Polarisation and Disagreement in Opinion Dynamics based on Confirmation Biases

Martin Prieto Alejandro

4th Year Project Report
Computer Science and Mathematics
School of Informatics
University of Edinburgh

2023

Abstract

Our beliefs and opinions shape our actions, which in turn impact political results, economic development, and even public health in the case of a pandemic. As such, understanding how beliefs evolve in collectives is crucial, and it is the primary objective of the field of Opinion Dynamics. Amongst other things, an understanding of these processes can help inform decisions in political spheres or in the design of social media platforms with the aim to mitigate the emergence of undesired opinion distributions and misinformation. Confirmation bias is the tendency to seek out and assimilate information in a way that is partial to prior beliefs and is known to have a great impact both in individual and collective opinion formation. In opinion dynamics, many different modelling assumptions have been made to account for confirmation bias. We classify three of these assumptions into the specific behaviours they model (*biased seeking*, *biased evaluation*, *biased assimilation*, and *belief perseverance*) and propose a new model which incorporates the classified assumptions in order to account for different confirmationally biased behaviours. We first consider the behavioural implications of the model in simple settings, finding that it can exhibit clustering of opinions and consensus characteristic of bounded confidence models, but also diversity within clusters (strong diversity) and noisy consensus often observed in the real world but rarely captured in the literature. After investigating the interplay between social network structure and *biased seeking*, we study the effects of the different biased behaviours on the emergence of polarised communities, concluding that *biased evaluation* is at the core of this phenomenon. Furthermore, the introduction of *biased seeking* and *biased assimilation* generates polarized communities that maintain some level of disagreement within themselves, a phenomenon observed in the real world but absent in other polarizing models. Finally, we validate the model through its ability to generate all qualitative categories of opinion distributions in networks from four social media platforms.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Martin Prieto Alejandro*)

Acknowledgements

I would like to thank my supervisor Dr. Peggy Series for her support and advice through this project.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

The study of Opinion Dynamics consists in investigating how opinions evolve in a collective. It has been a topic of interest for researchers from numerous fields such as sociology, psychology, political science, computer science, or even physics. In recent years, this field has gained significant importance due to the rise of social media platforms which give people the ability to express their opinions to a large audience, facilitating the rapid dissemination of information and spread of opinions. Although not negative in itself, this phenomenon has been shown to facilitate the spread of misinformation [9] which can lead to the widespread adoption of harmful ideas and beliefs. Other recent trends such as the rise of extremisms, and the increase in polarisation [51] reinforce the need to understand the underlying phenomena driving opinion changes. In particular, it is useful to understand what behavioral patterns contribute the most to the emergence polarisation, and extremisms. Amongst other things, a good understanding of the conditions leading to such outcomes can inform design decisions in social media platforms in order to reduce these undesired phenomena.

Confirmation biases are often considered to be the main culprit for our inability to reach a consensus, our susceptibility to misinformation, and the emergence of polarisation in the political sphere. Confirmation biases are a range of behaviours which represent the tendency to seek out and assimilate information in a way that is partial to prior beliefs. Amongst these behaviours, we can find *biased seeking*: the tendency to seek information that confirms ones beliefs, *biased evaluation*: being more critical of information that contradicts ones beliefs, *biased assimilation*: being more influenced by confirmatory than conflicting evidence, or *belief perseverance*: reluctance to change beliefs in the face of dis-confirmatory evidence.

Given the undeniable role confirmation biases play in opinion dynamics, we are interested in how the different confirmation biases impact collective opinion formation. We can categorize opinion distributions that we observe in the real world into four categories: *consensus*, *dissensus*, *clustering*, and *polarisation*. Given these possible outcomes, we are interested in knowing: How does the presence of the mentioned biased behaviours impact the outcome of collective opinion formation? How does the structure of social interactions affect disagreement in the presence of these biased behaviours? Which biased behaviours contribute the most to the emergence of polarisation? Are these biased behaviours sufficient to explain the range of opinion distributions observed in the real world?

1.2 Project Objectives

The main objectives of this project are:

1. Developing a model of opinion dynamics which explicitly accounts for *biased seeking*, *biased evaluation*, *biased assimilation*, and *belief perseverance*
2. Determining the effects of the introduced biases on the collective evolution of opinions, and studying the patterns that emerge from the presence of several biased behaviours.
3. Investigating the effect of social structure on the emerging dynamics in the presence of confirmation biases.
4. Determining the role of each biased behaviour in the emergence of polarisation.
5. Validating the ability of the model to reproduce opinion distributions observed in real life, through social structures taken from four different social media datasets.

Instead of proposing an ad-hoc model, we build on previous literature by taking existing assumptions for modeling confirmation bias [68, 57, 22, 28], categorizing them, combining them, and extending them to more complex settings. We study the emergent behaviours by means of computer simulation, which entails the complete implementation of the proposed model, and of the different experiments we wish to run.

1.3 Outline of Thesis

In Chapter 2, we will provide a short overview of the field of opinion dynamics, distinguishing between three main types of models: Non-Bayesian, Bayesian and Bayesian-inspired, and presenting common approaches for modelling the structure of social connections. In Section 2.2, we look at previous attempts at modeling confirmation bias in opinion dynamics, and we present the three modelling assumptions which we will draw upon to model the different components of confirmation bias. In Chapter 3, we introduce the proposed model, along with simplifying assumptions necessary to make the model computationally tractable in more complex settings. We also introduce the metrics for *Diversity* and *Disagreement* we will use to describe the opinion space throughout the thesis. Similar versions of these metrics have been used in the literature to capture properties of the opinion space [69, 22]. In Chapter 4, we study the results of running simulations with the proposed model in different experimental settings, in order to better understand the effects the biased behaviours in collective opinion formation. Finally, in Chapter 5, we discuss how this work relates to current research, consider the limitations of the model, and make suggestions for further study.

