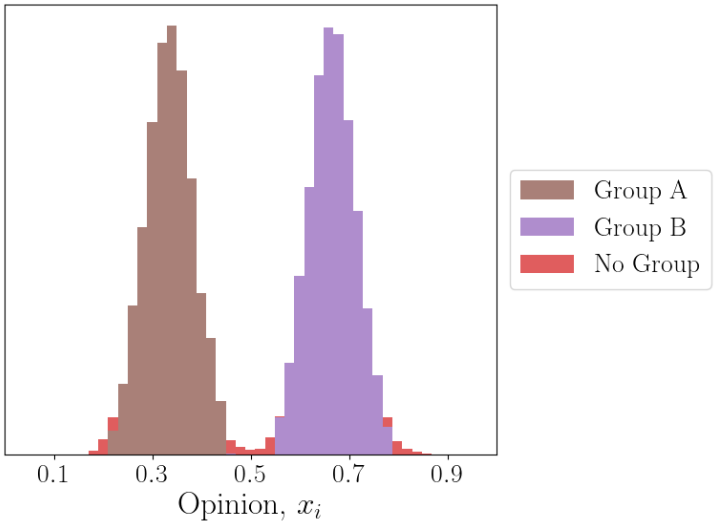


Figure 3.3: An example assignment of group identification using HDBSCAN on a bimodal distribution in the opinion space. The method identifies two groups and leaves edges cases with no group assignment, which aligns with an intuitive reading of the distribution.

3.3 The Consequences of Group Identification

Regarding the implementation detailed above, the first issue to note is that group labels given by the group identification rule are determined solely by the opinion distribution of agents. However, in reality, multiple political and social dimensions form group identity; the constituent parts may be associated political parties, ethnic groups, professional groups, age groups, just to name a few (Bodenhausen et al. 2012). Reducing identity labelling to just opinion similarity (in closeness and density) is clearly a significant simplification of the group identification process, but it is necessary to avoid introducing further variables, and therefore complexity, into a first proposition of a model. Given that opinions evolve over time in the model, it follows that group identification should evolve over time as well; it would be strange for the opinions of agents to reach consensus without also reaching one group identitification. Group identification is updated every iteration in this work's implementation, however a more complex model could aggregate identifica-

tion over a number of previous iterations and thus adjust the "stickiness" of group identification.

By splitting each model parameter, E, T, and R, into sub-parameters that depend on group label, there is an implication that group identification plays a role in the perception of others for each parameter. At a broad level that can be applied to all of the parameters, if the aspect of exposure operates both intra-group and inter-group then biases for in-group favoritism and out-group discrimination will apply (Tajfel 1974). More specifically, the existence of group-dependent exposure has been shown to exist in digital spaces (O'Callaghan et al. 2015; Bakshy et al. 2015; González-Bailón et al. 2023), as well as in physical spaces (Novelli et al. 2010); while perceived similarity with others, or consideration of others as in-group, has been found to increase the likelihood of sharing information (Baek et al. 2025). For tolerance, there is work showing a negative relationship between positive identification of the other as out-group and the positive evaluation of out-group arguments (Eschert and Simon 2019) – that is, agents identified as out-group are less likely to have their arguments reacted to positively. Finally, for responsiveness, there is evidence that people are more likely to respond to in-group opinions (Masson et al. 2016) and that feeling close to another increases assimilation of content shared by them (Balietti et al. 2021).

The clustering methods used for the group identification rule produces a global consensus across the population of agents. As such, the agroupment of the population assumes that agents have knowledge of the distribution of opinions which is clearly not realistic, although members of the public do make estimations about the beliefs of the population (Fields and Schuman 1976). The method employed in the rule also assumes that all agents agree on the definition of groups, so the labelling of in-group/out-group is symmetric for both agents in an interaction; however this is not necessarily the case, and that is without mentioning how people with multiple social identities might interact (Roccas and Brewer 2002). Despite these drawbacks the clustering example using HDBSCAN (Figure 3.3) returns a useful labelling of groups.

Other options for a group identification rule may be considered. There is a range of alternative clustering algorithms that could be used instead of HDBSCAN. However, most alternatives face the same issues of assuming population agreement and population-wide observation of opinions. Furthermore, most require pre-determination of the number of clusters present in the data which is a drawback since the number of groups may evolve during simulation experiments.

A different framework to clustering algorithms that can decide social boundaries between groups is presented by Yang et al. (2021). The method is motivated by reasoning from an agent's perspective (as opposed to the population perspective of clustering). Each agent follows an error minimization process at an individual-level and at a group-level. First, for an agent i, the error between actual opinion difference and expected opinion difference given group boundaries is minimized; where expected opinion difference is either 0 when agents i and j are of the same group, or, when they are of different groups, the opinion difference between the mean opinion of agent i's group and the mean opinion of agent j's group. Then, at a group-level, the agent i compares group boundaries with other group members and each moves their respective boundaries to minimize error between each one's boundary definition until consensus is reached within the group, so this step is minimizing global error for the group – this is akin to group members agreeing with others on a boundary. The result of the method (named here as MinError) can be seen in Figure 3.4, and full details of the method can be found in the original article by Yang et al. (2021). The principal drawback of the MinError method is that it is necessary to specify the number of groups prior to identifying the groups, and a secondary drawback is that all group members must agree on a boundary as in HDBSCAN. So, while the MinError method is motivated by social cognition, and therefore seemingly more fitting for a social process (group identification), HDBSCAN remains preferable.

Finally, to address the issue of group-wide agreement on boundaries, I developed an Ego-Centric group identification method whereby each agent has their own definition of group identification because there was not an appropriate solution to this problem given the novelty of how group identification is introduced to an agent-based model here. The starting point for the Ego-Centric method is to make use of the extended capabilities of HDBSCAN to provide a fuzzy, or soft, clustering to the population, rather than a hard category as in the group identification rule detailed above. As a result, each agent i is assigned a vector of probabilities dependent on their opinion $\gamma(x_i)$, for which the k-th element is the probability that the agent belongs to cluster k. By comparing $\gamma(x_i)$ and $\gamma(x_j)$ for agents i and j, it is possible to distinguish a fringe member of a cluster from member whose opinion is close to the mean opinion of the cluster. Therefore fuzzy method identifies areas of relative high density, as in the standard usage, then gives a value to where an agent is in relation to these densities. By calculating $\|\gamma(x_i) - \gamma(x_j)\|$ across x_i, x_j, it is possible to determine a likelihood of agent

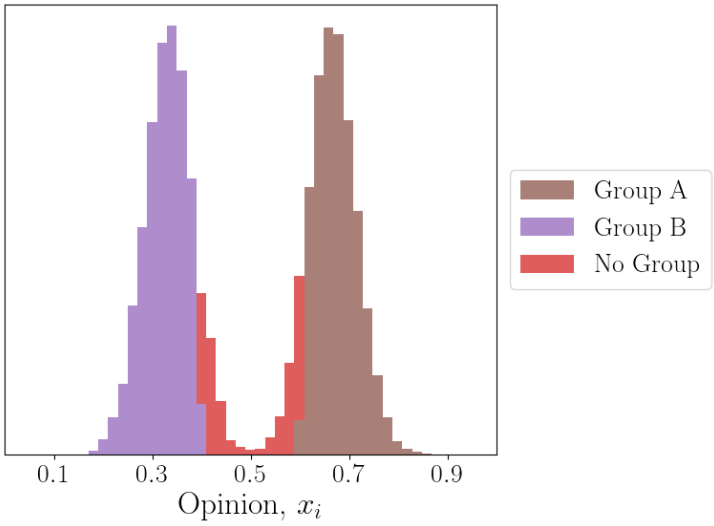


Figure 3.4: An alternative group identification method motivated by social cognition: the MinError Method (Yang et al. 2021). Two groups are identified and only agents in the middle of the opinion space are labelled as 'No Group'. However the process requires prior knowledge of the number of groups.

i considering agent j as in-group.

The probability of being considered in-group should also strictly decrease as opinion difference increases, since those agents that are further away in the opinion space should not be considered as more similar in terms of group identification. It can happen that if the opinions of i and j are on either side of a mode then $\|\gamma(x_i) - \gamma(x_j)\|$ is smaller than if j's opinion were at the mean of the mode, despite j's opinion being further away in opinion space. To ensure the probability is strictly decreasing, the set of agents $M_{i,j}$ that are those agents whose opinions lie between x_i and x_j in the opinion space. It is then stipulated that the probability that i considers j as in-group cannot exceed that between i and any member of $M_{i,j}$.

$$P(i \text{ considers } j \text{ in-group}) = \min_{\forall m \in M_{i,j}} (1 - \|\gamma(x_i) - \gamma(x_m)\|).$$

The clustering for three example agents can be seen in Figure 3.5. Each agent has clearly different understanding of which other agents are to be considered as in-group. An 'Extremist' on the left-hand-side of the space

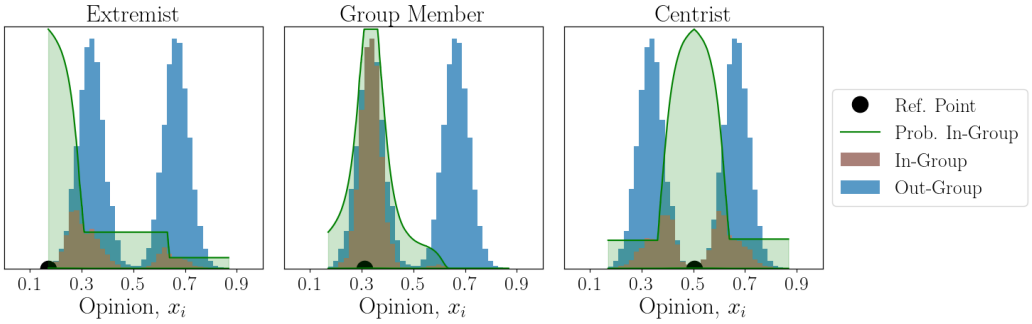


Figure 3.5: A second alternative group identification method for which individuals have their own unique group definitions: the Ego-Centric Method. Three example profiles from the opinion distribution are shown which each treat different parts of the population as in-group: the 'Extremist' considers 16% of the population as in-group, for the 'Group Member' it is 39%, and for the 'Centrist' it is 22%. The underlying histogram is coloured according to in-group/out-group identification, with the probability of considering another agent in-group shown by the green curve as a function of opinion difference.

considers most of the nearest mode to be in-group, while the 'Group Member' considers almost all of the mode to be in-group (not quite all of the mode since chance of being in-group is a strictly decreasing probability). The 'Centrist' sits between the two modes and treats them equally as either those to the left, or right, may be considered as in-group. Ultimately, this method is not employed in the model due to the extra layer of complexity it adds when trying to extract population-level behaviour for agents. The method could be usefully employed in future exploration of groups in agent-based models.

3.4 Simulation Procedure

The three rules that make up the Attraction-Repulsion Model with the group identification extension – the interaction, opinion update, and group identification rules – provide all that is needed to run simulation experiments upon a population of N agents. To begin an experiment, the opinions, x_i, of the agents are distributed according to a starting distribution. Simulations presented for results in Chapter 5 will start with a bimodal distribution in order

to ensure a group dynamic, given that group influence is the focus of the results. Next, the group identification rule is run and labels are assigned to the agents. The exposure rule is next, so agent i observes the opinions of the population, then agent $i+1$ will observe the opinions of the population, and so forth. Once all opinions have had the opportunity to be observed, agents' opinions are then updated by the opinion update rule. Finally, the iteration count is incremented as the model cycles from the group identification rule onward again.

All opinion updates are synchronous, so agent i updates their opinion at the same time as agent j updates their opinion, and then the next iteration of the simulation begins so group identification is evaluated anew, followed by potential observation of the newly updated opinions in the population according to the exposure rule, until opinion update happens once more. Given that the exposure rule is not deterministic, the simulations are executed twenty times to obtain average behaviour for the system.

For the given model parameters E, T, and R, an experiment consists of selecting one of the parameters to split into an in- and out-group version and searching the parameter space of the group-dependent parameters while the non-group-dependent parameters are kept constant. For example, $E = 0.1, T_{\text{in}} \in [0,1], T_{\text{out}} \in [0,1], R = 0.1$.

Algorithm 1 lays out the procedure for simulation, note that the rules of the model are referred to by name rather than detailed, since they are already explained in Section 3.2.

The convergence of the population's opinions to some steady state is key to knowing how long to run the model for. Each simulation is run up to a maximum of one thousand iterations, by which time the opinions of the population have come to a steady state; although a simulation may stop early if opinions have not changed over one hundred consecutive iterations ($T = 100$ in Algorithm 1). Figure 3.6 shows how the polarisation of opinions – assessed by the Duclos-Esteban-Ray measure, detailed in the next paragraph – evolves over the course of iterations. By three hundred iteration steps, the simulations approach a stable state. Iterations are allowed to continue to confirm that simulations are indeed stable.

Polarisation of the opinions of the agents at each time step is assessed by the Duclos-Esteban-Ray (DER) measure (Duclos et al. 2006). This measure provides a continuous single-valued, and therefore easily comparable, index of polarisation according to what the authors term an "identity-alienation" framework. The reasoning behind this is to represent two important aspects

Algorithm 1 Simulation experiments under the ARM for some set of parameters follow these steps.

$\{x_1, \ldots, x_N\} \leftarrow$ starting distribution
iterations $\leftarrow 0$
while *iterations* $<$ *iterations_limit* **do**
  GroupIdentificationRule to assign group identity to agents
  for x_i in $\boldsymbol{x}$ **do**
    InteractionRule(x_i)
  for x_i in $\boldsymbol{x}$ **do**
    OpinionUpdateRule(x_i)
  assign $\boldsymbol{x}$ updated opinion synchronously
  iterations $+= 1$
  if $\{x_1, \ldots, x_N\}$ unchanged over the T previous iterations **then**
    break

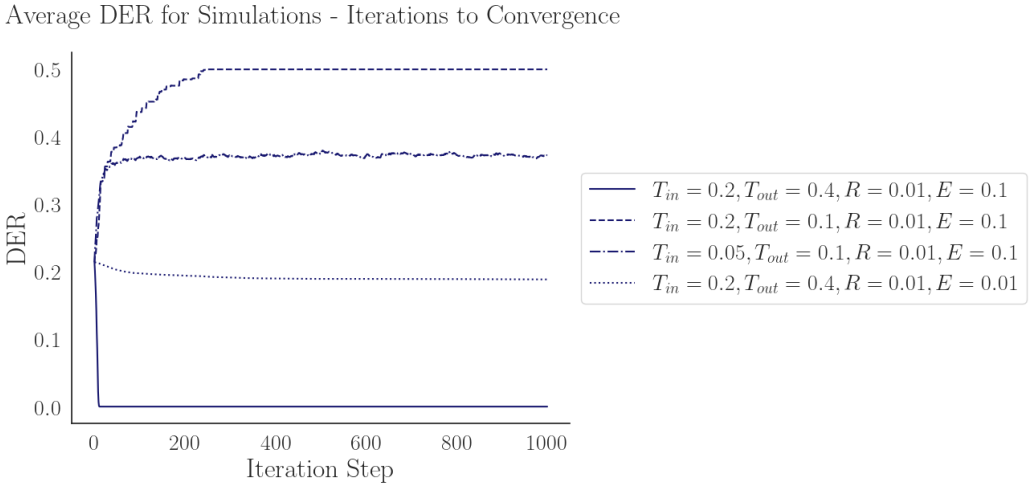


Figure 3.6: Convergence trajectories under different model parameter combinations. $R = 0.01$ means that opinion change is slow and hence the system takes longer to reach a steady state than when R is higher, so these are slow examples of convergence. Exploration of the implication of DER values for certain parameter sets is found in Chapter 5.

of polarisation: how dense modes in the data are (identity), and how spread the distribution is (alienation). The formula, as previously stated in Chapter

2, is

$$P_\alpha(f) \equiv \int \int f(x)^{1+\alpha} f(x') \|x' - x\| dx' dx. \tag{3.3}$$

where x and x' are points in the opinion space, $f(x)$ is the density at x, and α is a parameter that is set to 0.5. In order to use the measure on the discrete observations returned by the simulation, $f(x)$ is approximated by kernel density estimation and the sample based estimator for $P_\alpha(f)$ is used, as described in Section 4 of Duclos et al. (2006). For a bimodal starting distribution, such as that in Figure 3.3, the value of DER is 0.21, which represents mild polarisation. The minimum value of 0 is achieved at total consensus and the maximum value of 0.5 is achieved when half of the population is at one extreme of the space with the other half at the other extreme. When DER values are provided for end-states of the experiments, it is an average of the final one hundred iterations.

3.5 Illustrating Group Identification

The Attraction-Repulsion Model will be extensively tested in Chapter 5 to gain insight into how group identification affects the evolution of polarisation in the agent population. In this section, a first glimpse is given to illustrate how the dynamics of the model change when group identification is present.

The first steps of the opinion change for a simulated population can be seen in Figure 3.7 for the case where group identification is present in the simulation and for the case where no group differentiation is made by agents and all are treated equally. In the group identification example tolerance of opinions of those agents belonging to an out-group is lower than tolerance towards in-group agents, therefore repulsion between agents of the two groups occurs and moves the population towards polarisation while opinions within each group are consolidated towards an in-group consensus. When no group identification exists, each agent is treat the same so $T = T_{\text{in}} = T_{\text{out}}$. In this no group example, the original in-group tolerance is adopted globally as so the population arrives at consensus given most interactions are attractive – the opposite behaviour to polarisation found in the group identification case. The example here assumes in-group treatment is applied to all other agents when no group identification is present, the general T could be calculated in a different manner but the consideration and treatment of groups clearly results in different behaviours.

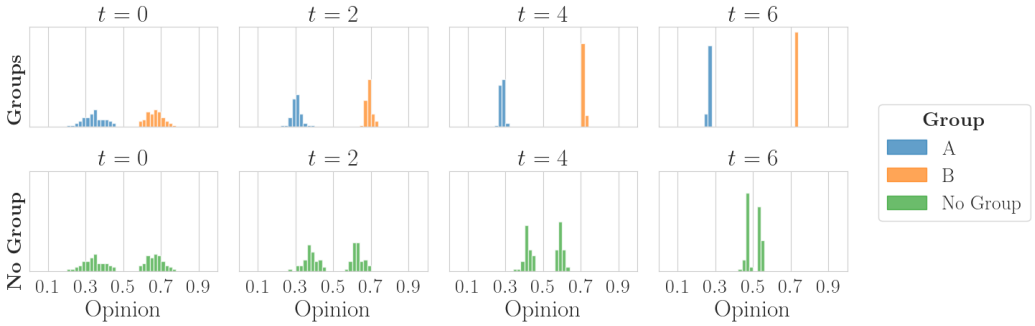


Figure 3.7: Comparative model simulation over initial time steps with and without group identification. In the group case, parameters are $T_{\text{in}} = 0.5$, $T_{\text{out}} = 0.2$, $E = 0.1$, $R = 0.01$. When group identification is not present, and each agent is considered as an equal individual, the parameters are $T = T_{\text{in}} = T_{\text{out}} = 0.5$, $E = 0.1$, $R = 0.01$, resulting in consensus rather than polarisation. The consideration and different treatment of the out-group leads to a considerably different outcome for the agents' opinions.

The development of average group opinions over time can be considered alongside the development of individual agents' opinions. This presents a novel level of analysis for the model. Group size, group mean opinon, and group opinion spread, can all be visually assessed by a alluvial diagram, as seen in Figure 3.8. Each rectangle of the visualisation shows group size and mean opinion as rectangle length and vertical position, respectively. Group opinion spread could be added to the alluvial diagram, represented as rectangle width or replacing group size as rectangle length, for legibility it has not been included currently, thus prioritising group size and group opinion mean as represented dimensions. The colour of the rectangle ranges from dark blue/red at the extremes to light grey at the opinion space midpoint.

This view of the behaviour allows categorisation of types of dynamics in the model at a broad level by considering whether groups diverge, converge, or remain static. For the example case, the groups drift towards complete polarisation resulting in both groups being pushed towards the extreme of the opinion space. Full characterisation of polarisation types achieved by the model is discussed in Chapter 5. It is not possible to compare with the case of no group identification, since the group labels are necessary to create the diagram.