Chapter 2

Background

2.1 Opinion Dynamics

2.1.1 General Overview and Non-Bayesian models

The field of opinion dynamics focuses on modeling how opinions evolve in a collective over time. Typically studying what conditions lead to one of four stable states: *consensus*, *dissensus*, *clustering of opinions*, and *polarisation*. Agent Based Models (ABM) naturally lend themselves to model this kind of phenomena, allowing modelers to define opinion update rules at the agent level and study the dynamics that emerge from the interactions of such agents. Although efforts have been made to model opinion dynamics in continuous time [16, 20], the complexity of these models, and the discrete nature of the available data, has lead to a majority of discrete time-models. Such models are often defined by specifying the opinion x_i of an agent i at time $t+1$ as a function of its opinion and neighbouring agents opinions at time t. For instance, a simple but greatly influential model (the DeGroot model)[27] defines opinion updating as

$$x_i(t+1) = w_{i1}x_1(t) + w_{i2}x_2(t) + \ldots + w_{in}x_n(t) \tag{2.1}$$

where w_{ij} is the weight agent i assigns to agent j's opinion. This model fits into one of the two main categories of models: continuous opinion models, where the opinion of agent i is given by a real number (usually within an interval). Many of the models in this category are extensions of the DeGroot model. Some notable examples are the Friedkin and Johsen model [35], the Altafini model [10, 11] the Hegselmann-Krause model [53, 65] or the Deffuant-Weisbuch model [26, 80] amongst many others. The second category of models is discrete opinion models, where agents take discrete (usually binary) opinions. Amongst these are the voter model [23, 43], the Sznajd model [71] and the majority rule model [36]. Although we have mentioned several seminal models, countless other models have been proposed and studied. In an attempt to track the progress in the field, surveys and reviews classify and summarise the findings from Opinion Dynamics models [63, 6, 38, 30, 12, 19]. The model proposed in this thesis is a model in discrete time with continuous opinions. Nevertheless, instead of representing opinions through a single number, we use probabilistic (Bayesian-inspired) opinions which we will explore in the next section.

2.1.2 Bayesian and Bayesian-Inspired Models

The models mentioned in the previous section are Non-Strategic or Non-Bayesian. That is, they don't describe the optimal way in which an agent should update its belief (opinion) in order to learn some underlying truth. Rather, they simply attempt to describe how humans interact by specifying simple mathematical rules that relate to how people behave, or that generate desirable properties through the process of opinion formation [6]. We now look at Bayesian and Bayesian-inspired models as they will be the focus of this work.

2.1.2.1 Bayes' rule

At the center of Bayesian models is Bayes' rule, a standard result from probability theory which describes the computation of conditional probabilities. Given two random events h and s, Bayes' rule describes the relationship between the independent probabilities of h and s ($P(h), P(s)$), the probability of h given s ($P(h|s)$), and the probability of s given h ($P(h|s)$) through the equation:

$$P(h|s) = \frac{P(h|s)P(h)}{P(s)} \tag{2.2}$$

where $P(s) \neq 0$. This result can also be extended to continuous random variables: given two continuous random variables θ and S, with probability density functions (pdf) f_θ and f_S, Bayes' rule states the following:

$$\begin{align} f_{\theta|S=s}(x) &= \frac{f_{S|\theta=x}(s) f_\theta(x)}{f_S(s)} \tag{2.3} \\ &\propto f_{S|\theta=x}(s) f_\theta(x) \tag{2.4} \end{align}$$

where $f_{\theta|S=s}(x)$ and $f_{S|\theta=x}(s)$ are conditional densities, and since the marginal density $f_S(s)$ is not dependent on x (its essentially a normalizing constant), $f_{\theta|S=s}(x)$ can be derived by normalizing $f_{S|\theta=x}(s) f_\theta(x)$ (Eq. 2.4).

2.1.2.2 Bayesian Models of Cognition

Bayesian models of cognition can provide a formal account of how rational agents should reason in situations of uncertainty. These models have become increasingly popular in many areas of cognitive science as they are able to represent several elements of human cognition [46]. By attributing meaning to the variables involved in Bayes rule, they provide a way to calculate how agents should update their beliefs in light of new evidence. Take the *prior probability* $P(h)$ to be the degree of belief of an agent in a hypothesis h and the *posterior probability* $P(h|s)$ is the degree of belief in h conditioned on the observation of s. Using Bayes rule (2.2) and marginalization, one can compute the *posterior* from the *prior* using the equation

$$P(h|s) = \frac{P(s|h)P(h)}{\sum_{h' \in \mathcal{H}} P(s|h')P(h')} \tag{2.5}$$

where $\mathcal{H}$ is the *hypothesis space* (set of hypothesis considered by the agent) and $P(s|h)$ is the *likelihood* (probability of the data given the hypothesis)[54]. This process of updating the probability of a hypothesis based on new data is called Bayesian inference.

As we will be dealing with continuous opinions, we also present an example of Bayesian inference with a continuous hypothesis space. Suppose an agent is trying to estimate some continuous parameter θ based on an observation s. In order to capture the agents uncertainty, θ is taken to be a random variable. Thus, the agent now has a *prior* belief which is the pdf f_θ of θ. For the agents belief about θ to be updated as a result of the observation s, s must be related to θ in some way. To keep things simple, we will consider s to be an observation of a normally distributed random variable $S \sim N(\theta, \sigma_S^2)$. This captures the idea that s is a noisy observation of the value of θ, where σ_S represents the noisiness or reliability of the observation. We can compute the *posterior* belief conditioned on the observation of s through Eq. 2.4, where the *likelihood* $f_{S|\theta=x}(s)$ is the pdf of S for fixed s and taken as a function of θ. We will simplify the notation of the likelihood $f_{S|\theta=x}(s)$ to $\mathcal{L}(\theta|s)$, the posterior $f_{\theta|S=s}(x)$ to $f(\theta|s)$, and the marginal density $f_\theta(x)$ to $f(\theta)$. With these simplifications, we can capture the updating of beliefs about continuous variables through the equation:

$$f(\theta|s) \propto \mathcal{L}(\theta|s) f(\theta) \tag{2.6}$$

Although the examples presented here are simple, the Bayesian framework can be expanded to more complex scenarios. Building on top of this framework, has allowed to establish theories of human inductive learning [73], visual scene perception [83], language processing and acquisition [21], and causal learning [70] amongst other topics.

2.1.2.3 Bayesian frameworks in opinion dynamics

With the growing popularity and success of Bayesian models in cognitive science, researchers have sought to apply them to opinion dynamics to build more realistic models of opinion updating at the agent level. Two categories of models that draw from the Bayesian framework can be distinguished.

The first category consists of models derived purely from Bayesian principles, which require a more mathematical analysis [33, 4, 6, 5]. In these models, agents observe some noisy signal about the world as in the previous section, but also interact with other agents to introduce the social component of opinion formation. However, as strictly Bayesian agents must model and update their beliefs not only about the variable of interest but also about the processes influencing other agents, these models can become complicated quickly. As a result, the study of such models is limited to very simple contexts such as sequential interactions.

Given the limitations of strict Bayesian models, and the inability to extend them to more complex and realistic settings, some modelers make simplifying assumptions or intentional changes to the framework. These decisions are motivated by the desire to make the models computationally tractable, the recognition that humans may not be able to meet the demands of Bayesian inference, and the observation that individuals behave irrationally in many situations. This leads us to the second category of models which we will denote as Bayesian-inspired models [57, 17, 68, 7]. These models retain many of the features of Bayesian models, including probabilistic opinions and an updating rule that depends on some truth-related likelihood function. However, they may modify or simplify the updating process in some way. The original model we propose can be included in this second category.

2.1.3 The topology of social interactions

Understanding the formation of opinions through interactions between agents involves not only modeling the interactions themselves but also considering the structure of the interactions. We finish this short overview of the field of opinion dynamics by looking at common approaches for modeling the structure of social interactions.

In a society, individuals don't interact or observe the actions of every other individual. Rather, they are only directly in contact with a subset of all individuals. This phenomenon can be captured through an interaction network which describes who is interacting with whom. Agents sit on vertices (nodes) of the network, and the edges define possible interactions between agents (i.e. neighbouring agents) [19].

To capture the complex topology of real life social interactions, modelers embed agents in artificial complex networks. These networks try to capture mathematical properties of real life networks such as average distance between randomly chosen vertices, degree distribution (where the degree of a vertex is its number of connections) or density of edges. Some of the most widely used are i) *Erdos–Rényi random graphs* [32]: constructed by creating an edge between each pair of vertices with probability p and characterised by a Binomial degree distribution, ii) *Small-World networks* [79]: characterised by the average distance between vertices growing logarithmically with the size of the network, and iii) *Scale-Free networks* [15]: constructed through preferential attachment of vertices and characterised by a power-law degree distribution. Nevertheless, these artificial networks differ in many aspects from real social networks. The field of Social network analysis (SNA) can help bridge this gap by investigating social structures through the use of networks and graph theory. Some of the findings in SNA have inspired new attempts at artificially capturing the features of real social networks [74, 47].

A different approach for obtaining more realistic networks to study models in, is to capture the structure of real life networks in datasets of vertices and edges. These datasets can then be used to study the model of interest in the same way as artificial networks. Although various types of real networks can be examined, social media networks of friendships and connections are a reliable source of data and capture many aspects of social interactions. The data for several of the networks used in the literature [14] is publicly available online [2].

2.2 Confirmation Bias in Opinion Dynamics

When looking at the evolution of beliefs in collectives, it is essential to understand how belief formation works at the individual level and what cognitive phenomena might influence this process. Cognitive biases are systematic errors in thinking that occur when people are processing and interpreting information. Several of these biases have an effect on opinion formation, but perhaps the most notable is confirmation bias.

2.2.1 Confirmation Bias

Confirmation bias is the search for and assimilation of information in a way that is partial to prior beliefs [61]. On top of being very well known and widely accepted, the behavioural implications of this bias make it specially relevant to the field of opinion dynamics. Confirmation bias is an umbrella term which encompasses different behaviours. For the scope of this dissertation, we will consider the following: i) *biased evaluation*: being more critical of information which contradicts ones beliefs [52, 72], ii) *biased assimilation*: being less influenced by opposing than confirmatory information [55], iii) *biased seeking*: seeking information that confirms ones beliefs [66], iv) *belief perseverance*: the tendency to preserve ones beliefs in the face of dis-confirmatory information [13]. These four confirmation biases are amongst the most important and empirically validated. However, it is worth pointing out that many more confirmation biases have been identified [50].

It is also worth noting that it is hard to establish a clear division between these behaviours, as there is some overlap. For instance, *biased assimilation* and *belief perseverance* are strongly linked, since being less influenced by opposing information will lead to preserving ones beliefs. As a result, we will use the same modelling assumption for both and will refer to these behaviours simply as *biased assimilation* (see 2.2.3).

2.2.2 Explicit and Implicit modelling of the bias

Numerous models, both Non-Bayesian and Bayesian-inspired, introduce some form of confirmation bias in the opinion updating process. Some explicitly account for confirmation bias [28, 68, 34, 22, 62, 25, 56, 8], while others can be argued to implicitly model the biased behaviour through concepts like bounded-confidence and stubbornness of agents [82, 58, 65].