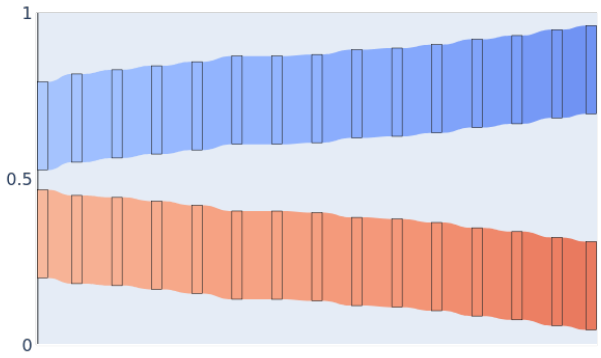


Figure 3.8: Evolution in time of the groups observed in a simulation for parameters $T_{\text{in}} = 0.5$, $T_{\text{out}} = 0.2$, $E = 0.1$, $R = 0.01$. Each rectangle represents a group and is centred at the mean opinion of the group. The length of the rectangle represents group size, which is equal and unchanging in this simple example but more complex behaviours see splintering or combining groups which change the rectangle size. The starting position of the groups is unstable, as they drift to the extremes of the opinion space.

3.6 Complexity of the Agent-Based Model

For each iteration of the model the computational complexity is analysed in order to assess how the model scales with increasing the size of the agent population. The implication is then how would the model handle large empirical datasets built from real-world data, for example, empirical populations of social media users with information subscriptions networks linking them and inferred positions on continuous opinions scales (Ramaciotti et al. 2022).

Time complexity of the agent-based model, presented in Algorithm 1, with n agents is as follows: assigning opinion is $\mathcal{O}(n)$ since the assignment operation is applied to each of the agents, group identity as identified by HDBSCAN is $\mathcal{O}(n \log(n))$ (McInnes and Healy 2017), interactions are $\mathcal{O}(n^2)$ given that every agent has the possibility to talk to all other agents, then the worst case scenario of the number of updates is an each agent has interacted with all other agents so $\mathcal{O}(n^2)$, finally the history check is $\mathcal{O}(n)$. Therefore

overall time complexity of the model is $\mathcal{O}(n^2)$, driven by the exposure rule and the update rule. Note that removing the step of identifying groups would not actually reduce complexity.

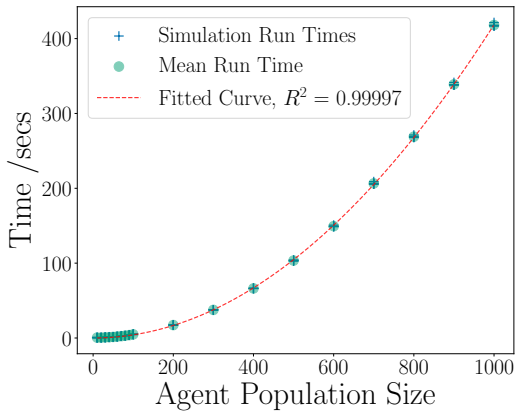


Figure 3.9: Increasing computational time is needed for larger populations. Model simulations were run five times for different population sizes n, with constant parameters $T_{\text{in}} = T_{\text{out}} = E = R = 0.2$, to provide average experiment runtimetime; a quadratic curve is then fitted to the average length of time, matching with the described $\mathcal{O}(n^2)$ complexity.

The space necessary to run the model depends on the number of agents n but also the number of time steps that are saved as history – this is denoted by h. Space needed for the model history is $\mathcal{O}(h \cdot n)$ as all agents' opinions are saved. The initial opinion assignment has complexity $\mathcal{O}(n)$. The comparison against previous iterations for convergence is implemented by a comparison against the previous iteration's opinions, then a counter is incremented if t and $t-1$ are identical (up until the limit T), and then iteration t replaces $t-1$ in memory for the comparison check of the next iteration, so complexity for this step is $\mathcal{O}(n)$. All interactions and therefore possible updates in an iteration are temporarily saved to be updated at the end of the iteration, as such this step uses $\mathcal{O}(n^2)$ auxiliary space given that each of the n agents could interact with all of the other $(n-1)$ agents. The model is then either $\mathcal{O}(n^2)$ or $\mathcal{O}(h \cdot n)$ if $h < n$ or $h > n$, respectively. The interactions between agents and opinion updates are the points necessitating the most computation, therefore to improve efficiency of the model these steps must be changed or approximated. The interaction

In its current state, the model is one-dimensional however if it were extended to a d-dimensional space – that is agents hold multiple opinions represented as a d-dimensional vector and compare their own vector against those that they interact with then the complexity would increase to $\mathcal{O}(d \cdot n^2)$.

In the worst case, an agent will interact with all $n-1$ other agents – and this becomes increasingly likely as the population moves towards consensus. To gain further insight, it is useful to consider the expected number of interactions for an agent, which will depend on the distribution of opinions and the exposure parameter.

Expected Number of Interactions

To address the question of the number of expected interactions across agents, first consider the case for agent i. The probability that agent i observes agent j's opinion is (the same as Equation 3.1):

$$w_{ij} = 0.5^{\|x_j - x_i\|/E}.$$

Then the expected number of interactions of agent i, or the expected number of other agents j that are observed, is the sum of these probabilities since each possible observation is a Bernoulli random variable:

$$\mathbb{E}[d_{\text{in}}(i)] = \sum_{j \neq i} w_{ij}.$$

Each agent j is possibly observed in turn and then effectively removed from the set of possible interactions for an iteration, so who agent i observes is a series of Bernoulli trials with probability w_{ij} and it is not necessary to count the number of ways that agent j could be chosen as a neighbour because each j is trialled only once for each i. If all agents are at consensus and have the same opinion, then each $w_{ij} = 1$, so the expected number of interactions of agent i is $n-1$, which is the worst case for model complexity as previously stated. At the other extreme, as w_{ij} tends towards 0 (for example, in the case where E tends towards 0), then the expected number of interactions of agent i is 0.

The random variables x_i and x_j may be considered as independent and identically distributed (i.i.d.) following a known continuous probability distribution $p(x)$ – certainly, at $t = 0$ this is true, and a distribution may be

approximated by Gaussian kernels for each subsequent time step. The expected number of interactions of an agent i with opinion x in the continuous distribution is:

$$\mathbb{E}[d_{\text{in}}(i)|x] = \mathbb{E}_x \left[\sum_{j \neq i} w_{ij} \right] = \sum_{j \neq i} \mathbb{E}_x[w_{ij}],$$

by the linearity of expectation. Since the random variables x are i.i.d., each $\mathbb{E}_x$ is identical and so the summation can be simplified to:

$$\mathbb{E}[d_{\text{in}}(i)|x] = (n-1) \cdot \mathbb{E}_x[w_{ij}].$$

With x and another agent's opinion x', an evaluation of the expectation, for some $p(x)$, can be written:

$$\mathbb{E}[d_{\text{in}}(i)|x] = (n-1) \cdot \int_0^1 0.5^{\|x'-x\|/E} \cdot p(x')dx'. \tag{3.4}$$

Notice that the expected number of interactions is a factor of n, the complexity of interactions is therefore confirmed as $\mathcal{O}(n^2)$ since this is the expected number of interactions for each of the n agents. Equation 3.4 may be evaluated numerically or considered for special cases such as $x \sim \text{Beta}(\alpha, \beta)$.

Furthermore, it is possible to find the average number of interactions across all x:

$$\mathbb{E}\left[\mathbb{E}[d_{\text{in}}(i)|x]\right] = (n-1) \cdot \int_0^1 \int_0^1 0.5^{\|x'-x\|/E} \cdot p(x') \cdot p(x)dx'dx.$$

For numerical insight at this point, a Beta distribution can be substituted in the place of $p(x)$, and evaluated to give the average number of interactions for a distribution of x as a function of the parameter E. If $\alpha = \beta = 1$, then this is equivalent to a uniform distribution on the [0,1] interval.

Figure 3.10 displays the numerical evaluation of the average degree under a uniform distribution (special case of the Beta distribution) for different values of the exposure parameter with $n = 100$ agents. When $E = 1$, $\mathbb{E}[d_{\text{in}}] = 79.6$, falling to 24.4 for $E = 0.1$.

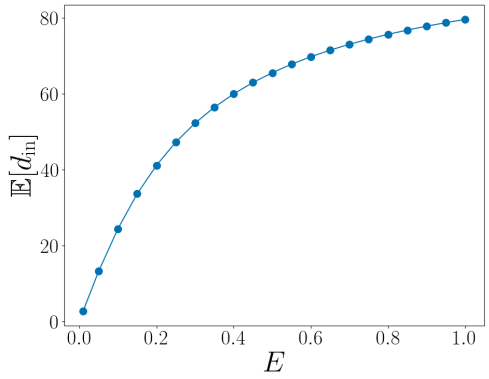


Figure 3.10: Expected number of interactions for $x \sim \text{Beta}(1, 1)$, while varying E. As E increases so does the expected number of interactions for some agents i in the Attraction-Repulsion Model.

Chapter 4

Empirical Distributions in Opinion Dynamics

A possible next step for the agent-based model is to connect to real-world distributions of opinions – for example, an opinion distribution estimated from social media data (Ramaciotti et al. 2022), which would lead to grounding the model in a wider context outside of traditional simulation scenarios. However, the volume of computations necessary to scale the model to the many agents present in large populations arises as a limiting factor (see the complexity analysis presented in Section 3.6). With large N and a large parameter space for models, the computational resources (both time and processing power) necessary to evaluate the models and calibrate parameters to fit empirical distributions becomes intractable. Furthermore, it is an open question as to how a distribution of opinions may be translated into the opinions of a population of N agents.

The agent-based model of the previous chapter is a useful tool, however, the issues of how the process could cope computationally when the number of agents, N, becomes large, and how to rigorously translate empirical distributions into the functions that define the exposure and opinion update rules, limit the capabilities of the model.

Scaling micro-level dynamics to assess macro-level populations, and the complexity problems that are incurred, is also found when modelling collections of particles, such as gases, in physical contexts. A classic approach to addressing these challenges is to take the mean-field limit of the system as the number of agents, N, tends to infinity (Braun and Hepp 1977). Under this approach the focus changes from following individual agents and their

positions to knowing the distribution of agents within the population. The probability of observing an agent with opinion x at time t is defined as $\mu_t(x)$, and this becomes the quantity to model (rather than change of an agent).

4.1 Mean-Field Limit of Attraction-Repulsion Model Equations

Prior to deriving the mean-field limit, it is necessary to clearly state the equations for the model that is being approximated. A general formulation of an agent-based opinion dynamics model with N agents that interact in a pairwise manner may be written as:

$$\frac{d}{dt}x_i^N(t) = \frac{1}{N}\sum_{j=1}^{N} w_{ij}\phi(x_i^N(t), x_j^N(t)), \qquad (4.1)$$
$$t \geq 0,\ i = 1, \ldots, N,$$

where $x_i^N(t) \in \mathbb{R}^d, d \geq 1$ is the opinion of agent i at a time t and $\phi(\cdot)$ is the interaction function determining opinion update. Furthermore, this formulation includes an interaction weight $w_{ij} \in \mathbb{R}$ which can moderate the effects of interactions. Here, w_{ij} represents the potential network structure of interactions between agents, or in the context of the previous section it could be a probability of interaction between agents i and j, or reflect a different interaction weighting. While it is conceivable that w_{ij} could be subsumed into ϕ in the case that the interaction weight depends solely on the opinions x_i and x_j, there are cases when interaction between agents i and j may not depend on solely opinions at time t. For example, the network structure governing interactions between agents may be static or develop at a different rate to opinions.

For the case of the Attraction-Repulsion Model, the opinion update rule in Equation 3.2 maps to the interaction function (that is $\phi(x_i, x_j) = +R(x_j - x_i)$ when $\|x_j - x_i\| \leq T$, or $\phi(x_i, x_j) = -R(x_j - x_i)$ when $\|x_j - x_i\| > T$); and the exposure rule of Equation 3.1 can either be considered for the interaction weight ($w_{ij} = 0.5^{\|x_j - x_i\|/E}$), or subsumed into $\phi(x_i, x_j)$ given that both are a function of x_i and x_j – in which case, w_{ij} would be a constant or free to represent further network structure. The purpose of w_{ij} is to represent network structure, as such the original exposure rule may be considered from

two perspectives: either it defines a weighted network between agents, or it is a weight on the exchange of opinions. Given that the exposure rule evolves over time and is a function of opinions, it will be subsumed into $\phi(x_i, x_j)$ for the modelling presented later.

To develop the equations governing the behavior of the mean-field approximation, the mass distribution of individuals in the opinion space at a time t, $\mu_t^N(x)$, and its limit as N tends to infinity, $\mu_t(x)$, must be defined. For a discrete case of N agents, this quantity can be expressed as the sum of Dirac functions at the positions of the agents,

$$\mu_t^N(x) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i^N(t)}(x). \tag{4.2}$$

This probability is the sum of the observations at opinion x divided by the size of the population. As N tends to infinity this yields the continuous distribution, so $\lim_{N\to\infty} \mu_t^N(x) = \mu_t(x)$.

There is now a key distinction to make: either the system (population of agents) is *exchangeable*, or it is *non-exchangeable*. If two agents in the system can be exchanged without changing the dynamics of the other particles then they are exchangeable. The system is exchangeable if there is no network structure, $w_{ij} = 1$ in Equation 4.1, since it is possible to shuffle the labels of all agents without impacting the dynamics of the system. If agents are *exchangeable*, they may also be called *indistinguishable* or *unlabelled*.

However, the exchangeable case presents a flaw when used to model social systems. Particles being indistinguishable may be suitable for modelling the reality of gases, but fails to capture important aspects of social systems. At this point, the analogy between particles and persons becomes strained. When an individual is interested in a population, it is likely that there is some information – perhaps demographics or group identity – that would be of interest to track through simulation, this is only possible in the non-exchangeable context.

The development of the mean-field approximation will follow the exposition detailed in Ayi and Duteil (2024). It is useful to begin with the more straightforward case (and classic case in physical contexts) that treats agents i and j as exchangeable ($w_{ij} = 1$). Considering the derivative with respect to time of the integral across the domain of a test function $f \in C_c^\infty(\mathbb{R}^d)$, *i.e.* a smooth/infinitely differentiable compactly supported function, then

the following derivation provides a path towards a solution.

$$\begin{aligned}
\frac{d}{dt}\int_{\mathbb{R}^d} f(x)d\mu_t^N(x) &= \frac{d}{dt}\left[\frac{1}{N}\sum_{i=1}^{N} f(x_i^N)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{d}{dt}x_i^N \cdot \nabla f(x_i^N) \\
&\qquad \text{[by the chain rule]} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\phi\left(x_i^N(t), x_j^N(t)\right)\cdot \nabla f(x_i^N) \\
&\qquad \text{[by substituting Equation 4.1, with } w_{ij}=1\text{]} \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi\left(x,y\right)\cdot\nabla f(x)d\mu_t^N(x)d\mu_t^N(y), \\
&\qquad \text{[by moving from discrete to continuous context]}
\end{aligned}$$

for which the solution to μ_t^N for an exchangeable system is deduced as a solution to the Vlasov-type equation; note that a Vlasov-type equation describes the evolution in time of a distribution function (Dobrushin 1979). To be clear, an equation of Vlasov-type is

$$\partial_t\mu_t + \nabla\cdot(V[\mu_t]\mu_t) = 0, \qquad V[\mu_t](x) = \int_{\mathbb{R}^d} w(\|x-y\|)(y-x)d\mu_t(y).$$

Relating this to the case presented here, the solution is:

$$\partial_t\mu_t(x) + \nabla\cdot\left[\left(\int_{\mathbb{R}^d}\phi(x,y)d\mu_t(y)\right)\mu_t(x)\right] = 0. \tag{4.3}$$

The Equation 4.3 is the solution for an exchangeable system, showing how $\mu_t(x)$ evolves over time.

Next, attention is turned to the solution for the non-exchangeable system. As mentioned, under the non-exchangeable condition the agents maintain an identity so cannot be considered as equivalent, which results in a longer derivation and increased mathematical complexity. It is necessary to redefine $\mu_t(x)$ which is currently the probability of finding an agent that has opinion x at time t, no matter which agent. In order to capture the individuality of agents, an identity parameter $\xi \in [0,1]$ is introduced so that the quantity

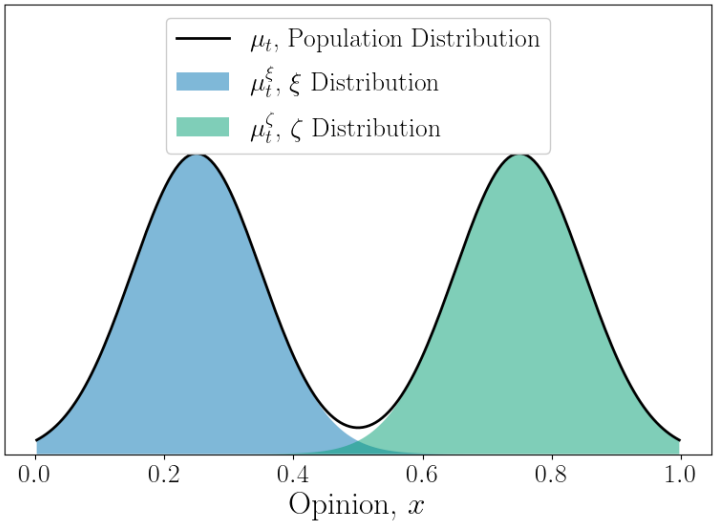


Figure 4.1: A bimodal distribution of opinions with two identity parameters, ξ and ζ. The normalized sum of two Gaussian distributions centred at 0.25 and 0.75, both with standard deviation 0.1, represents the distribution of a population's opinions. In simulation it is possible to model the progression of both modes by defining an identity for each.

of focus becomes $\mu_t^{\xi}(x)$, the probability of finding an agent with a labelled identity ξ at opinion x at time t.

It is useful to consider ξ in the context of taking the graph-limit, or *graphon*, of the graph that describes the edges present in the system (Lovász 2012). So as the size, N, of a graph, G_N, tends to ∞, the edge weight between one identity ξ and another ζ is defined as

$$w(\xi,\zeta) = \begin{cases} 1 & \text{if } (\xi,\zeta) \in \left[\frac{i-1}{N}, \frac{i}{N}\right) \times \left[\frac{j-1}{N}, \frac{j}{N}\right) \text{ and } (i,j) \in E(G_N), \\ 0 & \text{otherwise.} \end{cases}$$

This provides a definition for the edges that construct the graph between the population, if an edge exists between i and j in the original graph then the edge is mapped between their respective labels in the graphon. The definition may be broadened for $w \in \mathbb{R}$, to consider weighted edges. An intuitive understanding of the identity ξ is to consider it as one of a set of average profiles within the population, please see Figure 4.1 for an example. A discussion and explanation of the extension to graph limits for opinion

dynamics may be found in Prisant et al. (2024).

The graph structure stipulated by w means that taking the limit of the graph as N tends to infinity must be possible. A condition for this is that the graph must be dense, that is, the number of edges grows proportionally to the square of the number of vertices. Although there are alternative graph limit formulations from recent results to define the limit of a graph sequence which is sparse, such as L^p-graphons and graphops (Borgs et al. 2019; Backhausz and Szegedy 2022), as explained in Ayi and Duteil (2024). For the current w formulation, the density requirement means assuming the global structure of the graph when the size becomes large. Two examples can be seen in Figure 4.2, where rules for the graph function are given in terms of the population size N. Further details on convergence criteria can be found in Lovász (2012).

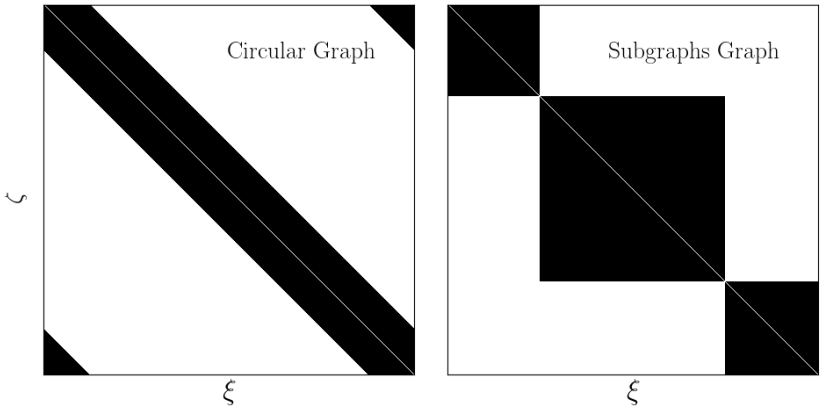


Figure 4.2: Pixel representations of the adjacency matrix of example graphs that could be have a limit taken and be represented by w, $w(\xi, \zeta) = 1$ is coloured black. On the left, a circular graph where agent i is connected to neighbours $i - \alpha, \ldots, i - 1, i + 1, \ldots, i + \alpha$ for $\alpha = N/8$. On the right, there are three subgraphs that are themselves complete, the boundaries of the subgraphs are at nodes $N/4$ and $3N/4$.

A consequence of introducing the parameter ξ is that, when modelling the evolution of the system in Section 4.2, it is possible to follow the development of not just the distribution as a whole but also average profiles of individuals that constitute the population. From a deterministic viewpoint, $\mu_t(\Delta\xi \times \Delta x)$ represents the mass of agents with labels in $\Delta\xi$ and positions in Δx at time t.

Recent works (Ayi and Duteil 2024; Jabin et al. 2025) show that in the setting of μ_t^ξ the solution for an inexchangeable mean-field limit of the system is

$$\partial_t \mu_t^\xi(x) + \nabla_x \cdot \left[\left(\int_\Omega \int_{\mathbb{R}^d} w(\xi, \zeta) \phi(x, y) \mu_t^\zeta(y) dy d\zeta \right) \mu_t^\xi(x) \right] = 0, \tag{4.4}$$

where Ω is the domain of the interval, typically $[0, 1]$. Despite the increased difficulty presented by including identities, reflected in the relative novelty of solutions when compared to exchangeable versions, the non-exchangeable system is used for the modelling that follows given its preferable nature for social contexts.

4.2 Finite Volume Method for Attraction - Repulsion Model

Commonly applied in Physics and Applied Mathematics, the finite volume method is a framework to discretise differential equations for the purpose of numerical simulation, that has its roots in modelling physical phenomena. The crux of the method is the use of the divergence theorem to transform volume integrals that have a divergence term into surface integrals. Then changes to the studied quantity of interest may be evaluated as fluxes only at volume surfaces rather than flux across an entire volume. Simply put, the flux into a volume depends on the flux out of the adjacent volumes – what goes out of one volume, must go into its neighbouring volume.

For the finite volume method, there are two features that make the method attractive to use. The first is that the method is locally conservative so all flux is accounted for in the method, critically this aligns with Equation 4.4 which is a conservation law – that is, the population is neither created nor destroyed during simulation. While a second advantage is that the mesh of finite volumes may be unstructured, which provides flexibility in model construction.

Aside from the finite volume method, the finite difference method and the finite element method also exist as methods to discretise differential equations by splitting Ω into intervals. Neither of the two alternatives conserve fluxes unless additional constraints are applied, so they are avoided for use with the conservation law discussed.

I will now give a brief overview of the finite volume method, then place it in the context of the Attraction-Repulsion Model used earlier. Following this, I will provide some validations demonstrating behaviour of the model prior to using it for simulations.

4.2.1 Method Overview

The finite volume method provides the framework for moving from the continuous space of the mean-field approximation into the discrete space used for computation and simulation of conservation laws that govern a continuous quantity, q. The quantity is typically energy or mass, but in the context of this work it is the distribution of agents with identity ξ along an opinion scale, that is μ_t^ξ. It will then be possible to approximate the solution for $\mu_t^\xi(x)$ forwards in time for $t \geq t_0$, where t_0 is a starting time (Eymard et al. 2000).

A general statement of a conservation law at a point in space, x, and a time, t, may be written as

$$\partial_t q(x,t) + \nabla \cdot \mathbf{F}(x,t) = f(x,t). \tag{4.5}$$

Where $\partial_t q(x,t)$ denotes the partial derivative of the quantity with respect to time, $\nabla \cdot \mathbf{F} = \frac{\partial F_1}{\partial x_1} + \cdots + \frac{\partial F_d}{\partial x_d}$ is the divergence of the flux $\mathbf{F}$ along each of the d dimensions of the space, and the function f is a "source term" that allows for the possibility of density loss or creation. The terms on the left-hand side of Equation 4.5 are the change in the quantity over time and the force acting on the quantity for each point x at time t. In this work, the source term, f, is set to 0, so the overall size of the quantity remains constant and as such the divergence of flux, $\mathbf{F}$, accounts for all changes in the quantity, q, over time. From this continuous description, it is necessary to explicitly discretise both space, x, and time, t, as follows.

For space discretisation, a mesh of cells (the finite volumes) is introduced on the domain Ω of the opinion space in $\mathbb{R}^d$. Starting with a one-dimensional case, consider a grid of points $x_i, i \in \{0, \ldots, M+1\}$, such that

$$0 = x_0 < \ldots < x_{i-1} < x_i < x_{i+1} < \ldots < x_{M+1} = 1.$$

A cell C_i for $i \in \{1, \ldots, M\}$ is then defined by reference to the mid-points, $x_{i+1/2}$, of the original grid of points, x_i.

$$C_i =]x_{i-1/2}, x_{i+1/2}[, \qquad x_{i+1/2} = (x_i + x_{i+1})/2, \qquad \Delta x_i = x_{i+1/2} - x_{i-1/2}.$$

This may be generalized to d dimensions, however the modelling that follows remains in one-dimension so this is not discussed here.

Time discretisation is achieved by introducing time steps that constitute an increasing sequence t_n, $n \in \mathbb{N}_0$ (the set of nonnegative integers) with $t_0 = 0$. The time steps may be defined as constant and regular by letting each time step be $\Delta t \in \mathbb{R}_{>0}$ and thus $t_n = n\Delta t$. Or, as is done in the simulations that follow, the time steps may be variable to ensure convergence as rates of change of the modelled quantity change over time. The $n+1$ time step is defined by the Courant-Friedrichs-Lewy (CFL) condition which depends on the size of the flux betweens cells:

$$t_{n+1} = t_n + \min_i \left(\frac{\beta \Delta x_i}{\max\{|\lambda_{i-1/2}, \lambda_{i+1/2}|\}} \right) \tag{4.6}$$

where β is the Courant number, set to 0.49 in later simulations to ensure convergence, and $\lambda_{i+1/2}$ is the wave speed calculated at $x_{i+1/2}$ – the precise definition in the model context is found in Section 4.2.2, but the wave speed may be considered as the flux at $x_{i+1/2}$ unweighted by the quantities q_i and q_{i+1} on either side of the cell boundary, such that $\mathbf{F}$ is of the form $\lambda_i \nabla q_i^n$.

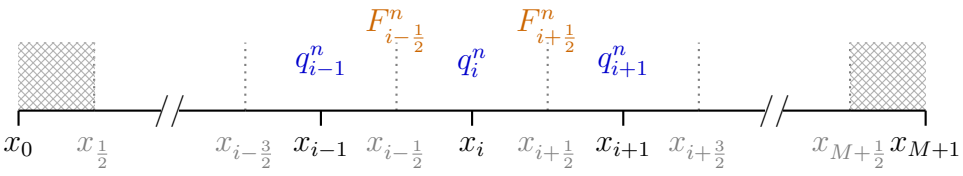


Figure 4.3: Discretisation of the simulation domain for the implementation of the finite volumes method. Schematic representation of the mesh cells with grid points, x_i, quantities being simulated, q_i^n, and fluxes, $F^n_{i\pm\frac{1}{2}}$ between cells. The shaded zones at the extremes of the domain show the region between the ghost points and the edge of the space covered by cells, C_i, that is not simulated.

In light of both discretisations, the following notations for the discretised quantity and flux are also adopted: $q_i^n = q(x_i, t_n)$, $F_i^n = F(x_i, t_n)$. See Figure 4.3 for a visual representation of the discretised space. Note that the cells C_i cover the opinion space between $x_{1/2}$ and $x_{M+1/2}$, therefore there is half a cell at each extreme which is not covered. In other words, there are as many cells, C_i, as there are grid points, x_i, that are not on the edge of the opinion

space. The points x_0 and x_{M+1} are then ghost points. With sufficient M, this does not pose a problem in simulation as the size of the ghost interval, $\Delta x_i/2$, decreases.

Implementing the time and space discretisations into Equation 4.5, setting the source term to 0, and using the divergence theorem on the flux divergence term, yields

$$\int_{C_i} \frac{q_i^{n+1} - q_i^n}{\Delta t_n} dx + \int_{\partial C_i} \mathbf{F}_i^n \cdot \boldsymbol{n}_{C_i}(x_i) d\gamma(x_i) = 0. \tag{4.7}$$

where $\boldsymbol{n}_{C_i}(x_i)$ is the unit normal vector to the surface ∂C_i at point x_i outward to the cell C_i, $\Delta t_n = t_{n+1} - t_n$, and $d\gamma$ is the integration symbol for ($d-1$-dimensional) Hausdorff measure on the considered boundary. In a one-dimensional setting, the surface is the boundary between cells along the scale and so the unit normal vector points either consistently up the scale or down the scale, since that is perpendicular to the cells.

To continue, the one-dimensional case is taken for clarity and ease; as such the finite volume scheme for solving Equation 4.7 is of the form

$$\begin{aligned}
\int_{C_i} \frac{q_i^{n+1} - q_i^n}{\Delta t_n} dx &= - \int_{\partial C_i} \mathbf{F}_i^n \cdot \boldsymbol{n}_{C_i}(x_i) d\gamma(x_i) \\
\frac{\Delta x_i}{\Delta t_n} \left(q_i^{n+1} - q_i^n \right) &= - \left(F_{i+1/2}^n - F_{i-1/2}^n \right) \\
q_i^{n+1} &= q_i^n - \frac{\Delta t_n}{\Delta x_i} \left(F_{i+1/2}^n - F_{i-1/2}^n \right),
\end{aligned} \tag{4.8}$$

which provides a value for the quantity in cell C_i at the next time step t_{n+1} given knowledge of the system at the current time step. This is the central part to the scheme – it allows the simulation to move forward in time.

The boundaries of the interval are key points to consider since the mass must be conserved with no loss of opinions outside of the domain of the opinion space. This may be written as

$$\partial_t \int_\Omega q_i^n(x) dx = 0.$$

By integrating Equation 4.5 over the domain of the opinion space Ω, a relationship with the flux can be stated:

$$\partial_t \int_\Omega q_i^n(x) dx + \int_\Omega \nabla \cdot \mathbf{F}(x,t) dx = 0.$$

Then, applying the divergence theorem, the equation becomes

$$\partial_t \int_\Omega q_i^n(x)dx + \int_{\partial\Omega} \mathbf{F}(x,t)\cdot \boldsymbol{n}\, dx = 0.$$

To ensure that mass is conserved, it must be that $\mathbf{F}(x,t)\cdot\boldsymbol{n} = 0$ at the boundary, $\partial\Omega$. If $\mathbf{F}$ is of the form $\lambda_i \nabla q_i^n$ (as it is in the implementation that follows), there are two ways to achieve mass conservation, either the wave speed causing the flux is forced to be equal to zero or the derivative of the quantity, q_i^n, is equal to zero. The simpler approach is impose that λ_i is zero at the boundary, rather than approximating the gradient of the quantity at the boundary. Hence the boundary condition is imposed such that the wave speed, and therefore flux, is zero at the boundary to conserve mass.

A consequence of the chosen boundary condition is that once mass hits the boundary it then cannot leave. In a real-world context this would mean that once some part of the population hits such extreme opinions, the opinion will not change (sometimes termed as 'stubborn' in opinion dynamics models). It is shown later that despite the stickiness of the boundaries, the model reproduces the behaviour found in an agent-based model.

The final part needed to fully describe the finite volume method is to determine the value of the flux at the boundary of each cell of the mesh. That is, calculating $F^n_{i+1/2}$ given the fluxes, F^n_i, F^n_{i+1}, that are already known. There are several existing methods that can be used to define the numerical flux at the boundary and in this work the Rusanov flux is chosen as it is relatively simple (Bouchut 2004) which makes it computationally efficient. Therefore,

$$F^n_{i+1/2} = \lambda^n_{i+1/2}\, q^n_{i+1/2} - \frac{|\lambda^n_{i+1/2}|}{2}\left(q^n_{i+1} - q^n_i\right). \tag{4.9}$$

where $\lambda^n_{i+1/2}$ is the *wave speed* discussed earlier (the effect of interaction at $x_{i+1/2}$ without the weight of quantities q^n_i, q^n_{i+1}) – what this means in the opinion dynamics model will be precised in the next section. An intuitive understanding of the Rusanov flux is that there are two parts: the centred flux at $q^n_{i+1/2}$ which is the main flux term for a smooth solution, while the $\left(q^n_{i+1} - q^n_i\right)$ term accounts for *numerical dissipation*. Dissipation is essentially smoothing if the difference between q^n_{i+1} and q^n_i is large in order to stabilise the scheme near discontinuities. This completes the overview of the method, it now remains to connect it with the mean-field limit of the Attraction-Repulsion Model.

4.2.2 Method Implementation

The finite volume method must now be placed in the context of the mean-field solution (Equation 4.4), this section details how the numerical simulations of the continuous version of the discrete agent-based model can be achieved. An inspiration, although in a different context and with different fluxes, may be found in Audusse et al. (2016).

As well as the time and space discretisation previously described, the identity ξ must be discretised. This is achieved in much the same way as the space discretisation of x by defining $\xi_j, j \in \{1, \dots, P\}$ grid points on the domain Ω with $\xi_{j+1/2} = (\xi_j + \xi_{j+1})/2$, and $\Delta\xi_j = \xi_{j+1/2} - \xi_{j-1/2}$.

Therefore, by substituting the notation of the mean-field limit in Section 4.1 into Equation 4.8 which provides a value for the quantity in cell C_i at the next time step t_{n+1}, the solution to the finite volume method can be written as

$$\mu_i^{\xi_j,n+1} = \mu_i^{\xi_j,n} - \frac{\Delta t_n}{\Delta x_i}\left(F_{i+1/2}^{\xi_j,n} - F_{i-1/2}^{\xi_j,n}\right), \tag{4.10}$$

where $\mu_i^{\xi_j,n+1}$ is the probability of finding an agent in cell C_i with a labelled identity ξ_j at time t_{n+1}, and $F_{i+1/2}^{\xi_j,n}$ is the flux at the boundary between cells C_i and $C_i + 1$ for agents with labelled identity ξ_j at time n.

The Rusanov flux is defined as,

$$F_{i+1/2}^{\xi_j,n} = \lambda_{i+1/2}^n \mu_{i+1/2}^{\xi_j} - \frac{|\lambda_{i+1/2}^n|}{2}\left(\mu_{i+1}^{\xi_j,n} - \mu_i^{\xi_j,n}\right), \tag{4.11}$$

where the flux speed, $\lambda_{i+1/2}^n$, exerted at the $x_{i+1/2}$ edge of cell C_i, and the interpolated boundary mass $\mu_{i+1/2}^{\xi_j,n}$, are

$$\lambda_{i+1/2}^n = \sum_{k,l} \Delta x_{k+1/2}\Delta\xi_l \cdot w_{i+1/2,l}\phi(x_{i+1/2}, x_{k+1/2})\mu_{k+1/2}^{\xi_l,n}, \tag{4.12}$$

$$\mu_{i+1/2}^{\xi_j,n} = \frac{\mu_{i+1}^{\xi_j,n} + \mu_i^{\xi_j,n}}{2}. \tag{4.13}$$

The Equation 4.12 for flux speed can be understood as a discretisation of the integrand in the solution for an inexchangeable mean-field limit (Equation 4.4) multiplied by the associated discretisation weights $\Delta x_{k+1/2}$ and $\Delta\xi_l$.

The interaction function ϕ that will be focused on is the Attraction-Repulsion Model defined in Axelrod et al. (2021) and used in the previous

agent-based model (Equation 3.1, 3.2), for completeness, that is,

$$\phi(x_i, x_j) = \begin{cases} R(x_j - x_i)\theta(x_i, x_j), & \text{for } \|x_i - x_j\| \leq T, \\ -R(x_j - x_i)\theta(x_i, x_j), & \text{for } \|x_i - x_j\| > T, \end{cases}$$

where θ is the exposure rule,

$$\theta(x_i, x_j) = (1/2)^{\frac{|x_i - x_j|}{E}}.$$

The *tolerance*, T *responsiveness* R, and *exposure* E, parameters are thus all included in the interaction function. The weights between identities, w, are initially set to 1 for simplicity given that θ is akin to an edge weight based on opinion for a complete graph. An alternative formulation would be to mediate exposure between identities rather than opinions, thus varying values of w for ξ. ζ, and is a future avenue of work.

The flux at the boundaries is imposed as being equal to zero to ensure mass conservation, as discussed in the method overview, this happens at $i = 1/2$ and $i = N - 1/2$.

4.3 Finite Volume Simulation Procedure

The simulations in Sections 4.4 and 5.2 were run using a Fortran code implementing the finite volume scheme of the mean-field approximation. The Fortran implementation was developed at Inria and Sorbonne Université (Laboratoire Jacques Louis Lions) by Nathalie Ayi, Francesco Cornia and Jacques Sainte-Marie. Additional details can be found in a forthcoming publication.

The Fortran code was then adapted to the specificities of the Attraction-Repulsion Model. Precisely, the interaction function ϕ was changed to implement the exposure rule as in Equation 3.1, the opinion update rule as in Equation 3.2, as well as Attraction-Repulsion Model parameters (tolerance T, responsiveness R, and exposure E). Initial conditions were also coded in Fortran to implement different distributions of opinions, such as a Uniform distribution and a bimodal distribution as the sum of two Gaussian distributions.

The Fortran model was then put in a Python wrapper to facilitate multiple model runs under sets, S, of model parameter combinations and different starting distributions. This framework allows for the finite volume method model to be easily deployed by Python, while maintaining the performance

benefits of Fortran for the computations to simulate the opinion distribution in time.

When running simulations, parameter sets of interest are defined in a Python wrapper, which also creates output files for simulation data, and the wrapper then calls the routine in Fortran. The simulation procedure is described in pseudo-code in Algorithm 2.

Algorithm 2 Simulation procedure for the finite volume method under the ARM with parameter sets, S, of E, T, R, model parameters.

```
Compile Fortran model
Begin Python wrapper
for each s ∈ S do
    Pass s(E, T, R) to Fortran input file
    Create simulation output file
    Call Fortran model with input file
    while t < t_max do
        Compute time interval, Δt_n
        Compute fluxes, F_{i+1/2}^{ξ_j,n}
        Implement boundary conditions
        Update density, μ_i^{ξ_j,n+1}, and save
```

4.4 Finite Volume Simulation Validation

Now that the method and its implementation have been described, it must be tested to ensure its usefulness for simulating social systems. A complete verification of the stability of the finite volume method would require proceeding analytically – that is, functional analysis verifying that the functions are conservative and that errors are not propagated through the simulation process. While another approach is to conceive a number of settings in which there are concrete expectations as to how the simulation should behave. This is the approach taken to validate the model, which will test for the presence of expected behaviours in controlled settings. The expected behaviours are informed by the understanding of the Attraction-Repulsion Model from Chapter 3 and the original analysis in Axelrod et al. (2021). These controlled experiments will show that the model is behaving as expected under a variety of conditions, and so providing confidence in the results of applications.

Validation Experiment	Objective	Expected Behaviour	Starting Distribution	Param. Values
(1) Baseline	Baseline case from Axelrod et al.	Change in the model should be smooth.	Uniform	$T = 0.25$, $R = 0.25$, $E = 0.1$
(2) Response	Testing R	Lowering R will slow the rate of change, and outcome remains as in (1).	Uniform	$T = 0.25$, $R = 0.05$, $E = 0.1$
(3) Exposure	Testing E	Lowering E will slow the rate of change, and change the outcome of (1).	Uniform	$T = 0.25$, $R = 0.25$, $E = 0.02$
(4) Consensus	Testing attraction for T	Increasing T will collapse the distribution to consensus.	Uniform	$T = 1.0$, $R = 0.25$, $E = 0.1$
(5) Extremes	Testing repulsion for T	Lowering T will push the population to the extremes of the space.	Uniform	$T = 0.05$, $R = 0.25$, $E = 0.1$
(6) Groups	Consistency of groups	Groups will remain intact but distinct from each other	Bimodal	$T = 0.25$, $R = 0.25$, $E = 0.1$

Table 4.1: Validation protocol for the finite volume method. The stated experiments test if the fundamental behaviours of the model are as expected in order to demonstrate the model's validity.