Although simplistic Non-Bayesian models of confirmation bias exist [28, 22], it is useful to model biases with explicit modifications to Bayesian inference. The reason for this is that Bayesian reasoning can be considered a good baseline for rationality and allows biases to be modeled as deviations from this baseline. This Bayesian-inspired approach for modelling biases has been used in opinion dynamics [68, 62, 8], and is very common in many different contexts [37, 64].

Nevertheless, modelers in opinion dynamics rarely specify which behaviour of confirmation bias they are taking into account. Instead, they are often content with using the umbrella term *confirmation bias* to justify their modelling assumptions. In this project, we try to categorise some of the assumptions made in the literature, with regards to confirmation bias, into the corresponding behaviours. We will then draw from these assumptions to build a new Bayesian-inspired model with continuous opinions that accounts for the four biased behaviours presented in the previous section. The next three sections each present a type of assumption made to model confirmation bias and the corresponding behaviours associated with the assumption. For each assumption, we summarise the results that have been previously obtained.

2.2.3 Biased Assimilation and Belief Perseverance

Sobkowicz proposes a model of opinion updating which introduces a filter function FL applied to the Bayesian update as a generalized framework to account for irrational inference and cognitive biases [68]. In this model, upon observing s, an agent i with prior opinion f_i updates his belief to the posterior by following the equation:

$$f_i(\theta|s) \propto f_i(\theta)\mathcal{L}(\theta|s)FL_i(\theta) \tag{2.7}$$

where $\mathcal{L}(\theta|s)$ is the likelihood associated with some truth related observation s and FL_i is the filter function of the given agent. Sobkowicz goes on to define two filter functions to account for confirmation bias and motivated reasoning. The confirmation bias filter which we will focus on is simply taken to be the agents prior. In order to calibrate the strength of the bias Sobkowicz introduces a filtering efficiency parameter $0 \leq \lambda \leq 1$ leading to the following filter function:

$$FL_i(\theta) = \lambda f_i(\theta) + (1-\lambda)U(\theta) \tag{2.8}$$

where U is a uniform distribution. Sobkowicz investigates the effects of the confirmation bias filter in a very simple scenario without agent interactions. Instead, agents are repeatedly exposed to the same observation s, and they all take the observation to be equally noisy. The observed phenomenon in this study is that with low values for the filtering efficiency parameter λ, agents mean beliefs will converge to values close to s. On the other hand, for higher values of λ, agents beliefs are less affected by the repeated observation of s and stabilize at a greater distance from the information being presented to them (see Fig. 3.2). Given this emergent property and the dynamics of singular updates (see Fig. 3.1), we take the filtering to account for *assimilation bias* and *belief perseverance*. For simplicity, we will refer to the effects of this assumption as *assimilation bias*.

2.2.4 Biased Evaluation

A model proposed by Martins and later extended by Adams introduces the notion of *trust* within Bayesian updating with continuous opinions [57, 7]. The opinion $f_i(\theta)$ of an agent i is taken to be normally distributed. That is, $f_i(\theta) = N(\mu_i, \sigma_i^2)$ where μ_i can be taken to be i's belief about θ, and σ_i can be understood as i's uncertainty surrounding the belief. Here, the observation s is the mean belief μ_j of another agent drawn randomly from the population. Based on the observation, agents update their beliefs through a modified version of Bayesian inference. The main deviation from Bayesian updating stems from the *likelihood*, which is taken to be a mixture of a normal distribution with mean μ_j and a uniform distribution U. Thus the *posterior* is given by

$$f_i(\theta|s) \propto f_i(\theta)\mathcal{L}(\theta|s) \tag{2.9}$$

where

$$\mathcal{L}(\theta|s) = pN(\mu_j, \sigma_S^2) + (1-p)U(\theta) \tag{2.10}$$

and p is taken to be the level of *trust*. The choice of σ_S varies across the work of Martins and Adams, mainly taken to be either σ_i or σ_j (more on this in Section 3.1.2). The effect of this choice of likelihood, is that as the distance between beliefs $|\mu_i - \mu_j|$ increases, the observation μ_j becomes less informative. This is a product of the fact that a uniform likelihood is uninformative (i.e doesn't add any information about the parameter of interest).

As previously mentioned, Martins [57] introduces this modelling assumption thought the notion of *trust*. That is, agents are more likely to *trust* agents with opinions similar to theirs than agents with conflicting opinions. Generalising this to any source of information, it is easy to see how this modelling assumption is in line with the *biased evaluation* behaviour: agents are more critical of information contradicting their beliefs i.e. they consider the information to be less informative about the underlying variable.

Both Martins and Adams consider the setting of sequential interactions between agents. That is, at each time step, two agents are randomly chosen from the set of all agents, they share their opinions, and update their beliefs accordingly. In this context, a range of different assumptions is made by the two papers: taking $\sigma_S = \sigma_i$, taking $\sigma_S = \sigma_j$, taking agents uncertainty to remain constant, or updating the uncertainty with each belief update. The different settings lead to different outcomes, but the most notable result is that the introduction of p leads to the fragmentation of opinions into clusters, where the number of clusters depends on the strength of agents priors (how small their initial uncertainty is). Although polarisation is mentioned as one of the outcomes in Adams, this label is used to refer to the fragmentation of opinions into two clusters. For consistency with other literature on polarising models, we won't consider such an outcome to be polarising, but simply a case of clustering (See Sec. 3.3.2).

2.2.5 Biased Seeking

Finally, when modelling confirmation bias in large social networks, a common approach is to alter the listening structure or the network of interactions in a way that is partial to the agents belief. It is easy to see how this modelling assumption represents the search for information that confirms ones preexisting beliefs, and thus, *biased seeking*.