Table 4.1 details the protocol for validation, while the six experiments constituting the protocol are shown in Figures 4.4–4.10. The figures are presented as snapshots of the simulations over time up until $t = 200$, which is not sufficient time for convergence of all simulations but instead provides well-spaced snapshot times to check simulation behaviours. Snapshots are taken at the closest evaluation point to $t = 0, 40, 80, 120, 160, 200$; this may be at $t = 39$ rather than 40 because the flux speed $\lambda^n_{i+1/2}$ in an experiment is different depending on the parameter values, and so the CFL condition of Equation 4.6 determining t_{n+1} changes, which results in different time step evaluations for stable simulations. The vertical axis of each snapshot

is $\mu_t(x) = \frac{1}{P}\sum_{j=1}^{P}\mu_t^{\xi_j}(x)$ where $P = 2$, which is the probability of opinion x – on the horizontal axis – at time t. Each validation experiment begins with a uniform distribution, except the final experiment which necessitates a bimodal distribution. The space discretisation is set such that $M = 200$, and therefore there are 200 evenly spaced x_i in the opinion space.

For a baseline experiment, the default values of parameters T, R, and E, from Axelrod et al. (2021) are selected. Parameter values need not necessarily map between their work and the finite volume method model but top-level behaviour should be the same; and, in fact, some level of polarisation occurs as in the original article. This experiment is '(1) Baseline' in Table 4.1, and the first of the series of validation figures presented – Figure 4.4.

Changing responsiveness, R, the outcome is as expected: opinion change happens at a slower rate but in the same manner. This can be seen in Figure 4.5 since it is a slower version of the baseline experiment in Figure 4.4. Note that R is one fifth of the default value from baseline experiment (1), and the distribution in experiment (2) takes five times the amount of time to reach the same shape (comparing $t = 200$ in Figure 4.5 with $t = 40$ in Figure 4.4). For clarity, this validation experiment would fail if the expectation was not met, *i.e.* rate of opinion change was not slower.

Validation experiment (3) tests exposure, E. Resultant behaviour is more complex than for experiment (2) since lowering E slows the experiment *and* changes the behaviour. For an opinion x, relative exposure to opinions that are increasingly different reduces as the parameter also reduces. Figure 4.6 displays behaviour that is attractive in the local vicinity of an opinion and is not exposed to opinions that would result in repulsion. Local attraction dominates since the weight under the exposure rule (Equation 3.1) at an opinion difference of 0.25, where interaction flips from attraction to repulsions, is 0.0002 for $E = 0.02$ – significantly different to 0.1768 for the default $E = 0.1$.

The last parameter to test is tolerance, T, by checking that what should be either predominantly attractive or repulsive interactions are reflected as such in the experiments. Validations (4) and (5) investigate T by first increasing tolerance for total attraction across the opinion space and then decreasing the parameter to divide the population towards the extremes. Both cases behave as expected in Figures 4.7 and 4.8. In the case of low tolerance, when most of the population is repulsed by each other, one might expect to observe the mass at $x = 0.5$ to remain unchanged since the repulsion felt

from either side of this point should be balanced. However, this is not what is observed in Figure 4.8. The lack of a stationary middle point is clear when considering Equation 4.8, the addition of the flux terms is non-zero since $F^n_{i+1/2} = -F^n_{i-1/2}$ in this symmetric experiment and so the density will change. Closer inspection of the changes to the distribution under repulsion at the start and the end of the experiment are in Figure 4.9. The closer inspection reveals that the uniform distribution does not remain flat as might appear in the zoomed-out Figure 4.8. The curve of the distribution is expected, as x is more distant from 0.5 the imbalance of repulsion experienced in a direction becomes greater and so more mass is lost. Finally, The sudden jump in the distribution is at $x = 0.05$, which is where repulsion is only felt in one direction.

The final validation is to test that groups are consistent, and the results can be seen in Figure 4.10. The two modes quickly collapse to two distinct points, and then slowly move apart as they are repulsed by each other, which is exactly the desired behaviour.

To summarise the validation experiments, they provide evidence that the equation, the scheme, the implementation, and the simulation, are consistent with the expected behaviours, as informed by the understanding of the Attraction-Repulsion Model from Chapter 3 and the original analysis in Axelrod et al. (2021). The validations provide sufficient confidence with which to extend simulation work into considering applications of the method and having trust in the validity of the results and findings of the work pursued.

4.5 Complexity of Finite Volume Method for Attraction-Repulsion Model

For a grid of points on x consisting of M cells, the flux between each point on the grid and every other point (Equations 4.11, 4.12) is calculated and therefore has a time complexity in $\mathcal{O}(M^2)$. The flux calculation is also made across each of the identities, ξ, of which there are P. Since each identity is compared to the others, the time complexity from this operation is in $\mathcal{O}(P^2)$. Equation 4.13 to interpolate the density at the boundary is also necessary to execute in the model, however it does not effect the overall complexity since the operation is in $\mathcal{O}(M)$ time complexity because it requires one pass over the grid of points of size M. A dynamic group identification process, as in Chapter 3, is not currently implemented for the finite volume method model so no other steps need to be considered. Overall complexity is therefore in $\mathcal{O}(P^2 \cdot M^2)$, representing each identity ξ_j comparing to the other identities and the flux between each x_i being calculated. This holds for both time and space complexity.

Mapping between complexity inputs for the agent-based model and the finite volume method model is a complex problem because they measure different things. Therefore a formal link between the number of n agents and an approximation by M mesh cells and P identities is not clear and not provided here.

One approach to providing an effective complexity comparison between the models is to consider survey methodology. It is common for surveys to use scales (such as Likert scales) to divide the opinion space into a grid of options, ranging from five to eleven options. In the case of eleven options within the opinion space, the implied discretisation of the interval $[0, 1]$ would be approximately 0.091. If an eleven-step Likert scale is considered sufficient to capture the opinions of a population of $N = 10,000$ agents then this implies $M > 11$ is equivalent. In the simulations of Sections 4.4 and 5.2, $M = 200$ and so intervals are far smaller than the equivalent eleven-point survey scale that is deemed sufficient for $N = 10,000$ agents. This approximate comparison does not consider the density of individuals per cell for the survey or the simulation so should not be considered as a rigorous comparison.

The key advantage of the finite volume method model is that the complexity does not depend on the number of agents n, but on M and P. The bottle neck in the agent-based model is at the interaction and updating stage

which introduces $\mathcal{O}(n^2)$ complexity and this is avoided in the finite volume method model. The resulting model for large populations is still of squared complexity, but it is dependent on the precision of the grid of points in x and ξ rather than the size of the population. Therefore providing a framework with which to model large populations that an agent-based model could not manage. So, when the number of agents n is no longer manageable for simulation then the finite volume method model enables simulation to continue.

A further advantage of the finite volume method model is that it is deterministic so it only needs to be run once, while the agent-based model is stochastic in the exposure rule and so must be averaged over multiple experiment runs.

The efficiency improvements in the finite volume method allow for modelling much larger populations that are represented as distributions rather than individuals. One application is then the possibility of modelling large social media data or survey data; while the method also enables extensive exploration of model parameter space that is computationally infeasible under the associated agent-based model. In Section 5.2 the finite volume method will be employed to enable simulation across the Attraction-Repulsion Model parameter space in order to identify if the model can produce behaviour that is plausibly similar to that of empirical distributions.

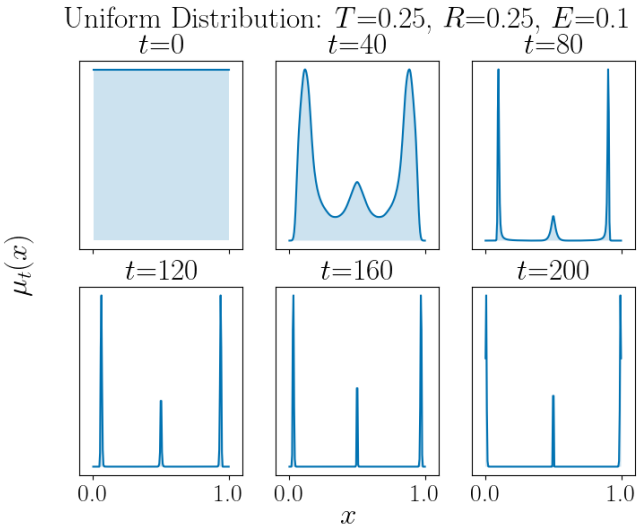


Figure 4.4: These are the default parameter values given in Axelrod et al. (2021). As the finite volume method considers the mesh on x rather than individual agents, it is not necessary that behaviour at parameter values exactly align. However, the variance of the distribution at $t = 200$ is 0.20, which is approximately the same as in their results.

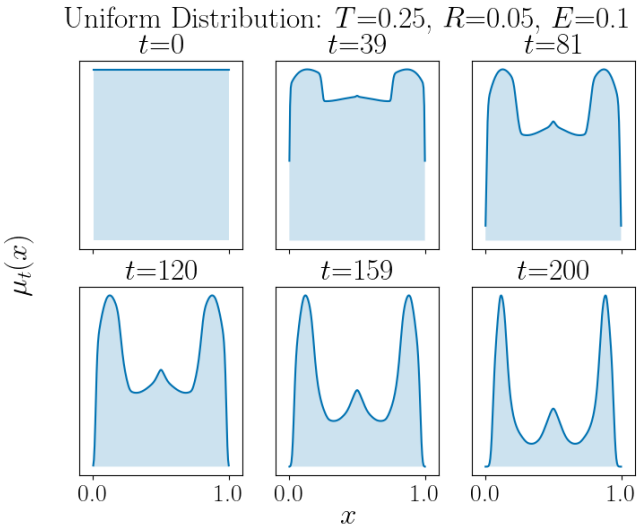


Figure 4.5: Lowering responsiveness results in a slower version of the baseline experiment in Figure 4.4. Shape and evolution of the distribution follow a similar behaviour.

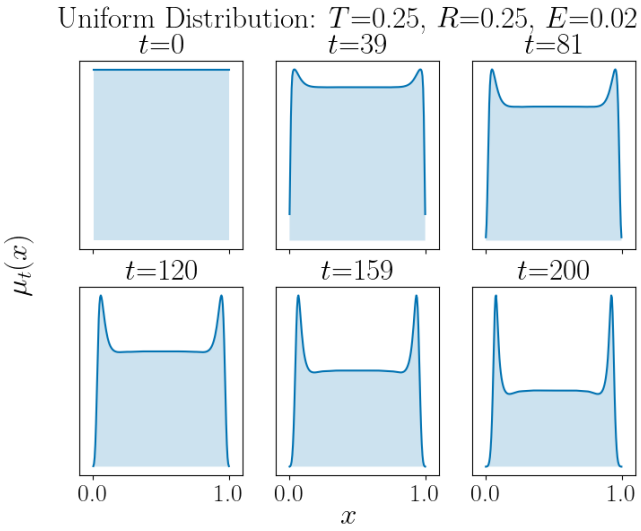


Figure 4.6: Lowering exposure slows the experiment, but also changes the behaviour. There is little evidence of the repulsion that causes the central mode in experiment (1) since exposure to 'distant' opinions is relatively more rare. Instead, the distribution at the limits bunches to local points.

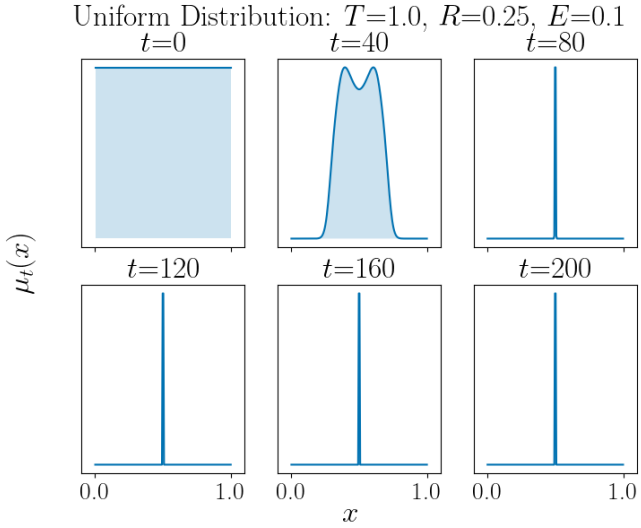


Figure 4.7: High tolerance results in the distribution collapsing to a shared consensus point at $x = 0.5$. The whole distribution is attracted to itself: initially creating two sub-groups, seen at $t = 40$ and then a sole group concentrated at a single opinion.

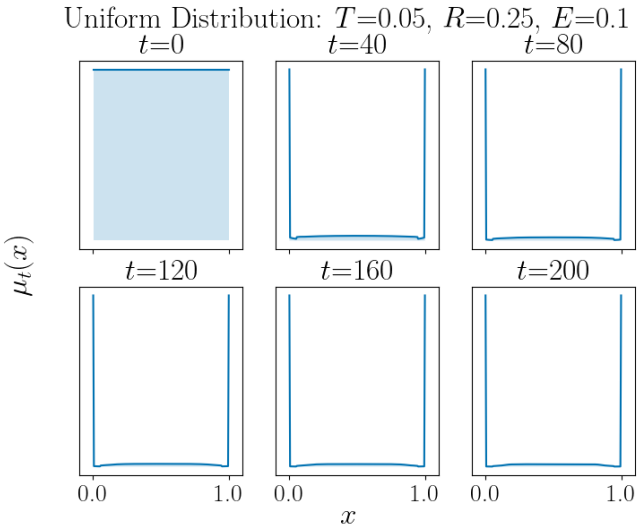


Figure 4.8: With low tolerance the population is pushed to the extremes. Please see Figure 4.9 for closer inspection of what happens to the density that lies between the extremes of the space.

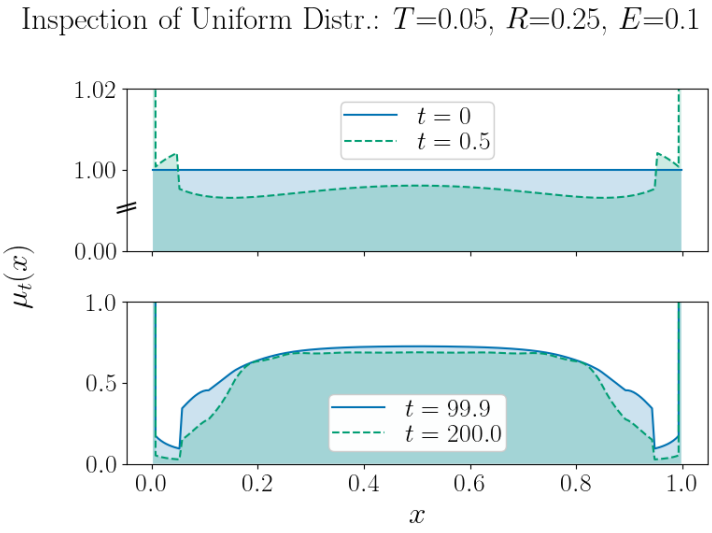


Figure 4.9: An inspection of certain time steps in Figure 4.8. Mass is lost from the centre of the distribution to the extremes. Mass is lost quicker as x moves away from the mid-point of the range, which reflects an increasing imbalance in repulsion between left-hand-side and right-hand-side, until attraction at the extremes of the space.

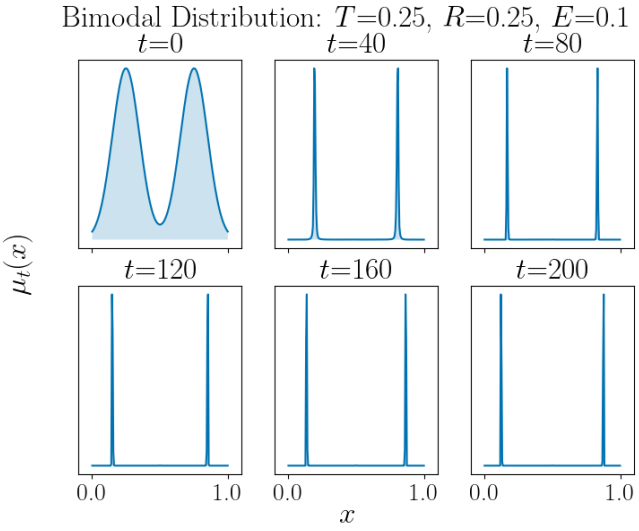


Figure 4.10: A bimodal starting distribution using the same parameters as in the baseline experiment (1) of Figure 4.4. The two groups collapse to their respective consensus points and then slowly separate to the extremes.

Chapter 5

Applications

This chapter presents results from the application and simulation of the two methodologies that have been described in Chapters 3 and 4. The two variations of the Attraction-Repulsion Model address two gaps in research, detailed in Chapter 2, within opinion dynamics relevant to a disconnection with social theory and a lack of empirical relevance, respectively.

First, the modified Attraction-Repulsion Model with group identification, presented in Chapter 3, will be treated. The purpose of including group identification in an opinion dynamics model is to reflect the importance of social identity in theory (Tajfel 1974) and to address the disquieting increase of affective polarisation (Garzia et al. 2023), which relies on in-group and out-group perception. This framing of group, rather than individual, offers a fresh perspective for opinion dynamics.

Simulations from the group identification model will be used to answer research questions: (a) where does different treatment of in-group and out-group result in opinion evolution tending towards polarisation or consensus, (b) is a recognition of in-group and out-group important for understanding polarisation, and (c) does the shift from a focus on individuals to groups impact existing framing of opinion dynamics research? Building on the high relevance of affective polarisation and group identity in the social science literature, this is an important exploration of the impact on existing models.

The second method, taking the mean-field limit of the original Attraction-Repulsion Model and applying the finite volume method as presented in Chapter 4, allows for the treatment of large populations (as represented by opinion distributions), in a step towards closing the empirical gap of opinion dynamics. More precisely, the empirical gap is the missing connection

between the well-developed theoretical understanding of models and limited experimental validation (Carpentras et al. 2022). This greatly impinges on the usefulness of the models when trying to draw conclusions that speak to modelling real-world questions of opinion distributions.

Application on the finite volume method will be focused on considering the falsifiability of models, otherwise stated as considering what the plausible parameter space of the model is – are distributions that result from simulation approximately similar to what is observed in the real world? The principal question of the application is then, is a framework provided by which opinion models can be falsified or deemed plausible for empirical data?

5.1 Group Dynamics of Polarisation

5.1.1 Experimental Protocol

Simulations of the group-dependent ARM, detailed in Chapter 3, will be undertaken with various parameter combinations to assess the impact of group identification within the model. The initial conditions of the model will be constant across simulations: the number of agents will be set to $N = 100$ as in Axelrod et al. (2021); the starting distribution of opinions will be bimodal (as in Figure 3.3) on the assumption that group identification is already present in the population and that a question for the model is the degree of exacerbation of polarisation under parameter combinations. For initialising the bimodal distribution, half of the agents draw opinions from a Normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ with $\mu_1 = 0.33$ and $\sigma_1 = 0.05$, while the remaining 50 agents draw opinions from a Normal distribution $\mathcal{N}(\mu_2, \sigma_2^2)$ with $\mu_2 = 0.67$ and $\sigma_2 = 0.05$. An alternative initialisation of opinions could be a unimodal or uniform distribution, however these cases would reflect one group or no group presence in the population; this is of less relevance when questioning the relevance of in-/out-group interaction differences. Note that the initial distribution of opinions is always randomly drawn for each simulation run of the model.

From this shared starting point for simulations, each parameter (*tolerance* T, *responsiveness* R, and *exposure* E) is taken in turn and split into an in-group and out-group version. The group-dependent parameters are varied in twenty 0.05 increments from 0.05 to 1.0, while the non-group-dependent parameters are kept constant at values within the same $[0.05, 1]$ range – this

is named as one experiment. Each simulation run within an experiment lasts up to one thousand iterations or is terminated early when opinions have not changed over one hundred consecutive iterations, and each simulation run is repeated twenty times in order to arrive at average behaviour given the non-deterministic nature of interaction choices between agents.

Nine experiments are then run for each group-dependent parameter pair to investigate the range of system behaviours. In other words, when tolerance is group-dependent how does the role of the group-dependent parameter change when responsiveness is low/medium/high and when exposure is low/medium/high. The precise set of parameters explored for experiments are:

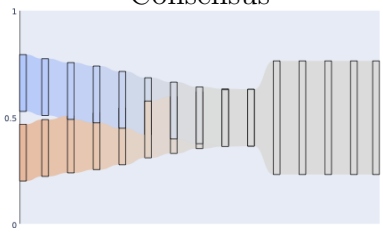
- Group-dependent tolerance (GDT) simulations:
  $T_{\text{in}} \in [0,1], T_{\text{out}} \in [0,1], R \in \{0.01, 0.1, 0.25\}, E \in \{0.01, 0.1, 0.25\}$;
- Group-dependent responsiveness (GDR) simulations:
  $T \in \{0.01, 0.1, 0.4\}, R_{\text{in}} \in [0,1], R_{\text{out}} \in [0,1], E \in \{0.01, 0.05, 0.25\}$;
- Group-dependent exposure (GDE) simulations:
  $T \in \{0.01, 0.1, 0.4\}, R \in \{0.01, 0.05, 0.25\}, E_{\text{in}} \in [0,1], E_{\text{out}} \in [0,1]$.