This modelling assumption is usually accompanied by Non-Bayesian simplistic update rules [22, 28, 67, 44] and is often referred to as preferential rewiring. That is, the connections (edges) between agents are rewired in a way that gives rise to homophily, were agents are more likely to be connected to agents with similar opinions. While some models such as the Rewire Bounded Confidence Model [28] consider the rewiring as preliminary or independent to the process of opinion formation, the more realistic setting is that of adaptive or co-evolving networks, where the topology of social interactions and opinions both change through time. Such a setting gives rise to a sort of feedback loop where opinions influence the topology of the network, which in turn influences the change in opinions [76]. The usual process of rewiring goes as follows: when two agents i and j with respective opinions x_i and x_j interact, if the distance between the opinions ($|x_i - x_j|$) is greater than a given bound ε, there is a probability ρ that agents will break this connection and create a new connection with another (usually random) agent. As we will be dealing with probabilistic opinions, the distance between opinions can be taken to be the distance between opinion means $|\mu_i - \mu_j|$.

Given the range of different models that have been studied in co-evolving networks, it is hard to summarise the findings emerging from this modelling assumption. Nevertheless, preferential attachment is often linked to making consensus between agents more difficult to achieve [48]. When rewiring is introduced in the context of modelling confirmation bias, it has been shown that the social network will tend towards segregated, homogeneous communities [22, 67].

2.3 Proposed work

Our contributions to previous work can be summarised as follows:

1. Developing a new model which incorporates *biased assimilation* through Sobkowicz's filtering function, *biased evaluation* through the modified likelihood proposed by Martins, and *biased seeking* through rewiring social connections.
2. Studying the effects of *biased evaluation* and *biased assimilation* in the simple settings of single observations and fully connected networks.
3. Studying the effects of *biased seeking*, how they relate to the properties of the network structure and to the presence of the other biased behaviours.
4. Determining the role of each biased behaviour in the emergence of polarisation.
5. Validating the ability of the model to reproduce opinion distributions observed in real life, through social structures taken from four different social media datasets.

The models proposed by Sobkowicz [68] and Martins [57] have been studied for single observations and sequential interactions respectively. Thus, the aforementioned results emerging from these models are obtained in very simple and unrealistic settings. The new model we propose can shed light on the implications of both modelling assumptions in more complex settings, and the behaviours emerging from their interactions. On the other hand, the modeling assumption for *biased seeking* can only be considered in complex social networks often leading to segregated communities, with low disagreement within communities. In this case, we are interested in knowing how other biases and network structure might influence the emergence of the low disagreement observed in the literature. When considering the effects of each bias on polarisation, we will take a polarising outcome to entail both the fragmentation of opinions into two main clusters, and an overall increase in opinion diversity (See Sec. 3.3.2). Finally, we use network structure data from Twitter, Flickr, Advogato and Facebook to asses the qualitative validity of the model.

Chapter 3

Methods

3.1 Model

We start by presenting the model we will use, accounting for *biased assimilation* and *biased evaluation* through modifications to Bayesian inference, and introducing *biased seeking* through the rewiring of the social network. We first introduce the general framework for opinion updating, and then we introduce our assumptions about the reliability of observations σ_S.

3.1.1 General Framework

We start with a set of agents $N = \{1, 2, ..., n\}$, which are trying to estimate a parameter θ taking values in the range $[-1, 1]$. For instance, θ could be taken to be the level of honesty of a politician where -1 would indicate complete dishonesty, 1 indicates complete honesty and the rest of the values represent some point in between. Each agent i has a set of neighboring agents N_i determined by the graph of social interactions $G(V, E)$ with set of vertices (nodes) V, and set of edges E. Each node is an agent and each edge is a social connection with another agent. We take $G(V, E)$ to be a directed graph, accounting for the fact that social connections aren't always reciprocal (especially in social media).

At time $t = 0$, agents hold normally distributed prior beliefs about θ. As such, agent i's prior belief about θ is given by the pdf $f_i(\theta) = N(\mu_i(0), \sigma_i(0))$. At time each time step t, agent i observes the mean belief $\mu_j(t)$ of an agent j randomly selected from its neighbours ($j \in N_i$). The observed opinion is interpreted by agents to be a noisy measurement of the true value of the parameter θ. Thus, we will assume that agents take the opinions they observe to be sampled from a normal distribution with mean given by the true value of θ, and standard deviation σ_S given by the noisiness or reliability of the observation. Based on the observed opinion, agent i updates his belief to the *posterior* given by:

$$f_i(\theta|\mu_j(t)) \propto f_i(\theta)\mathcal{L}(\theta|\mu_j(t))FL_i(\theta) \tag{3.1}$$

where $FL_i(\theta)$ introduces the *biased assimilation* and is defined as:

$$FL_i(\theta) = \lambda f_i(\theta) + (1-\lambda)U(\theta) \tag{3.2}$$

and the modified likelihood introduces *biased evaluation* and is defined as:

$$\mathcal{L}(\theta|\mu_j(t)) \propto (1-\delta)N(\mu_j, \sigma_S^2) + \delta U(\theta) \tag{3.3}$$

After each iteration, the posterior becomes the new prior, and this process is repeated.