Once the experiments of simulations have run, the level of polarisation within the population is determined by the DER measure averaged over the final one hundred iterations (Equation 3.3) and used to report on outcomes. The application of the group-dependent model will be unfolded by first discussing behaviours possible in simulations, and then diving into what conditions produce different levels of polarisation within experiments.

5.1.2 Group Polarisation Behaviours Achieved by the Model

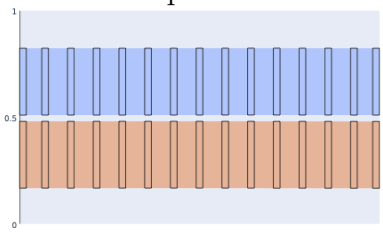
It is useful to qualitatively address the typical dynamics that can be found during model simulation by proposing a typology of combinations of some group properties of the system. The four dynamics displayed in Figure 5.1 are discerned from the range of simulation experiments found in the remainder of this Section 5.1, since they explain how the simulations arrive at their final states of polarisation. Convergence of the agents' opinions to consensus (Figure 5.1a) or divergence to complete polarisation (Figure 5.1c) are primary behaviours found in many works on opinion dynamics. Further to these two

Consensus



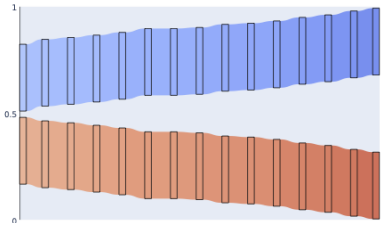
(a) $T_{\text{in}} = 0.2$, $T_{\text{out}} = 0.4$, $R = 0.01$, $E = 0.1$

Stable polarisation



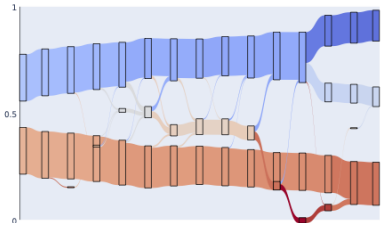
(b) $T_{\text{in}} = 0.2$, $T_{\text{out}} = 0.4$, $R = 0.01$, $E = 0.01$

Unstable polarisation (group drift)



(c) $T_{\text{in}} = 0.2$, $T_{\text{out}} = 0.1$, $R = 0.01$, $E = 0.1$

Unstable polarisation (fragmentation)



(d) $T_{\text{in}} = 0.05$, $T_{\text{out}} = 0.1$, $R = 0.01$, $E = 0.1$

Figure 5.1: Representative simulation of four macro-scale types of group dynamics arising from the simulation of the group-dependent Attraction-Repulsion Model, achieved with different values for the parameters of the model: in-group and out-group tolerance, responsiveness, and exposure. The four group behaviour types are groups converging (consensus), groups remaining separate with a constant distance between them (stable polarisation), groups diverging from each other (group drift), and groups splintering into sub-groups (fragmentation).

polar opposites, the model also displays stable but not complete polarisation (Figure 5.1b) and an unstable polarisation that is characterised by group fragmentation (Figure 5.1d) – a unique understanding of model behaviour afforded by the lens of group identification dynamics.

Consensus in Figure 5.1a is achieved when agents are exposed to opinions belonging to another group and they are sufficiently tolerant of the difference with out-group opinions, the resulting interaction is attraction. This process repeats until the population eventually arrives at a shared opinion and two groups become one.

Stable polarisation in Figure 5.1b is a result of low out-group exposure. If the opinions of an out-group are rarely observed then in-group interactions become the only important dynamic within the model. As such, the groups neither repel nor attract, but remain polarised to an extent which does not become more or less extreme as they remain in their own bubble.

Unstable polarisation can also occur within the model simulations. The first kind of unstable polarisation is group drift, seen in Figure 5.1c, in which groups move towards the extremes of the opinion space as they are exposed to out-group opinions that are above the tolerance threshold and so result in repulsion.

The second kind of unstable polarisation is group fragmentation, which is only understandable with a group identification framework such as that suggested in this group-dependent model extension. Figure 5.1d displays groups splitting from within as driven by in-group repulsion. Given that agents are not tolerant of the spread of in-group opinions, the group fragments and results in smaller groups of agents floating between the extreme groups. The splinter group agents may then evolve to rejoin the previous group through attraction/repulsion, join a new group, or remain in the middle of the opinion space. With multiple groups, the maximum level of opinion polarisation is never achieved since not all agents will reach the extremes of the opinion space.

5.1.3 Group-Dependent Tolerance Experiments

The experiments for the case of introducing group-dependent tolerance into the model can be found in Figure 5.2, each pixel on each heatmap panel experiment of the three-by-three grid is the average final DER polarisation of the agents' opinions across twenty simulation runs for a combination of $(T_{\text{in}}, T_{\text{out}}, R, E)$. Depending on the combination, resulting polarisation is

somewhere on the DER measure scale between 0 (low: consensus) and 0.5 (high: complete polarisation). Heatmaps representing experiments will be compared to each other, then insight into the group-dependent roles of T_{in} and T_{out} will be discussed.

When the responsiveness parameter, R, is fixed (comparing heatmap panels vertically), higher exposure tends to raise the out-group tolerance threshold, T_{out}, at which low polarisation occurs. Thus with relative increased exposure to agents with greater opinion differences, agents need a larger tolerance of out-groups for opinion consensus to occur.

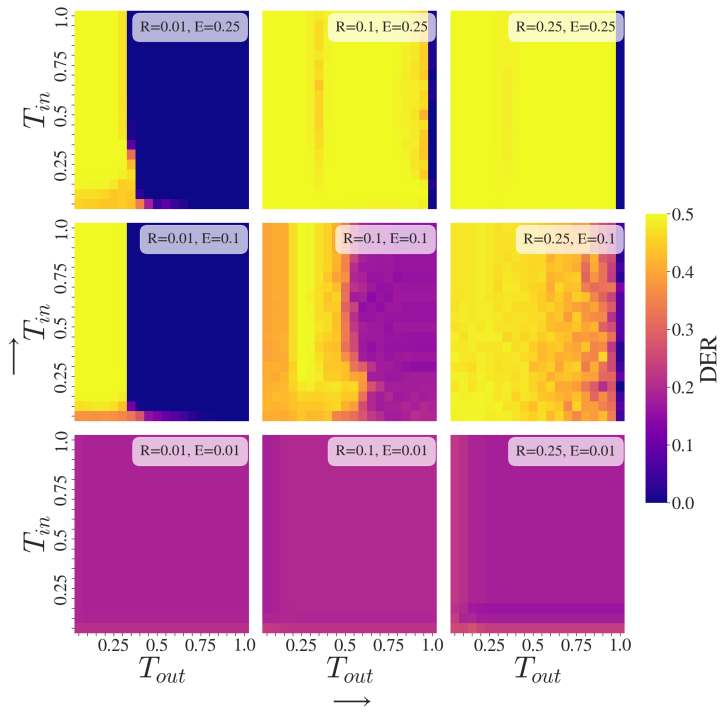


Figure 5.2: Group-dependent tolerance ($T_{\text{in}}/T_{\text{out}}$) experiments. For determining polarisation outcomes of the simulation, T_{out} is the more important parameter. Tolerance towards the out-group decides whether two groups will be attracted to each other or separate, although exposure E must be high enough that out-group opinions are observed often enough to impact the in-group.

If the exposure parameter, E, is instead considered as fixed (comparing heatmap panels horizontally), then increasing responsiveness results in in-

creased polarisation. With high responsiveness, agents explore the opinion space quickly and find the extremes at 0 and 1, from positions at the extremes of the opinion space it is difficult for an agent to be attracted to any out-group opinion unless T_{out} tends towards 1. Similar results to these findings for E and R are reported by Axelrod et al. (2021) in the original work without groups.

Changing focus to the in-/out-group parameters, $T_{\text{in}}/T_{\text{out}}$, out-group tolerance is a more influential parameter than in-group tolerance. The vertical boundaries between areas of high polarisation and low polarisation – for example, in the ($R = 0.01$, $E = 0.1$) experiment where $T_{\text{out}} > 0.3$ – tend to define system behaviour transitions between groups merging or groups diverging. So, tolerance of out-group is a key predictor of the resulting opinion distribution from a simulation.

In-group tolerance can still impact polarisation outcomes if it is sufficiently low that it causes in-group fragmentation. This effect can result in multiple groups spread between the extremes meaning that maximum polarisation is not achieved. An example of parameters for which in-group fragmentation occurs can be seen at $T_{\text{in}} = 0.05$ with $T_{\text{out}} \leq 0.5$ in the ($R = 0.01$, $E = 0.1$) experiment, where DER is approximately 0.3. For higher R values, in-group fragmentation still occurs early in the simulation but higher responsiveness results in larger opinion changes meaning that middle ground agents are able to arrive at the extremes of the opinion space and break out of the middle ground.

Some experiments in Figure 5.2 – such as the ($R = 0.25$, $E = 0.1$) experiment for $0.5 \leq T_{\text{out}} < 1.0$ – show non-smooth transitions between final polarisation states when changing parameter values (this also occurs later, in Figures 5.3 and 5.4 for group-dependent responsiveness and group-dependent exposure). This non-smoothness will be termed as *pixelation*. It is due to volatility in the final DER measure of the polarisation of the agents across the twenty simulation runs, which is caused by one of two reasons: (1) First, group fragmentation may occur which results in different possible configurations of middle ground agents and thus a variety of possible DER values. (2) The second cause of pixelation is that under certain parameters the simulation has several possible stable states, or "local minima", that are arrived at with non-negligible probability and so the average outcome for simulations is less predictable.

An example of the second cause is the experiment with parameters $T_{\text{in}} = 0.6$, $T_{\text{out}} = 0.4$, and ($R = 0.1$, $E = 0.1$); seventeen out of the twenty simu-

lations end with DER over 0.40 which is equivalent to the agent population being split between the two extremes of the opinion space in approximately equal group sizes. While the other three simulations end in less polarised (stable) distributions, the extreme being one simulation with a 0.25 DER value due to eighty-four agents having opinion 1 and the rest having opinion 0. This range of simulation outcomes is due to parameter combinations that can result in large opinion changes for agents each iteration, making the system less predictable. An alternative consideration of this cause, is that the standard deviation of final state DER polarisation across the twenty simulations is higher in the regions of Figure 5.2 where pixelation is more present. A further reflection of this is the sharp boundary in polarisation outcomes when varying T_{out} at $R = 0.01$, since low responsiveness causes only small, gradual, changes in opinion for agents.

5.1.4 Group-Dependent Responsiveness Experiments

The group-dependent responsiveness parameter, explored in Figure 5.3, governs the magnitude of attraction or repulsion when agents update their opinion in response to interactions. Low values of R mean slow convergence or divergence of opinions, while high values result in agents pushing to the extremes of the opinion space quickly when repulsed, but it can also result in agents under attraction overshooting the opinion that they are attracted towards. An example of overshooting would be if an agent were to observe a different opinion value twice with $R > 0.5$ while not observing other opinions; the agent would change their original opinion by over half of the opinion difference twice and so overshoot the referenced new opinion. Although overshooting is unlikely in real settings, it is presented here for a full exploration of the model.

When exposure is low, $E = 0.01$, (the lower horizontal band of three experiments in Figure 5.3) there is a tendency for polarisation to decrease as tolerance increases, no matter the level of $R_{\text{in}}/R_{\text{out}}$. In the experiments where $T \in \{0.1, 0.4\}$, there is a smooth transition between higher polarisation for $R_{\text{in}} > 0.5$ and lower polarisation for R_{in} values below 0.5. The understanding for this is that increasing R_{in} results in strong agent reactions to the frequent interactions that are occurring within a group, akin to increased energy in a system, so that the population's opinions do not settle to a constant value.

For intermediate exposure levels, $E = 0.05$, large out-group responsiveness ($R_{\text{out}} > 0.5$) brings polarisation, although this is less certain an outcome

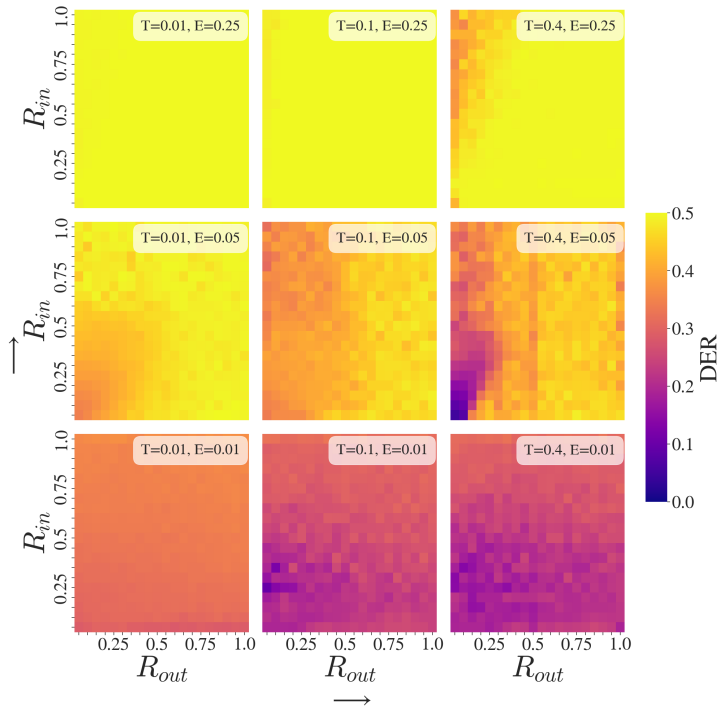


Figure 5.3: Group-dependent responsiveness ($R_{\text{in}}/R_{\text{out}}$) experiments. In contrast to the sharp behaviour boundaries of ($T_{\text{in}}/T_{\text{out}}$) experiments, the magnitude of attractive or repulsive response results in gradual changes to the final polarisation of the population. Responsiveness to out-group opinions R_{out} is more influential than treatment of the in-group.

as tolerance increases. For in-group responsiveness, pixelation makes it unclear if low R_{in} results in lower polarisation in experiments $T = 0.1, 0.4$, however it is the case that lower R_{in} reduces polarisation for the $T = 0.01$ experiment.

Under high exposure, $E = 0.25$, only simulations with low out-group responsiveness ($R_{\text{out}} < 0.2$) and high tolerance ($T = 0.4$) result in low polarisation, which arises from agents neither reacting strongly to repulsive out-group nor overshooting attraction to out-group. However, in most simulations frequent exposure to agents with greater opinion differences results in polarisation.

Generally, increasing in-group or out-group responsiveness will increase the polarisation of the agents' opinions, with the exception of the ($T =$

$0.4, E = 0.25$) experiment under low out-group responsiveness. The group-dependent responsiveness experiments do not exhibit the same sharp behaviour transitions as the group-dependent tolerance experiments, reflecting that responsiveness is a gradual change in magnitude of opinion update rather than the sharp threshold of attraction/repulsion governed by tolerance.

5.1.5 Group-Dependent Exposure Experiments

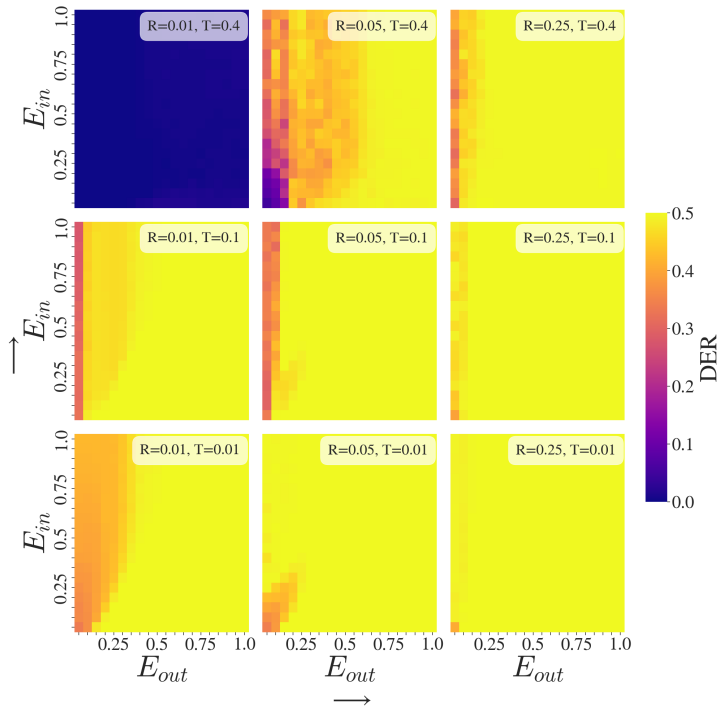


Figure 5.4: Group-dependent exposure ($E_{\text{in}}/E_{\text{out}}$) experiments. Polarisation is more likely for increased out-group exposure E_{out}, when the polarising effect of repulsion to different out-group opinions is not tempered by either high tolerance T or larger exposure to in-group E_{in}. The case of over-exposure to in-group opinions and under-exposure to out-group opinions can be considered as a similar scenario to filter bubbles.

Preferential exposure towards in-group agents over out-group agents resembles some aspects of a recommender system in social media settings, hinting at the much discussed topic of filter bubbles, while also emphasising

classic social dynamics such as homophilly. Of course, a recommender system may also include content-based recommendations or data on behavioural traces (*e.g.* clicks, likes, and shares, on a social platform), which are not present in group-dependent exposure.

For low tolerance experiments ($T = 0.01$) in Figure 5.4, an increase in E_{out} results in an increase of repulsion between groups as out-group repulsive interactions are more common. As tolerance increases, low polarisation final states become possible, and low polarisation are more likely arrived at when responsiveness is low too.

E_{in} can act as a balance to E_{out} when responsiveness is low. In this case the increased repulsion from increased exposure to out-group opinions that are different is countered by increased attraction to in-group opinions, and thus the net effect of the repulsion is reduced by the increased in-group attraction to the increased group average opinion.

The ($R = 0.01$, $T = 0.4$) experiment is the only experiment in Figures 5.2, 5.3, and 5.4, that reaches consensus for all combinations of the group dependent parameter. This shows that sufficient tolerance with small opinion changes from iteration to iteration lead to a stable, and predictable, consensus.

Across the group-dependent exposure experiments, when polarisation is possible, an increase in E_{out} increases the polarisation of the population's opinions. In a similar manner to group-dependent tolerance, it is clear that treatment of out-group is important to understand the potential for polarisation within the model.

5.1.6 The Impact of Groups in Opinion Dynamics

Introducing group identification to agents changes how the dynamics of a model may be understood. Rather than being focused on dynamics at the level of the individual, populations can be analysed by how they are structured by groups and how these structural elements combine or diverge to produce consensus or polarisation, as in Figure 5.1. This is a new direction within opinion dynamics that can provide further insight in future research.

Treatment of out-group is raised as a key factor in understanding polarisation outcomes of the population. Ultimately, relationship to out-group determines whether opinion consensus or polarisation occurs due to groups being attracted to or repulsed by one another. This finding sits alongside the current emphasis placed on affective polarisation by colleagues in social

sciences that understands the level of out-group hate as a central part of polarisation.

In-group dynamics also provide insight into population behaviours. If a group's spread of opinion is larger than in-group tolerance then the group will fragment from within, resulting in an increase in polarisation. There is also the balancing effect of more in-group exposure than out-group exposure which can reduce potentially polarising effects.

The relative roles of tolerance T, responsiveness R, and exposure E, have been discussed. Tolerance ultimately decides the eventual movement towards consensus or polarisation by an agent's attraction to, or repulsion from, other agents. Responsiveness has the role of introducing uncertainty of outcome under the non-deterministic model where large opinion changes can influence subsequent levels of polarisation. Exposure determines whether an agent observes relatively more different opinions or not, with the result that polarisation cannot occur when exposure outside of similar opinions does not occur (which is approximately similar to a filter bubble).

Considering the initial research questions posed on the relevance of introducing group identification for agents, it is clear that in-group/out-group differentiations can explain polarising behaviour and also present a promising lens with which to connect individual agents and an understanding of groups from social theory.

5.2 Falsification of Opinion Dynamics Models

The mean field approximation of the Attraction-Repulsion Model paired with the finite volume method now renders it feasible to search the parameter space of an opinion dynamics model such as the Attraction-Repulsion Model. Given the significant efficiency improvements explained in Chapter 4, simulation for the three-dimensional (T, R, E) parameter space of the model can be assessed rather than two-dimensional slices where one parameter is kept constant, as in Axelrod et al. (2021).