To account for the *biased seeking*, at each time step, if the distance between the observed mean opinion μ_j and the agents mean opinion μ_i is bigger than a given threshold ε, there is a probability ρ that the connection from i to j will be rewired. Given that we are working with probabilistic opinions, we will take $\varepsilon = 6 \times \sigma_i(0)$. Note that initially, agent i considers that the true value of θ is in the range $(\mu_i(0) - 3\sigma_i(0)\ ,\ \mu_i(0) + 3\sigma_i(0))$ with 0.997 probability. If he considers that agent j has a similar uncertainty, then a distance between their mean beliefs greater than $6\sigma_i(0)$ indicates that there is virtually no overlap between their beliefs, and we can use this as a basis for disappearance of that social link. Thus after agents i observes j's opinion, if

$$|\mu_i(t) - \mu_j(t)| > 6\sigma_i(0) \tag{3.4}$$

there is probability ρ that the edge from i to j will be removed, and a new edge will be created from i to another random agent (not already connected to i). The choice of $\sigma_i(0)$ instead of $\sigma_i(t)$ stems from the fact that agents uncertainties will approach 0 especially quickly with the introduction of *assimilation bias*. As such, using $\sigma_i(t)$ would result in agents rewiring every connection.

In short, the proposed model can be described by 3 parameters:

- *biased assimilation* parameter λ: controlling the filtering efficiency of FL.
- *biased evaluation* parameter δ: controlling how uninformative the likelihood becomes for disagreeing agents.
- *biased seeking* parameter ρ: controlling the rewire probability between disagreeing agents.

3.1.2 Determining noise

When looking at the general framework, we mention that each observation is associated with some level of noise σ_S. Different choices are made for this value in the works of Martins and Adams [57, 7]. Martins takes σ_S to be equal to the agents uncertainty $\sigma_i(t)$. On the other hand, Adams introduces the notion of "shared uncertainty", were σ_S is taken to be the observed agents uncertainty $\sigma_j(t)$.

Both of these assumptions deviate from the idea of learning that Bayesian models introduce. In Bayesian models, an agent will reduce its uncertainty as it makes more observations. As a result, after making many observations (for higher time steps), it will be less swayed by a single new observation. With the assumption introduced by Martins [57], since $\sigma_S = \sigma_i(t)$, an agent will be equally swayed by an observation at any time step. Since all agents uncertainties decrease over time, taking $\sigma_S = \sigma_j(t)$ has a similar effect.

Furthermore, with the introduction of the *assimilation bias*, agents uncertainties will decrease faster, making $\sigma_i(t)$ and $\sigma_j(t)$ bad candidates for the reliability of the observation σ_S. Instead, we will take σ_S to be the agents initial uncertainty $\sigma_i(0)$. Intuitively, with this assumption, an agent considers the reliability of his observations to be equal to his uncertainty about the variable before the learning process starts. Although still far from perfect from a Bayesian standpoint, this assumption implies that as agents make more observations, their opinion will be less affected by a single observation. Thus, some notion of learning typical of Bayesian models is still preserved.

3.2 Simplifying Assumptions

The model we have presented leads to an update rule that becomes too computationally expensive in complex settings. This is mainly a product of the introduction of the filter function FL used to model *biased assimilation*. Thus, we here make some simplifying assumptions which allow us to derive an analytical solution to the update rule, while also showing that the emergent behaviour that arises from varying the parameter λ remains the same as for the model proposed by Sobkowicz.

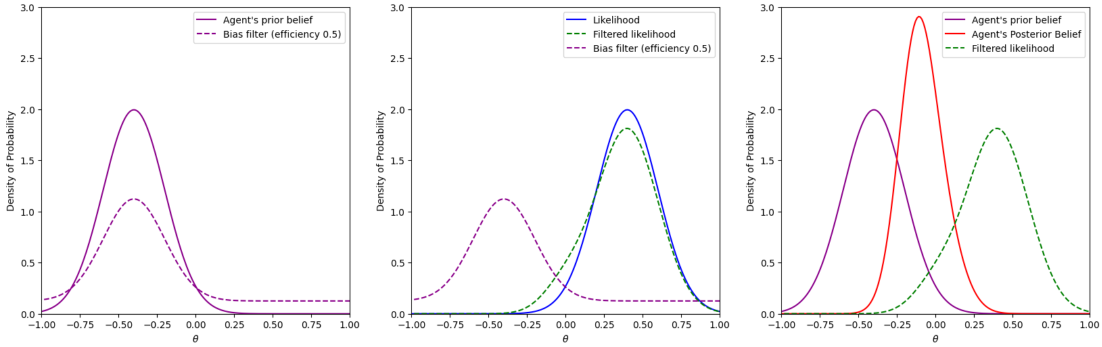
3.2.1 Computational challenges of Sobkowicz's model

Under the updating rule proposed by Sobkowicz (Eq. 2.7), even if normally distributed priors are assumed, the posteriors won't be normally distributed. Further, intermediate distributions required to make the calculations won't be normally distributed either. Note that given the commutativity of the product operation, Sobkowicz's update can be interpreted in two ways:

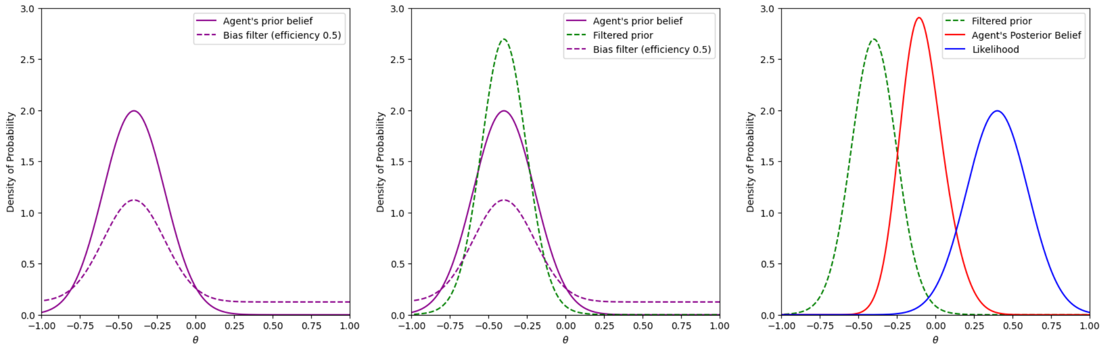
- Taking the product of the filter function and the likelihood, and then taking the product of the resulting filtered likelihood and the prior (Fig. 3.1a)
- Taking the product of the filter function and the prior, and then taking the product of the resulting filtered prior and the likelihood (Fig. 3.1b)

Both orders of derivation are identical and result in the same posterior. We see in Fig. 3.1 that the resulting posterior under Sobkowicz's updating is skewed right with respect to a normal distribution. Although less clear in the filtered prior case, the intermediate distribution (in green) is also not normal.