Broad exploration of the model parameter space produces a lot of evidence which can then be analysed to approach the potential falsification of the model or discover realistic behaviour. The definition for realistic behaviour that will be used here is that the opinion distribution can change a little from iteration to iteration but it is almost stable given that real-world opinion change is slow moving (Smith 1994) and attitude changes are typically long-term trends (Charlesworth and Banaji 2022) – this property will be termed *quasi-stable*.

The hypothesis for potential falsification is thus: The opinion dynamics model is useful if the model is able to produce a quasi-stable distribution that resembles the initial distribution under a set of parameters. Evidence, to potentially falsify the claim that the Attraction-Repulsion Model (the opinion dynamics model in question here) is useful, is produced by simulating the model and analysing the properties of these simulations for quasi-stableness. If the simulated distribution is similar to the initial distribution then the simulation will be considered *plausible*.

5.2.1 Experimental Protocol

The falsification procedure is presented in the following three steps: (1) all (T,R,E) combinations are simulated until they are stable, which is up to $t = 600$, (2) a stopping time $t_c \leq 600$ for the simulation is found, which will be referred to as the critical stopping time, when the rate of distribution change is small (precise criterion below), and (3) the difference between the starting distribution μ_0 and the quasi-stable distribution at the critical stopping time μ_c is calculated and compared to find what parameter combinations result in stable simulated distributions that resemble the initial distribution.

Simulations will begin with a bimodal distribution that is the sum of two Gaussian distributions centred at 0.25 and 0.75, both with standard de-

viation 0.1. The bimodal distribution provides an already mildly polarised opinion distribution which has similarities to distributions of interest in the real world. Further simulations for a unimodal and uniform starting distribution can be found in Appendix A. This starting distribution is similar to the modified agent-based model simulations of Section 5.1, however in that case the two Gaussians are centred at 0.33 and 0.66 with standard deviation 0.05. The resulting difference is that the boundary for behaviour change between consensus and polarisation is at $T = 0.5$ in Figure 5.10, while it appears around $T_{\text{out}} = 0.33$ in Figure 5.2 reflecting the different distances between modes of the initial distribution. The number of identities P is equal to two, but this does not feature as a factor in the falsification procedure currently, future work could consider what is realistic behaviour for identities. The number of cells M is 200, and therefore 200 evenly spaced x_i in the opinion space.

The results presented here are part of ongoing work, there may be future improvements to the complex task of determining what criteria reflect realistic behaviour. While alternative models to the Attraction-Repulsion Model could also be considered within the analysis pipeline by adjusting the interaction function ϕ in the general formulation of an agent-based opinion dynamics model (Equation 4.1). Therefore, while only the Attraction-Repulsion Model is analysed here, the process can be easily adapted to consider the plausibility of further models.

5.2.2 Plausibility and Falsification of Simulation

To begin, it is necessary to precisely define what a plausible simulation is by a real-world property. Not withstanding some external shock such as war, observed opinion distributions typically change by small amounts while maintaining consistency with previous observations (Fiorina and Abrams 2008; Richter et al. 2024). So, it seems that if a simulated distribution remains broadly similar over time then it is realistic. Replicating this property offers a definition for the plausibility of a simulation when compared to reality. There are two traits that characterise this property: the first is quasi-stability from iteration to iteration during simulation, and the second is a non-large deviation between the quasi-stable distribution and the initial distribution.

To put these two criteria into operation, a measure of difference between distributions must be chosen since it is necessary to compare μ_0 and μ_c as well as determine quasi-stability iteration to iteration. The Wasserstein distance

is chosen to measure the difference between distributions because it measures distribution difference as a combination of shape and location, *i.e.* do the distributions have similar spread and are the respective means or modes at similar points of the domain.

The metric is otherwise known as the earth mover's distance. It gained this alternative name due to an intuitive explanation of the measure: if two distributions a and b are represented by two distributions of piles of earth (soil), then the Wasserstein distance is the minimum work done to move earth in distribution a such that it now represents distribution b. The Wasserstein distance has been used previously to compare opinion distributions, but instead as a measure of polarisation by way of comparison of the observed distribution to a reference distribution with all mass split between the extremes of the opinion space (Lee and Sobel 2024).

Other measures of difference, or distance, between distributions are the Kolmogorov-Smirnov test statistic and Jensen-Shannon distance. The Kolmogorov-Smirnov approach compares cumulative distributions rather than mass distributions (such as an opinion distribution) and relies on calculating the point of maximum difference between cumulative distributions, therefore only considering the point of greatest difference between cumulative distributions rather than the comparative shapes. The Jensen-Shannon distance is a metric from information theory which does take into account the relative shapes of the distribution which could be alternatively used to compare the simulated opinion distributions. Wasserstein distance is preferred because it is well-defined in the case of the distribution being a Dirac delta function while the Jensen-Shannon distance is not – the relevancy being that consensus would take this shape for a continuous distribution of opinion so the Jensen-Shannon distance is avoided in the discrete simulated case too (Gibbs and Su 2002).

Using the Wasserstein distance, the metric to determine the plausibility of a simulation is $W(\mu_0, \mu_c)$. If the Wasserstein distance between the initial distribution and the quasi-stable distribution is less than 0.1 (meaning that the distributions are similar) then we will say the simulation is plausible, while if the distance is large then the model simulation must be false with regards to the realistic behaviour criteria.

The critical stopping time t_c at which the simulated distribution is considered quasi-stable, and compared to the initial distribution, must also be defined. An unchanging stable distribution, as observed at the end of simulation for $t = 600$, is unlike the desired slow distribution change observed in

reality; however, a rapidly changing distribution is also unrealistic. Therefore a threshold k for the value of the Wasserstein distance between iterations t_{n-1} and t_n must be decided upon to determine the condition for quasi-stability. The $n+1$ time step is defined by the CFL condition in Equation 4.6 which is dependent on the wave speed, potentially affecting the temporal evolution of metrics. The quasi-stable distribution μ_c is found at the first observation time t_n such that $W(\mu_{t_{n-1}}, \mu_{t_n}) \leq k$. This leaves the possibility that $W(\mu_{t_{n-1}}, \mu_{t_n})$ could increase above k in the future, but similarity from one observation to the next observation is chosen as the desired criteria in this case.

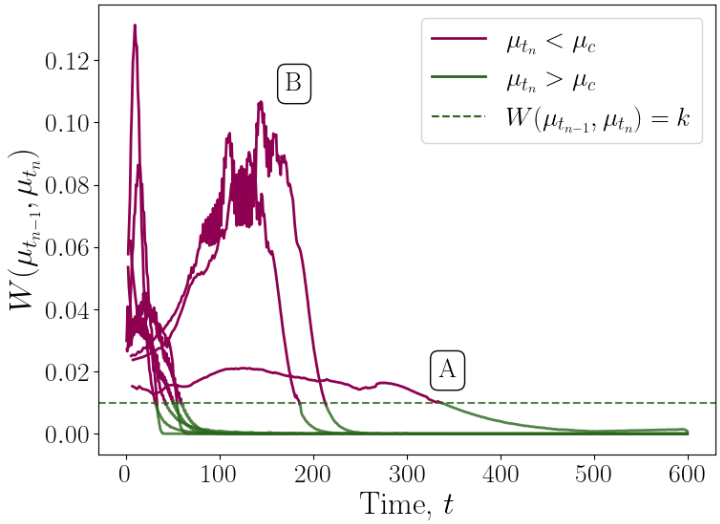


Figure 5.5: Evolution of the Wasserstein distance between $\mu_{t_{n-1}}$ and μ_{t_n} up to $t = 600$. Note that the plotted lines are smoothed with a moving average of window size five to improve legibility, the smoothing is not used in following results. The critical distribution μ_c is found at the first instance where $W(\mu_{t_{n-1}}, \mu_{t_n}) \leq 0.01$, marked horizontally on the figure. Curves are for $R = 0.1$, while $(T, E) \in \{0.1, 0.5, 0.9\} \times \{0.1, 0.5, 0.9\}$ to present a range of possible (T, E) combinations. Time snapshots of the curves labelled with 'A' and 'B' can be found in Figure 5.6

The critical threshold k is set equal to 0.01 with the desire that μ_c is relatively stable but still changing. Figure 5.5 displays how $W(\mu_{t_{n-1}}, \mu_{t_n})$ develops over time, with the threshold $k = 0.01$ marked. All shown simulations are with $R = 0.1$ since smaller R results is slower change of simulations,

so they are the simulations for which t_c will be larger. At $k = 0.01$, the simulations are settling into stability while further change is still possible, thus matching the desired behaviour for real-world similarity and being a suitable threshold for quasi-stability.

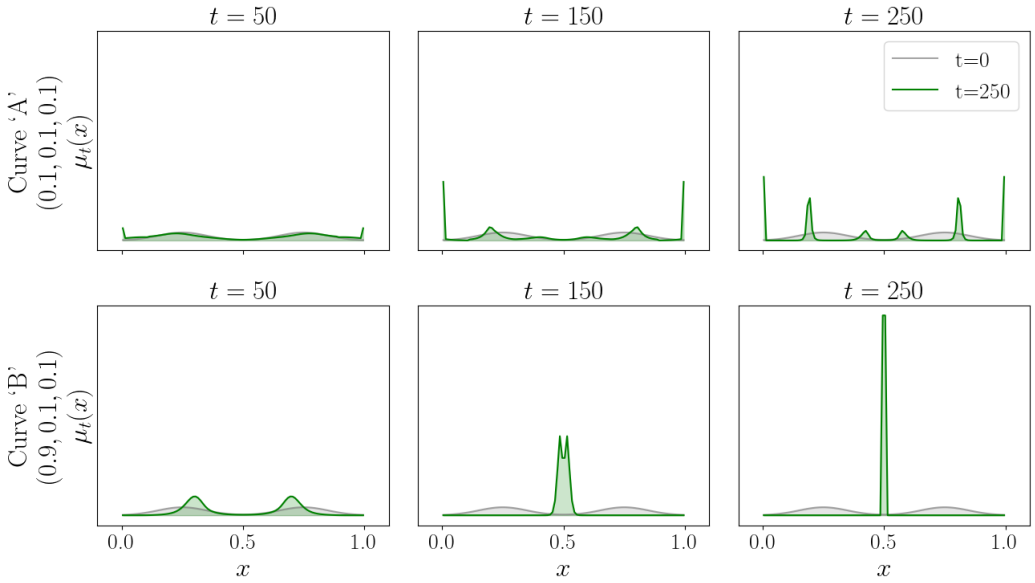


Figure 5.6: Time snapshots simulations for curves 'A' and 'B' from Figure 5.5 for (T, R, E) parameter simulations equal to (0.1, 0.1, 0.1) and (0.9, 0.1, 0.1), respectively. Curve 'A' represents a polarising simulation, while curve 'B' is a simulation that results in consensus. The grey curve in each panel is the initial distribution μ_0 for reference.

There is not a guarantee that the quasi-stable distribution μ_c has not changed significantly from the initial distribution μ_0 in the time up until t_c. The initial plausible simulations will be inspected if they are found – this is the approach taken here since the amount of the parameter space producing plausible simulations is small – and discounted as being plausible if the distribution deviates to no longer resemble the initial distribution before arriving at μ_c. In future work, a continuous check throughout the simulation procedure could be implemented such that if $W(\mu_0, \mu_{t_n})$ is too large for any t_n then the simulation is no longer plausible, although this would require precising a second threshold value along with k.

Time snapshots of $W(\mu_{t_{n-1}}, \mu_{t_n})$ curves labelled 'A' and 'B' from Figure

5.5 can be seen in 5.6. The 'A' case maintains lower $W(\mu_{t_{n-1}}, \mu_{t_n})$ values over time but arrives at t_c later than the 'B' case. This can be seen in the snapshots where simulation 'A' appears to have changed less than simulation 'B' at $t = 150$.

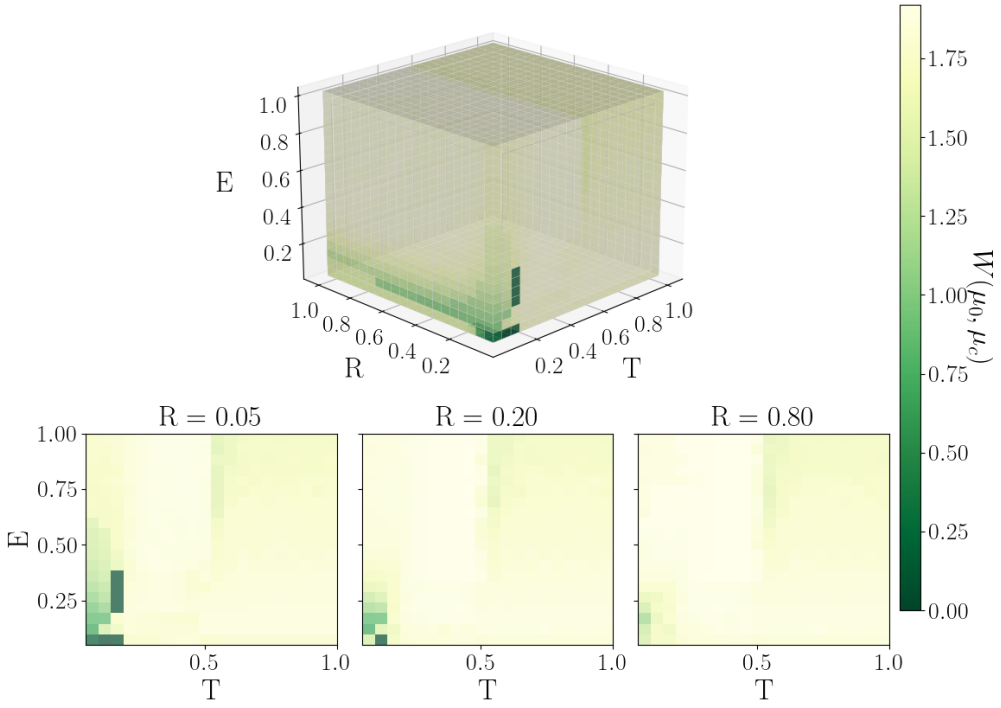


Figure 5.7: Comparing the initial distribution μ_0 with the quasi-stable distribution μ_c for simulations within the (T, R, E) parameter space. Two-dimensional slices of the cube are provided to provide further insight into the space. Model parameter combinations produce plausible simulations where $W(\mu_0, \mu_c)$ is towards 0, such as the regions ($T = 0.15$, $R = 0.05$, $0.2 \leq E \leq 0.35$) and ($T \leq 0.15$, $R \leq 0.1$, $E = 0.05$).

Now that μ_c is defined for each (T, R, E) simulation, $W(\mu_0, \mu_c)$ can be measured across the parameter space to assess where model simulations are plausible and where they appear to be false with regards to the criteria set

out in this work. The plausible simulation space is specific to μ_0 given that the initial distribution is the base comparison for the Wasserstein distance with μ_c.

Figure 5.7 displays $W(\mu_0, \mu_c)$ across the model parameter space. Some structural elements are present throughout the two-dimensional slices, such as the boundary at $T > 0.5$ where consensus will occur and the region where ($T \leq 0.15$, $E \leq 0.2$) which produces neither complete consensus nor polarisation. Comparison with the final polarisation state of simulations can be found later in Figure 5.10.

The two regions with the lowest $W(\mu_0, \mu_c)$ – ($T = 0.15$, $R = 0.05$, $0.2 \leq E \leq 0.35$) and ($T \leq 0.15$, $R \leq 0.1$, $E = 0.05$) – both have a μ_c that is found quickly (see Figure 5.8 for snapshots of μ_c with relevant critical times t_c for these two regions, and Figure 5.9 for t_c of the full space), within the first few iterations. With μ_c found quickly, the simulated distribution does not deviate away from the initial distribution and return to the shape so the concern of discounting plausibility is not borne out.

In the long-run the simulations in these regions converge to different distributions, but they produce plausible distributions in the early time steps of simulation while the model under different parameters does not. Short-term plausibility of the model is relevant to real-world distributions where similarity from one observation to the next observation is desirable. The simulation for $(T, R, E) = (0.1, 0.05, 0.05)$ is plausible over a far longer time, although this is due to low R and E meaning not much change in the distribution.

Of the two low $W(\mu_0, \mu_c)$ zones, the short-term plausible ($T = 0.15$, $R = 0.05$, $0.2 \leq E \leq 0.35$) is the more interesting region since interaction occurs between different parts of the opinion distribution and some balance between attraction and repulsion occurs, it seems that low responsiveness is also key to this possible outcome.

Where $W(\mu_0, \mu_c)$ is approximately equal 1, in the ($T \leq 0.1$, $E = 0.15$) region of Figure 5.7, the simulations follow a process of flattening from the initial bimodal distribution since the low tolerance spreads the distribution and exposure is high enough that repulsive interactions take place with regularity. At μ_c in this parameter region the distribution has mass at either extreme of the opinion space and less peaked modes in their approximate initial positions. Following the simulations through to their final distribution results in a number of point masses at intervals along the $[0, 1]$ opinion space. For example, under ($T = 0.1$, $R = 0.5$, $E = 0.15$) six point masses, including at $x = 0$ and $x = 1$, are found in the opinion space at $t = 600$.

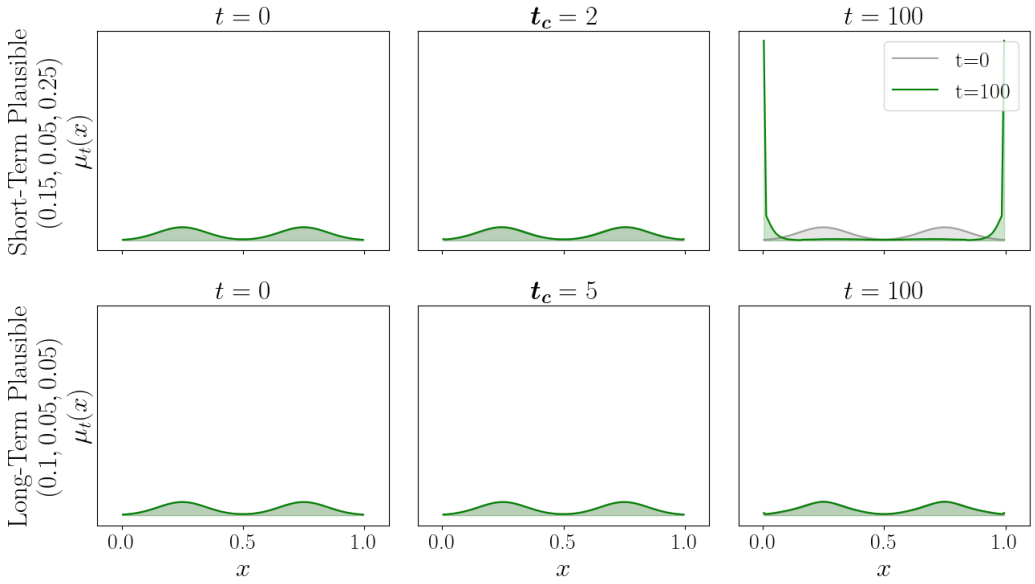


Figure 5.8: (T, R, E) simulations that achieve low $W(\mu_0, \mu_c)$ evaluated at $t = 0, t_c, 100$. Some simulations are short-term plausible meaning that they are initially quasi-stable but move away from the initial distribution over time. Other simulations are long-term plausible meaning that after the quasi-stable simulation they remain similar to the initial distribution, typically due to low exposure and responsiveness so opinion change is minimal.

In Figure 5.9, the time t at which μ_c is achieved is shown for the model parameter space. For most simulations μ_c is found for $t < 100$, while low exposure ($E = 0.05$) paired with either low responsiveness ($R = 0.05$) or low tolerance ($T = 0.05$) can result in slower convergence to the quasi-stable distribution. The exception is for the two plausible simulation regions where μ_c is arrived at quickly.

The majority of the model parameter space produces simulations that converge to quasi-stable distributions μ_c but they do not resemble the initial distribution μ_0. So they are not useful under the conditions set out here and the model can be determined as false for these parameters.

Results for alternative starting distributions, such as the Uniform distribution or a unimodal distribution generated from a single Gaussian distribution, can be seen in Figures A.1 and A.2 in Appendix A with varying regions of plausible simulation – highlighting that the model is falsified with respect

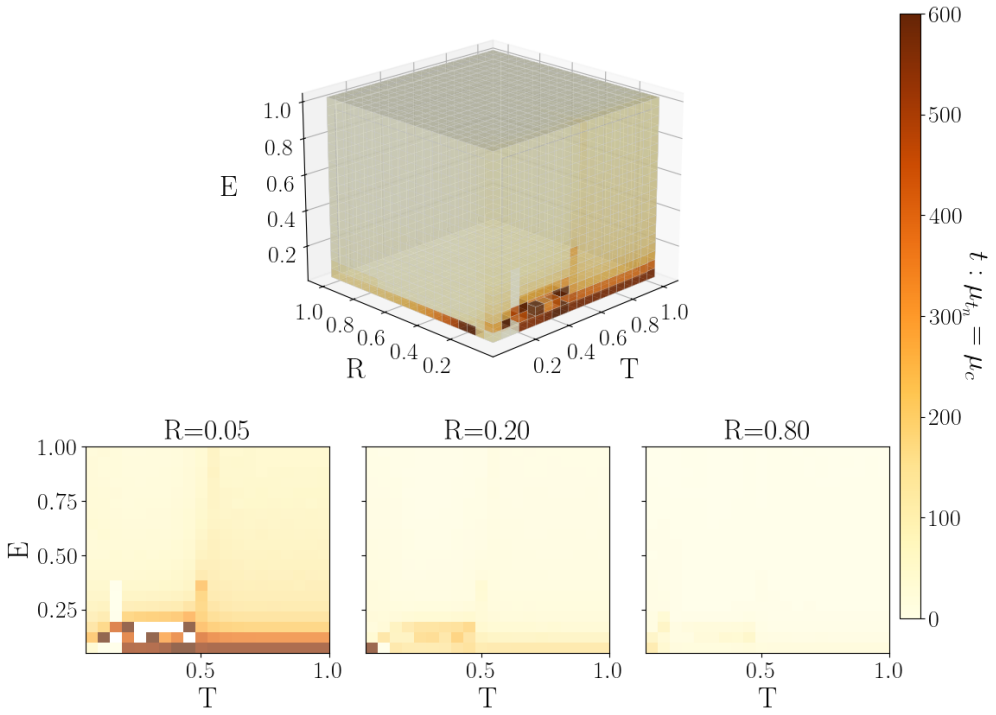


Figure 5.9: Critical time t_c at which μ_c is achieved within the (T, R, E) parameter space. Low exposure E paired with either low tolerance T or low responsiveness R tend to result in slower time of arrival to the quasi-stable distribution μ_c, apart from for the two low $W(\mu_0, \mu_c)$ regions identified in Figure 5.7.

to parameters dependent on the initial distribution.

5.2.3 Polarisation of the Parameter Space

Polarisation outcomes, as measured by the DER measure, evaluated when simulations are run to convergence up until $t = 600$ for the original model without group-dependent extension can be seen in Figure 5.10. The polarised/non-polarised regions of the space correspond to the expected be-

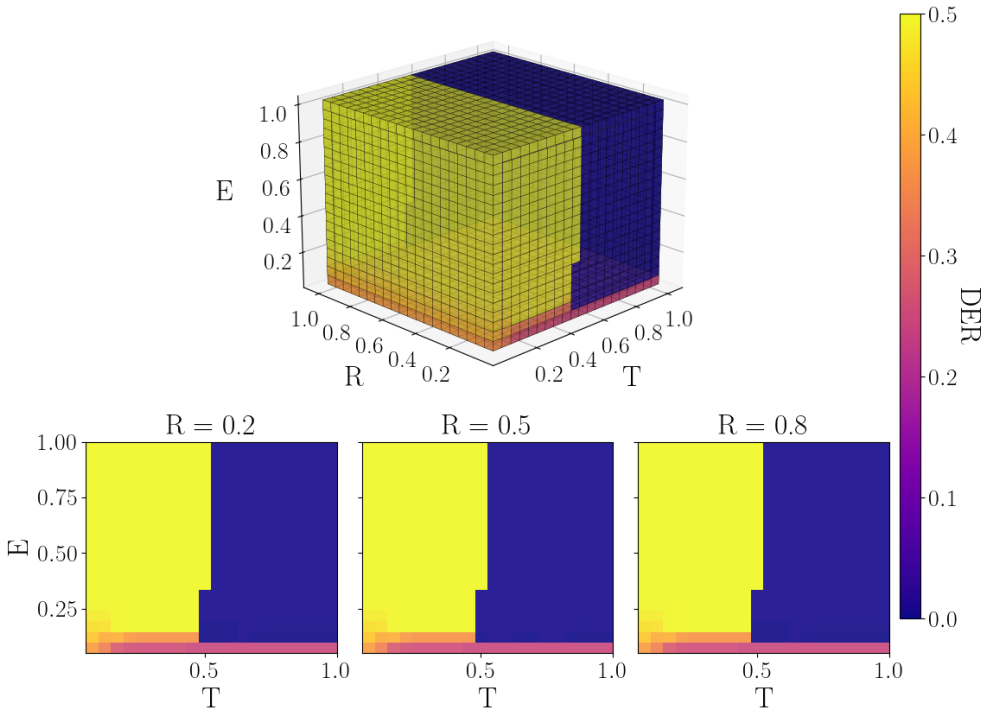


Figure 5.10: Polarisation outcomes of the three-dimensional parameter space of the Attraction-Repulsion Model without group-dependent extension, initialised with a bimodal distribution. Tolerance T and exposure E structure the space with boundaries between lowly and highly polarised regions, while responsiveness R influences rate of opinion distribution change but does not impact the final distribution as seen in identical two-dimensional slices across R values.

haviour of the model as found in Axelrod et al. (2021) and discussed, albeit with the additional model element of group identification, in Section 5.1.

There is a sharp boundary at which tolerance changes opinion distribution outcomes from consensus to polarisation at $T = 0.5$. Low exposure values can limit polarisation in the case where $E \leq 0.1$ by limiting the scope of repulsive interactions. While responsiveness is a rate of change parameter for the

distribution so has no impact on final state polarisation outcomes, as seen in the identical two-dimensional slices of the three-dimensional parameter space for different R. The simulation is now deterministic so the increased noise of high R in the non-deterministic agent-based model is not a factor in this case.

The exact boundary position between polarised and non-polarised regions is dependent on the initial opinion distribution, however relative structure of the regions is constant. For example, low T and high E will produce polarisation while high T and high E will produce consensus, but while low T in Figure 5.10 is $T < 0.5$, the threshold for what is considered "low T" is variable with the initial distribution.

Given that interaction is no longer a binary outcome between agents, the pixelation observed in Figures 5.2, 5.3, and 5.4, does not occur in Figure 5.10. More precisely, since the exposure rule is an interaction weight in the finite volume method model, rather than a probability as in the agent-based model, increasing R does not have the effect of making polarisation outcomes less predictable as discussed in Section 5.1.3.

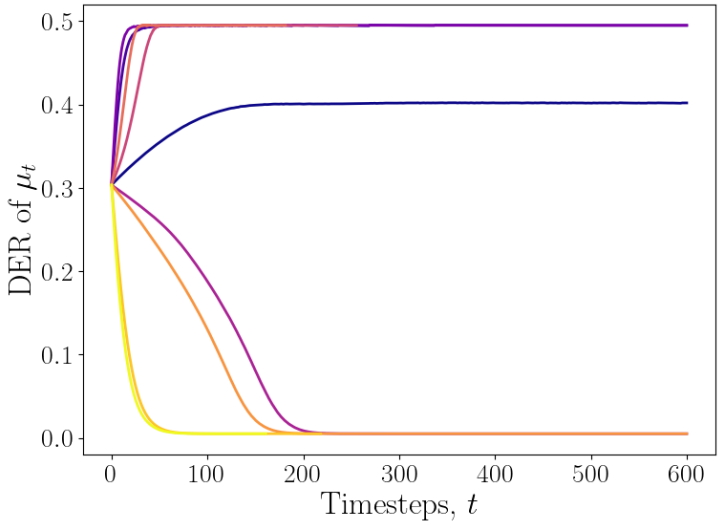


Figure 5.11: Convergence of DER values over simulations up to $t = 600$. Curves are for $R = 0.1$, since these are slow simulations, while $(T, E) \in \{0.1, 0.5, 0.9\} \times \{0.1, 0.5, 0.9\}$, the same as in Figure 5.5. The convergence curves for the agent-based model can be found in Figure 3.6.

Evolution of the DER measure over time can be observed in Figure 5.11.

The same set of parameter simulations is plotted as in Figure 5.5, again setting $R = 0.1$ for all shown simulations due to the slower convergence rate under lower responsiveness. By maximum simulation time ($t = 600$) the polarisation of the distribution is stable across parameter combinations.

Efficient simulation of opinion distributions rather than agents from the modelling techniques of Chapter 4 allow for exploration of the polarisation outcomes of the full three-dimensional model parameter space rather than the two-dimensional exploration in the previous work of Axelrod et al. (2021).

5.2.4 An Approach to Close the Empirical Gap

It has been discussed that the majority of opinion dynamics research focuses on the emergence of macro-properties – typically consensus or polarisation of the population – from the simulation of agent-based models. This is principally due to the difficulties of collecting consistent data over time and making an appropriate link between empirical observations and model mechanisms. The application of the theoretical techniques detailed in Chapter 4 opens a new approach for opinion dynamics whereby models can be falsified as well as used as a tool for the explanation of macro-level scenarios, as they are currently.

If the simulations produced by the model do not resemble the initial distribution at the point at which they are quasi-stable then this is taken as evidence that the model under these parameter conditions is false. The alternative is that the model is plausible if the quasi-stable distribution produced through simulation is similar to the initial distribution. It is possible to change the criteria by which the model is falsified if different evidence were to be required to deem simulations realistic. This approach provides a step towards closing the current empirical gap by providing a framework to describe realistic distributions rather than the extremes of consensus and polarisation.

The initial distribution presented here is the sum of two Gaussian distributions with separate means, thus it is a bimodal distribution. While it is a constructed distribution, it is an appropriate starting point to represent an already polarised distribution that may be found in reality. The same analysis can be undertaken with further distributions to potentially falsify the model for those distributions in the future.

For the bimodal example, regions of plausible simulations were identified to then be discussed. The framework can be adapted to real world settings

by changing μ_0 from a constructed bimodal distribution to an empirical distribution. This straightforward use of empirical data allows models to be falsified in an empirical setting and presents a potential fruitful avenue of future research in closing the current empirical gap in opinion dynamics. Therefore the framework detailed here provides a method by which models can be falsified or deemed plausible for empirical data.

Chapter 6

Conclusions

This thesis has demonstrated two complementary techniques that bring opinion dynamics closer to social reality: group identification and empirical falsification. Both strands of work stress the importance of placing opinion dynamics in real-world context, rather than remaining as theoretical models of opinion change that treat individuals as atoms (Jensen 2019). The purpose of this contextualisation is to produce more insightful models for the polarisation of opinions, which has been an increasing subject of interest in recent years (Finkel et al. 2020; Wagner 2021; Peralta et al. 2024).

The Attraction-Repulsion Model used in Axelrod et al. (2021) was taken as the opinion dynamics model of focus due to its incorporation of positive and negative social influence (Bail et al. 2018) and homophily in exposure to others (Barberá 2015). Two methodological approaches extended the model to place it in improved real-world context, with new functionality that answers to the call for the inclusion of groups and empirical validation.

The addition of group identification to determine opinion updates differently as a result of the perception of others as in-group or out-group facilitated a reflection on the importance of affective polarisation (Iyengar et al. 2012) and social identity (Tajfel 1974) for opinion modelling. While the application of a mean-field approximation and numerical simulation by the finite volume method facilitated the modelling of opinion distributions to reduce the empirical gap between models and validation with data (Carpentras 2023a).

Research objectives of the thesis were stated in Chapter 1. The first objective was to test for macro-properties of opinion dynamics models, such as similarity of simulated distributions to initial distributions, in order to

determine a framework that assesses models as plausible or false for a certain set of parameters and initial distribution. Macro-properties were chosen over micro-properties (such as model mechanism calibration) due to the difficulty of obtaining fine-grained temporal data of opinion change.

The second research objective was to present a model that allowed for research into social group dynamics. The purpose for doing so was to address a lack of modelling that asks how the treatment of others according to group identification can explain the polarisation of opinion; particularly since classic models rely on pairwise interactions with identical individuals (Starnini et al. 2025).

This manuscript began with an Introduction and Literature Review in Chapters 1 and 2 that (a) presented current shortcomings of opinion dynamics, (b) developed an understanding of opinion polarisation, and (c) identified research objectives to productively build on top of existing work. The underlying theory to address the research gaps was explained in Chapters 3 and 4; treating group dynamics and opinion distributions, respectively. Finally, the applications and results of the extended opinion dynamics models were explored in Chapter 5.

First results of the research conducted were focused on advancing modelling techniques. The Attraction-Repulsion Model can now approach questions pertaining to the group identification of individuals with group perception enabled by the application of HDBScan and the use of group-dependent parameters in the opinion update rule and the exposure rule. The second contribution to modelling theory is the presentation of a framework by which the Attraction-Repulsion Model can treat opinion distributions. It is key to note that neither type of model extension is unique to the Attraction-Repulsion Model, therefore both modelling advances should be applicable to any similar agent-based opinion dynamics model.

Further results were applications of the theory developed. When considering the work on the group dynamics of polarisation, the treatment of an out-group was identified as a central determining factor in a population's eventual level of polarisation – which aligns with understanding in social science (Yarchi et al. 2024). On the other hand, in-group dynamics were shown to potentially fragment a group from within if the group's spread of opinions was too large.

Results relating to the finite volume method model identified regions of plausible simulations in the model's parameter space for an initial bimodal distribution. The judgement of plausibility relies on criteria defined by quasi-

stable simulations and resemblance to the initial distribution, as measured by the Wasserstein distance. The impact of this is a step forward with research into whether models can reproduce empirical distributions.

The broad implications of these findings are two-fold. Reducing the empirical gap has been a principal concern of the opinion dynamics community in recent years (Flache et al. 2017). Therefore the results relating to the modelling of opinion distributions enable the opinion dynamics community to move towards empirical validation of the many existing models and their variations.

While the other broad implication is to place opinion dynamics in greater context and contact with social theory. It was noted that group identification in modelling of opinion change is particularly relevant to affective polarisation (Druckman et al. 2021). Accordingly, the research approach is an interdisciplinary effort to bridge the gap between social sciences and the mathematical modelling of opinion dynamics. One that integrates a salient feature of social influence (group perception and identification), and thereby highlights a key aspect of polarisation dynamics.

There are some limitations to the work presented. The criteria for plausible simulations are experimental and an early attempt at such a falsification process. Realistic behaviour is complex, potentially situation-dependent, and therefore hard to define. Future work could look to define plausibility and quasi-stability by different conditions. The difficulty of assessing realistic behaviour also reflects on the current implementation of group identification; future improvements could be made to the identification process since the use of the clustering algorithm to define groups is not socially motivated and assumes population-wide agreement of group identification. Initial steps were made towards this end, with alternative group identification methods being presented alongside the eventual implementation, and could be developed further. Incorporating social environment sensing in models is a developing and relatively underexplored research avenue that can offer a valuable contribution to existing models of social systems (Galesic et al. 2021), suggesting fruitful future development in this area.

Furthermore, the identities established as part of the inexchangeable mean-limit provide a future opportunity to include analysis of group identification within the empirical validation framework enabled by the numerical simulation of the model. This next step would be a unifying piece of work between the research objectives related to group dynamics and closing the empirical gap with opinion distributions to be able to provide even more

detailed findings and real-world context.

It was mentioned that the modelling techniques are not unique to the Attraction-Repulsion Model, future research could employ different models from the opinion dynamics model inventory (Proskurnikov and Tempo 2017). The models would likely produce some variation of behaviours found here and perhaps provide further insight as a result. Finally, application of the techniques to datasets could provide further new insight. While representative distributions taken to reflect reality (such as unimodal, bimodal, and uniform distributions) are useful, an application with a dataset would close the empirical gap even further.

The central message of this thesis is that opinion dynamics modelling must continue to strive forward in its push for empirical validation while maintaining a connection to key parts of social theory in order to produce increasingly relevant results. This work has contributed to this effort – demonstrating the importance of out-group treatment for polarisation outcomes and establishing search criteria for plausible simulations – thus extending the opinion dynamics and polarisation literature to provide a foundation for future research that further bridges theory and empirical evidence.

Bibliography

A. I. Abramowitz and K. L. Saunders. Is polarization a myth? *The Journal of politics*, 70(2):542–555, 2008.

E. Audusse, F. Bouchut, M.-O. Bristeau, and J. Sainte-Marie. Kinetic entropy inequality and hydrostatic reconstruction scheme for the saint-venant system. *Mathematics of Computation*, 85(302):2815–2837, 2016.

R. Axelrod, J. J. Daymude, and S. Forrest. Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*, 118(50):e2102139118, 2021.

N. Ayi and N. P. Duteil. Mean-field and graph limits for collective dynamics models with time-varying weights. *Journal of Differential Equations*, 299: 65–110, 2021.

N. Ayi and N. P. Duteil. Large-population limits of non-exchangeable particle systems. *Active Particles, Volume 4*, pages 79–133, 2024.

Á. Backhausz and B. Szegedy. Action convergence of operators and graphs. *Canadian Journal of Mathematics*, 74(1):72–121, 2022.

E. C. Baek, R. Hyon, K. López, M. A. Porter, and C. Parkinson. Perceived community alignment increases information sharing. *Nature Communications*, 16(1):5864, 2025.

C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

J. B. Bak-Coleman, M. Alfano, W. Barfuss, C. T. Bergstrom, M. A. Centeno, I. D. Couzin, J. F. Donges, M. Galesic, A. S. Gersick, J. Jacquet, et al. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27):e2025764118, 2021.

E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

D. Baldassarri and P. Bearman. Dynamics of political polarization. *American sociological review*, 72(5):784–811, 2007.

S. Balietti, L. Getoor, D. G. Goldstein, and D. J. Watts. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, 118(52): e2112552118, 2021.

S. Banisch and E. Olbrich. An argument communication model of polarization and ideological alignment. *Journal of Artificial Societies and Social Simulation*, 24(1):1–1, 2021.

S. Banisch and H. Shamon. Validating argument-based opinion dynamics with survey experiments. *arXiv preprint arXiv:2212.10143*, 2022.

S. Banisch and H. Shamon. Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate. *Sociological Methods & Research*, 54(1):187–236, 2025.

P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015.

F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics reports*, 874:1–92, 2020.

P. C. Bauer. Conceptualizing and measuring polarization: A review. *SocArXiv*, 2019.

G. V. Bodenhausen, S. K. Kang, and D. Peery. Social categorization and the perception of social groups. *The Sage handbook of social cognition*, pages 311–329, 2012.

C. Borgs, J. Chayes, H. Cohn, and Y. Zhao. An l^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5): 3019–3062, 2019.

F. Bouchut. *Nonlinear stability of finite Volume Methods for hyperbolic conservation laws: And Well-Balanced schemes for sources*. Springer Science & Business Media, 2004.

A. Bramson, P. Grim, D. J. Singer, W. J. Berger, G. Sack, S. Fisher, C. Flocken, and B. Holman. Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science*, 84(1):115–159, 2017.

W. Braun and K. Hepp. The vlasov dynamics and its fluctuations in the 1/n limit of interacting classical particles. *Communications in mathematical physics*, 56(2):101–113, 1977.

M. B. Brewer. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444, 1999.

K.-L. Brousmiche, J.-D. Kant, N. Sabouret, and F. Prenot-Guinard. From beliefs to attitudes: Polias, a model of attitude dynamics based on cognitive modeling and field data. *Journal of Artificial Societies and Social Simulation*, 19(4):2, 2016.

E. Bruch and J. Atwell. Agent-based models in empirical social research. *Sociological methods & research*, 44(2):186–221, 2015.

D. Carpentras. Why we are failing at connecting opinion dynamics to the empirical world. *Review of Artificial Societies and Social Simulation*, Mar. 2023a. URL `https://rofasss.org/2023/03/08/od-emprics`. Accessed 2025-09-24.

D. Carpentras. Why we are failing at connecting opinion dynamics to the empirical world. *Review of Artificial Societies and Social Simulations*, 2023b.

D. Carpentras and M. Quayle. Propagation of measurement error in opinion dynamics models: The case of the deffuant model. *Physica A: Statistical Mechanics and its Applications*, 606:127993, 2022.

D. Carpentras, P. J. Maher, C. O'Reilly, and M. Quayle. Deriving an opinion dynamics model from experimental data. *Journal of Artificial Societies and Social Simulation*, 25(4), 2022.

D. Carpentras, A. Lueders, P. J. Maher, C. O'Reilly, and M. Quayle. How polarization extends to new topics: An agent-based model derived from experimental data. *Journal of Artificial Societies and Social Simulation*, 26(3), 2023.