The distributions resulting from repeated updating are complex in the sense that they can't be captured by a predictable number of parameters (such as mean and variance). A common approach to deal with such distributions computationally, is to reduce them to probability mass functions which assign probabilities for θ, in intervals of a given size in $[-1, 1]$. The probabilities for all intervals can be stored in an array which approximately



(a) Filtered likelihood as intermediate distribution



(b) Filtered prior as intermediate distribution

Figure 3.1: Computational interpretations of belief updating as proposed by Sobkowicz: (a) taking the intermediate distribution (in green) to be the filtered likelihood $\mathcal{L}(\theta|s)FL_i$, (b) taking the intermediate distribution to be the filtered prior $f_i(\theta)FL_i$. Both cases lead to the same posterior (in red), which is not normal.

represents the distribution. To obtain the products of pdf's necessary for opinion updating, one can iteratively perform element-wise multiplication of the distribution arrays and normalize the result. This is in fact the way that Sobkowicz chooses to model distributions in his simulations (using intervals of length 0.1). Although this is a valid approach for the scope of Sobkowicz's work, each update of an agents opinion requires hundreds of operations. When dealing with large social networks, the computational requirements become far too great to perform a comprehensive study of the model.

3.2.2 Vectorisation

A common approach to optimize computations involving arrays or matrices is the use of vectorisation, that is, implementing array operations in such a way that the operations for each element of the array are preformed in parallel. We use the python library Numpy [41] to vectorise all operations. Although improvements are considerable with respect to an iterative method (Fig. 3.3), the computation time is still too great. Further, using arrays to capture probability density functions create problems other than computational complexity (See Appendix A). We will discuss model implementation further in Sec. 3.3.1.

3.2.3 Simplifying Assumptions for Analytical Solution

Given that vectorisation isn't enough for our computational needs, and added problems that come with using arrays to represent distributions, we propose a simplifying assumption which allows us to find an analytical solution and thus, faster and more precise calculations.

A widely used assumption in the field, is that of normally distributed opinions: a normal distribution is enough to capture some value of the opinion and the uncertainty that comes with it. Although an analytical solution is possible under this assumption (See Appendix B), the mixture likelihood introduced by *biased evaluation* makes the derivations far too complex. Thus, we opt for a slightly stronger assumption which is taking the intermediate distribution (Fig. 3.1 in green) to be normally distributed. This leads to a normal posterior since the product of normally distributed pdfs is itself normally distributed. As there are two possible intermediate distributions, there is a choice to be made with regards to taking the *filtered prior*, or the *filtered likelihood*. We will take the intermediate distribution to be the *filtered prior* given that it is not skewed, and thus more similar to a normal distribution. Further, taking this intermediate distribution leads to the simplified update which behaves closer to the model proposed by Sobkowicz (see in Fig. 3.2).

Thus, in the opinion update rule given by Eq. 3.1, we first compute the filtered prior fl_i given by:

$$\begin{aligned} fl_i(\theta) &\propto f_i(\theta) FL_i(\theta) && (3.5)\\ &\propto \lambda f_i(\theta)^2 + (1-\lambda) f_i(\theta) && (3.6) \end{aligned}$$

and then we take the product of the filtered prior and the likelihood $fl_i(\theta)\mathcal{L}(\theta|\mu_j(t))$ to obtain the posterior. Assuming fl_i is normally distributed: $fl_i(\theta) = N(\mu_i'(t), \sigma_i'^2(t))$, we derive its mean $\mu_i'(t)$ by taking the expected value of (3.6). The standard deviation σ_i' can then be derived and through the standard result for random variables $Var[\theta] = E[\theta^2] - E[\theta]^2$. We get:

$$\mu_i'(t) = \mu_i(t) \tag{3.7}$$

and

$$\sigma_i'(t) = \frac{\lambda \frac{\sigma_i(t)}{4\sqrt{\pi}} + (1-\lambda)\sigma_i^2(t)}{\frac{\lambda}{2\sqrt{\pi}\sigma_i(t)} + (1-\lambda)} \tag{3.8}$$

The opinion update rule is now given by

$$f_i(\theta|s) \propto fl_i(\theta)\mathcal{L}(\theta|s) \tag{3.9}$$

and we can obtain an analytical solution for this equation assuming that $f_i(\theta|s)$ is normally distributed. This solution is presented in Sec. 3.2.5.

3.2.4 Computational improvements, and preserved behaviour

We have shown that through our proposed simplifications, an analytical solution can be found for the process of updating with biased assimilation. To show that the proposed simplifying assumption doesn't have an effect on the qualitative behaviour of the model, we replicate the main result from Sobkowicz with regard to the confirmation (assimilation) bias filter. That is, we take agents to be exposed to the same observation $s = 0.6$ at each iteration with probability $p = 0.3$, which they will take to be sampled from a distribution with $\sigma_S = 0.4$. In other words, each iteration, agents update their beliefs using the likelihood $\mathcal{L}(\theta|s) = N(0.6, 0.4)$ with probability $p = 0.3$, and keep their old beliefs with probability $1 - p = 0.7$.