D. Cassells, L. Costantini, A. F. Ashery, S. Gadge, D. L. Pires, M. Á. Sánchez-Cortés, A. Santoro, and E. Omodei. A 72h exploration of the co-evolution of food insecurity and international migration. *arXiv preprint arXiv:2407.03117*, 2024a.

D. Cassells, L. Tabourier, and P. Ramaciotti. Modeling both pairwise interactions and group effects in polarization on interaction networks. In *International Conference on Complex Networks*, pages 43–54. Springer, 2024b.

D. Cassells, A. Vendeville, L. Tabourier, and P. Ramaciotti. Co-evolution of groups and opinions in agent-based models. *preprint under review*, 2025.

C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591–646, 2009.

T. E. Charlesworth and M. R. Banaji. Patterns of implicit and explicit attitudes: Iv. change and stability from 2007 to 2020. *Psychological Science*, 33(9):1347–1371, 2022.

X. Chen, P. Tsaparas, J. Lijffijt, and T. De Bie. Opinion dynamics with backfire effect and biased assimilation. *PloS one*, 16(9):e0256922, 2021.

W. Chu, Q. Li, and M. A. Porter. Inference of interaction kernels in mean-field models of opinion dynamics. *SIAM Journal on Applied Mathematics*, 84(3):1096–1115, 2024.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.

G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

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

D. DellaPosta, Y. Shi, and M. Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

P. DiMaggio, J. Evans, and B. Bryson. Have american's social attitudes become more polarized? *American journal of Sociology*, 102(3):690–755, 1996.

R. L. Dobrushin. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979.

D. J. Downey and M. L. Huffman. Attitudinal polarization and trimodal distributions: measurement problems and theoretical implications. *Social science quarterly*, 82(3):494–505, 2001.

J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan. Affective polarization, local contexts and public opinion in America. *Nature human behaviour*, 5(1):28–38, 2021.

A. Dubovskaya, S. C. Fennell, K. Burke, J. P. Gleeson, and D. O'Kiely. Analysis of mean-field approximation for deffuant opinion dynamics on networks. *SIAM Journal on Applied Mathematics*, 83(2):436–459, 2023.

J.-Y. Duclos, J. Esteban, and D. Ray. Polarization: concepts, measurement, estimation. In *The Social Economics of Poverty*, pages 54–102. Routledge, 2006.

S. Eschert and B. Simon. Respect and political disagreement: Can intergroup respect reduce the biased evaluation of outgroup arguments? *PloS one*, 14(3):e0211556, 2019.

R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi, et al. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121, 2022.

S. C. Fennell, K. Burke, M. Quayle, and J. P. Gleeson. Generalized mean-field approximation for the deffuant opinion dynamics model on networks. *Physical Review E*, 103(1):012314, 2021.

J. M. Fields and H. Schuman. Public beliefs about the beliefs of the public. *Public Opinion Quarterly*, 40(4):427–448, 1976.

E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, et al. Political sectarianism in America. *Science*, 370(6516):533–536, 2020.

M. P. Fiorina and S. J. Abrams. Political polarization in the American public. *Annu. Rev. Polit. Sci.*, 11(1):563–588, 2008.

M. P. Fiorina, S. A. Abrams, and J. C. Pope. Polarization in the american public: Misconceptions and misreadings. *The Journal of Politics*, 70(2): 556–560, 2008.

A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz. Models of social influence: Towards the next frontiers. *Jasss-The journal of artificial societies and social simulation*, 20(4):2, 2017.

S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5): 75–174, 2010.

J. B. Freeman and R. Dale. Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, 45(1):83–97, 2013.

N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990.

S. Galam. Sociophysics: A review of galam models. *International Journal of Modern Physics C*, 19(03):409–440, 2008.

M. Galesic, W. Bruine de Bruin, J. Dalege, S. L. Feld, F. Kreuter, H. Olsson, D. Prelec, D. L. Stein, and T. van Der Does. Human social sensing is an

untapped resource for computational social science. *Nature*, 595(7866): 214–222, 2021.

K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

D. Garzia, F. Ferreira da Silva, and S. Maye. Affective polarization in comparative and longitudinal perspective. *Public Opinion Quarterly*, 87(1): 219–231, 2023.

J. M. Gerson and K. Peiss. Boundaries, negotiation, consciousness: Reconceptualizing gender relations. *Social problems*, 32(4):317–331, 1985.

D. Geschke, J. Lorenz, and P. Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58 (1):129–149, 2019.

M. Gestefeld and J. Lorenz. Calibrating an opinion dynamics model to empirical opinion distributions and transitions. *Journal of Artificial Societies and Social Simulation*, 26(4), 2023.

M. Gestefeld, J. Lorenz, N. T. Henschel, and K. Boehnke. Decomposing attitude distributions to characterize attitude polarization in Europe. *SN Social Sciences*, 2(7):110, 2022.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

A. Gollwitzer, C. Martel, W. J. Brady, P. Pärnamets, I. G. Freedman, E. D. Knowles, and J. J. Van Bavel. Partisan differences in physical distancing are linked to health outcomes during the covid-19 pandemic. *Nature human behaviour*, 4(11):1186–1197, 2020.

S. González-Bailón, D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. M. Guess, et al. Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398, 2023.

S. Gurukar, D. Ajwani, S. Dutta, J. Lauri, S. Parthasarathy, and A. Sala. Towards quantifying the distance between opinions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 229–239, 2020.

R. Hegselmann. Cellular automata in the social sciences: Perspectives, restrictions, and artefacts. In *Modelling and simulation in the social sciences from the philosophy of science point of view*, pages 209–233. Springer, 1996.

R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):1–2, 2002.

M. J. Hetherington. Putting polarization in perspective. *British Journal of Political Science*, 39(2):413–448, 2009.

J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595 (7866):181–188, 2021.

M. Hohmann, K. Devriendt, and M. Coscia. Quantifying ideological polarization on a network using generalized euclidean distance. *Science Advances*, 9(9):eabq2044, 2023.

M. J. Hornsey. Social identity theory and self-categorization theory: A historical review. *Social and personality psychology compass*, 2(1):204–222, 2008.

I. Iacopini, G. Petri, A. Baronchelli, and A. Barrat. Group interactions modulate critical mass dynamics in social convention. *Communications Physics*, 5(1):64, 2022.

S. Iyengar, G. Sood, and Y. Lelkes. Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431, 2012.

S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146, 2019.

P.-E. Jabin, D. Poyato, and J. Soler. Mean-field limit of non-exchangeable systems. *Communications on Pure and Applied Mathematics*, 78(4): 651–741, 2025.

W. Jager and F. Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10(4):295–303, 2005.

P. Jensen. The politics of physicists' social models. *Comptes Rendus. Physique*, 20(4):380–386, 2019.

J. T. Jost, D. S. Baldassarri, and J. N. Druckman. Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature reviews psychology*, 1(10):560–576, 2022.

M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, et al. Social physics. *Physics Reports*, 948:1–148, 2022.

G. Karypis and V. Kumar. Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0. *University of Minnesota*, 1995.

I. V. Kozitsin. A general framework to link theory and empirics in opinion formation models. *Scientific reports*, 12(1):5543, 2022.

D. Kreiss and S. C. McGregor. A review and provocation: On polarization and platforms. *New Media & Society*, 26(1):556–579, 2024.

E. Kubin and C. Von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

M. Lamont and V. Molnár. The study of boundaries in the social sciences. *Annual review of sociology*, 28(1):167–195, 2002.

H. Lee and M. E. Sobel. The wasserstein bipolarization index: A new measure of public opinion polarization, with an application to cross-country attitudes toward covid-19 vaccination mandates. *arXiv preprint arXiv:2408.03331*, 2024.

S. A. Levin, H. V. Milner, and C. Perrings. The dynamics of political polarization, 2021.

C. C. Liu and S. B. Srivastava. Pulling closer and moving apart: Interaction, identity, and influence in the us senate, 1973 to 2009. *American Sociological Review*, 80(1):192–217, 2015.

C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

L. Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

G. Marks, D. Attewell, J. Rovny, and L. Hooghe. Cleavage theory. In *The Palgrave handbook of EU crises*, pages 173–193. Springer, 2020.

M. Mäs, A. Flache, K. Takács, and K. A. Jehn. In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization science*, 24(3):716–736, 2013.

L. Mason. "i disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American journal of political science*, 59 (1):128–145, 2015.

L. Mason. Ideologues without issues: The polarizing consequences of ideological identities. *Public Opinion Quarterly*, 82(S1):866–887, 2018.

T. Masson, P. Jugert, and I. Fritsche. Collective self-fulfilling prophecies: Group identification biases perceptions of environmental group norms among high identifiers. *Social Influence*, 11(3):185–198, 2016.

N. McCarty, K. T. Poole, and H. Rosenthal. *Polarized America: The dance of ideology and unequal riches*. mit Press, 2016.

L. McInnes and J. Healy. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 33–42. IEEE, 2017.

L. McInnes, J. Healy, S. Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

E. o. Merriam-Webster. 2024 Word of the Year: Polarization. `https://www.merriam-webster.com/wordplay/word-of-the-year`, 2024. [Accessed 24-09-2025].

M. Moussaïd, J. E. Kämmer, P. P. Analytis, and H. Neth. Social influence and the collective dynamics of opinion formation. *PloS one*, 8(11):e78433, 2013.

C. Musco, I. Ramesh, J. Ugander, and R. T. Witter. How to quantify polarization in models of opinion dynamics. *arXiv preprint arXiv:2110.11981*, 2021.

D. C. Mutz. Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, 96(1):111–126, 2002.

M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

H. Noorazar. Recent advances in opinion propagation dynamics: A 2020 survey. *The European Physical Journal Plus*, 135(6):521, 2020.

D. Novelli, J. Drury, and S. Reicher. Come together: Two studies concerning the impact of group relations on personal space. *British Journal of Social Psychology*, 49(2):223–236, 2010.

H. Olsson and M. Galesic. Analogies for modeling belief dynamics. *Trends in Cognitive Sciences*, 28(10):907–923, 2024.

D. O'Callaghan, D. Greene, M. Conway, J. Carthy, and P. Cunningham. Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, 33(4):459–478, 2015.

L. Peel, T. P. Peixoto, and M. De Domenico. Statistical inference links data and theory in network science. *Nature Communications*, 13(1):6794, 2022.

A. F. Peralta, J. Kertész, and G. Iñiguez. Opinion dynamics in social networks: From models to data. *arXiv preprint arXiv:2201.01322*, 2022.

A. F. Peralta, P. Ramaciotti, J. Kertész, and G. Iñiguez. Multidimensional political polarization in online social networks. *Physical Review Research*, 6(1):013170, 2024.

T. F. Pettigrew and L. R. Tropp. A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, 90(5):751, 2006.

K. T. Poole and H. Rosenthal. Patterns of congressional voting. *American journal of political science*, pages 228–278, 1991.

R. Prisant, F. Garin, and P. Frasca. Opinion dynamics on signed graphs and graphons: Beyond the piece-wise constant case. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 5430–5435. IEEE, 2024.

A. V. Proskurnikov and R. Tempo. A tutorial on modeling and analysis of dynamic social networks. part i. *Annual Reviews in Control*, 43:65–79, 2017.

A. V. Proskurnikov and R. Tempo. A tutorial on modeling and analysis of dynamic social networks. part ii. *Annual Reviews in Control*, 45:166–190, 2018.

H. Pérez-Martínez, S. Lamata-Otín, F. Malizia, L. M. Floría, J. Gómez-Gardeñes, and D. Soriano-Paños. Social polarization promoted by sparse higher-order interactions. *arXiv preprint arXiv:2507.12325*, 2025.

U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036106, 2007.

P. Ramaciotti, J.-P. Cointet, G. Muñoz Zolotoochin, A. Fernández Peralta, G. Iñiguez, and A. Pournaki. Inferring attitudinal spaces in social networks. *Social Network Analysis and Mining*, 13(1):14, 2022.

P. Ramaciotti, D. Cassells, Z. Vagena, J.-P. Cointet, and M. Bailey. American politics in 3d: Measuring multidimensional issue alignment in social media using social graphs and text data. *Applied Network Science*, 9(1):2, 2024.

S. Redner. Reality-inspired voter models: A mini-review. *Comptes Rendus. Physique*, 20(4):275–292, 2019.

S. Reicher, S. A. Haslam, and R. Rath. Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, 2(3):1313–1344, 2008.

A. Reiljan, D. Garzia, F. F. Da Silva, and A. H. Trechsel. Patterns of affective polarization toward parties and leaders across the democratic world. *American Political Science Review*, 118(2):654–670, 2024.

R. Rekker. The nature and origins of political polarization over science. *Public Understanding of Science*, 30(4):352–368, 2021.

F. Richter, C. Thiébaut, and L. Safra. Not just by means alone: why the evolution of distribution shapes matters for understanding opinion dynamics. the case of the french reaction to the war in ukraine. *Frontiers in Political Science*, 6:1327662, 2024.

J. Robison and R. L. Moskowitz. The group basis of partisan affective polarization. *The Journal of Politics*, 81(3):1075–1079, 2019.

S. Roccas and M. B. Brewer. Social identity complexity. *Personality and social psychology review*, 6(2):88–106, 2002.

M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues. Clustering algorithms: A comparative approach. *PloS one*, 14(1):e0210236, 2019.

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

C. R. M. A. Santagiustina, J.-P. Cointet, and P. R. Morales. Expressing one's identity online: Left-right and cross eu-country variation in self-representation in social media. *arXiv preprint arXiv:2501.05927*, 2025.

M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte. The many facets of community detection in complex networks. *Applied network science*, 2(1):4, 2017.

J. Sieber and R. Ziegler. Group polarization revisited: A processing effort account. *Personality and Social Psychology Bulletin*, 45(10):1482–1498, 2019.

T. W. Smith. Is there real opinion change? *International Journal of Public Opinion Research*, 6(2):187–203, 1994.

P. Sobkowicz. Whither now, opinion modelers? *Frontiers in Physics*, 8: 587009, 2020.

M. Starnini, F. Baumann, T. Galla, D. Garcia, G. Iñiguez, M. Karsai, J. Lorenz, and K. Sznajd-Weron. Opinion dynamics: Statistical physics and beyond, 2025. URL `https://arxiv.org/abs/2507.11521`.

P. Steiglechner, P. E. Smaldino, and A. Merico. How opinion variation among in-groups can skew perceptions of ideological polarization. *PNAS nexus*, 4 (7):pgaf184, 2025.

T. Szczepanska, P. Antosz, J. O. Berndt, M. Borit, E. Chattoe-Brown, S. Mehryar, R. Meyer, S. Onggo, and H. Verhagen. Gam on! six ways to explore social complexity by combining games and agent-based models. *International Journal of Social Research Methodology*, 25(4):541–555, 2022.

K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

C. S. Taber and M. Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

H. Tajfel. Social identity and intergroup behaviour. *Social science information*, 13(2):65–93, 1974.

K. Takács, A. Flache, and M. Mäs. Discrepancy and disliking do not induce negative opinion shifts. *PloS one*, 11(6):e0157948, 2016.

R. Toral and C. J. Tessone. Finite size effects in the dynamics of opinion formation. *arXiv preprint physics/0607252*, 2006.

P. Törnberg, C. Andersson, K. Lindgren, and S. Banisch. Modeling the emergence of affective polarization in the social media society. *PloS one*, 16(10):e0258259, 2021.

F. M. Turner-Zwinkels, J. van Noord, R. Kesberg, E. García-Sánchez, M. J. Brandt, T. Kuppens, M. J. Easterbrook, L. Smets, P. Gorska, M. Marchlewska, et al. Affective polarization and political belief systems: The role

of political identity and the content and structure of political beliefs. *Personality and Social Psychology Bulletin*, 51(2):222–238, 2025.

A. Vendeville. Voter model can accurately predict individual opinions in online populations. *Physical Review E*, 111(6):064310, 2025.

A. Vendeville, J. Royo-Letelier, D. Cassells, J.-P. Cointet, M. Crépel, T. Faverjon, T. Lenoir, B. Mazoyer, B. Ooghe-Tabanou, A. Pournaki, et al. Mapping the political landscape from data traces: multidimensional opinions of users, politicians and media outlets on x. *preprint under review*, 2025.

J. G. Voelkel, M. N. Stagnaro, J. Y. Chu, S. L. Pink, J. S. Mernyk, C. Redekopp, I. Ghezae, M. Cashman, D. Adjodah, L. G. Allen, et al. Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, 386(6719):eadh4764, 2024.

M. Wagner. Affective polarization in multiparty systems. *Electoral studies*, 69:102199, 2021.

M. M. Waldrop. Modeling the power of polarization. *Proceedings of the National Academy of Sciences*, 118(37):e2114484118, 2021.

I. Waller and A. Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268, 2021.

P. H. Westfall. Kurtosis as peakedness, 1905–2014. rip. *The American Statistician*, 68(3):191–195, 2014.

P. Windrum, G. Fagiolo, and A. Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007.

S. Wolfram. Cellular automata as models of complexity. *Nature*, 311(5985): 419–424, 1984.

V. C. Yang, T. van der Does, and H. Olsson. Falling through the cracks: Modeling the formation of social category boundaries. *PloS one*, 16(3): e0247562, 2021.

M. Yarchi, C. Baden, and N. Kligler-Vilenchik. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. In *Dissonant Public Spheres*, pages 185–226. Routledge, 2024.

Appendix A

Appendix

The $W(\mu_0, \mu_c)$ and final-state DER for the finite volume method model, as appearing in Figures 5.7 and 5.10, are presented here under a different initial distribution μ_0. In Figure A.1, μ_0 is a single Gaussian distribution centred at 0.5 with a standard deviation of 0.05 with support on the interval $[0, 1]$. While in Figure A.2, μ_0 is the Uniform distribution on the interval $[0, 1]$.

The parameter regions for which simulation is plausible ($W(\mu_0, \mu_c) < 0.1$) are different for the three different initial distributions, which highlights that the model is plausible under the parameter combinations *and* the initial distribution. The final-state polarisation measured by DER is also dependent on model parameters and the initial distribution – for example, consensus occurs for $T \geq 0.2$ in the unimodal case of Figure A.1 while consensus occurs for $T > 0.5$ in the bimodal case of Figure 5.10.

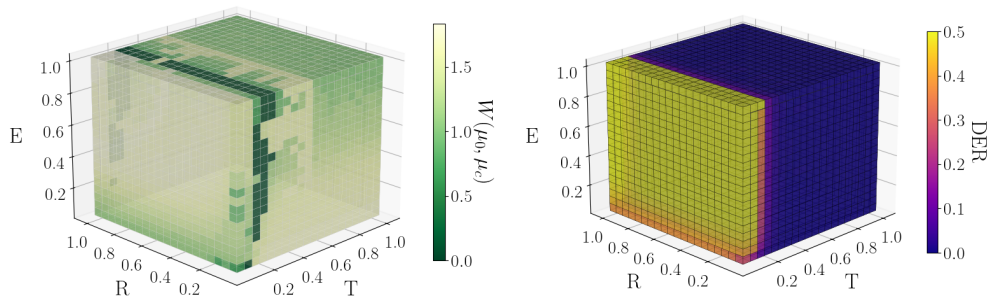


Figure A.1: The same methodology as in Figures 5.7 and 5.10, instead with μ_0 as a single Gaussian distribution centred at 0.5 with standard deviation 0.05 and support on the interval $[0, 1]$.

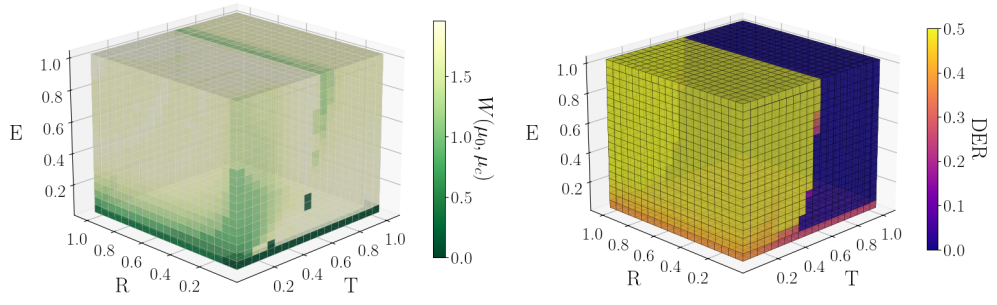


Figure A.2: The same methodology as in Figures 5.7 and 5.10, instead with μ_0 as a Uniform distribution with support on the interval $[0, 1]$.