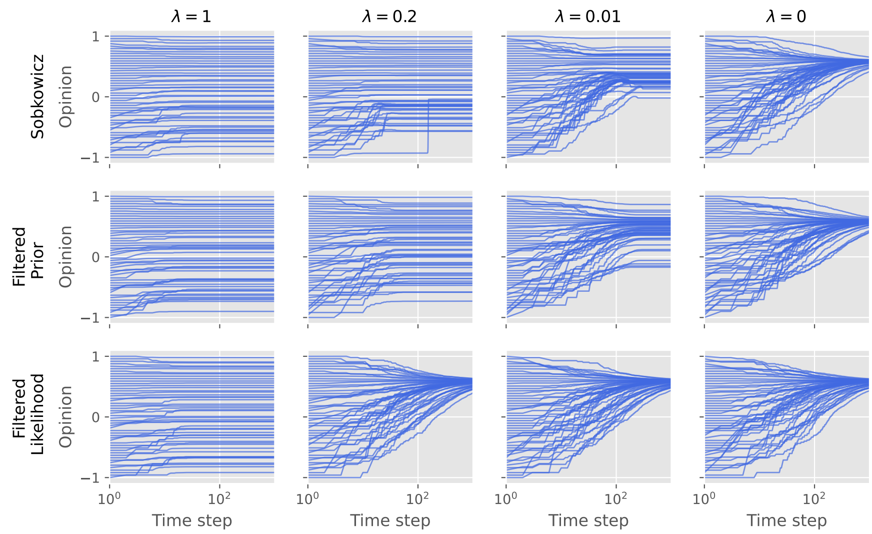


Figure 3.2: Evolution of mean beliefs of 50 agents repeatedly updating their beliefs through the likelihood $\mathcal{L}(\theta|s) = N(0.6, 0.4)$, and with varying *assimilation bias* strength λ. Simulations are run with Sobkowicz's model (top), with our simplified model taking the filtered prior as intermediate distribution (middle), and with our simplified model taking the filtered likelihood as intermediate distribution (bottom).

We observe that the qualitative behaviour of the model proposed by Sobkowicz is preserved with the introduction of our simplifying assumption (with filtered prior). That is, being exposed to the same signal repeatedly, agents beliefs will stabilize at a greater distance from the observation as the filtering efficiency λ increases (See Sec. 4.1 for further discussion of updating under a single observation). Taking the filtered likelihood as the intermediate distribution alters the sensitivity of λ, further justifying our choice of using the filtered prior as the intermediate distribution.

Fig. 3.3 shows how the computational times for simulations scale as a function of the number of agents simulated. We see that the introduction of an analytical update rule leads to much faster simulations which are therefore scalable to larger networks and more complex interaction patterns (rewiring).

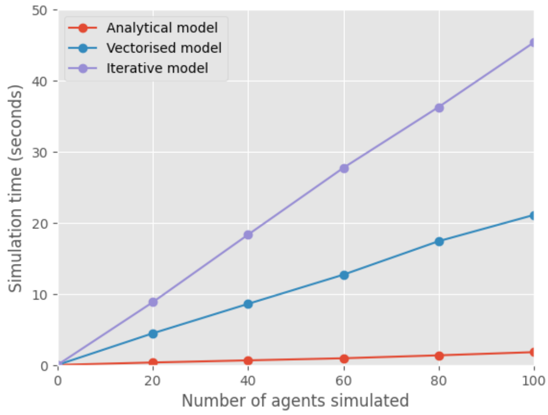


Figure 3.3: Computation times for simulations with varying numbers of agents and different implementation methods.

3.2.5 Final update rule

Given our simplifying assumptions, agents beliefs are captured by normal distributions. Thus, the process of opinion updating for a single agent can be described through the update of its mean and variance. At time step t take agent i to have the prior belief $f_i(\theta) = N(\mu_i, \sigma_i^2)$, when agent i observes the mean belief μ_i of agent j, it updates its opinion to the posterior given by

$$f(\theta|\mu_j) \propto fl_i(\theta)\mathcal{L}(\theta|\mu_j) \tag{3.10}$$

where $fl_i(\theta)$ is normally distributed, and $\mathcal{L}(\theta|\mu_j(t))$ is the mixture likelihood proposed by Martins. By assuming the posterior is normally distributed: $f(\theta|\mu_j) = N(\mu_i(t+1), \sigma_i^2(t+1))$, the mean and variance can be analytically derived as shown in Adams [7] through:

$$\mu_i(t+1) = \delta^* \frac{\mu_i\sigma_{i,0}^2 + \mu_j\sigma_i'^2}{\sigma_{i,0}^2 + \sigma_i'^2} + (1-\delta^*)\mu_i \tag{3.11}$$

$$\sigma_i(t+1) = \sigma_i'^2\left(1 - \delta^* \frac{\sigma_i'^2}{\sigma_i'^2 + \sigma_{i,0}^2}\right) + \delta^*\left(1 - \delta^*\left(\frac{\mu_i - \mu_j}{1 + (\sigma_{i,0}/\sigma_i')^2}\right)\right) \tag{3.12}$$

where

$$\delta^* = \frac{(1-\delta)\phi}{(1-\delta)\phi + \delta} \tag{3.13}$$

and

$$\phi = \frac{1}{\sqrt{2\pi(\sigma_i'^2 + \sigma_j^2)}} e^{-\frac{(\mu_i - \mu_j)^2}{2(\sigma_i'^2 + \sigma_{i,0}^2)}} \tag{3.14}$$

with σ_i' derived as per Eq. 3.8.

3.3 Implementation, simulations and metrics

In the previous sections, we have introduced the proposed model, and simplifying assumptions that allow us to run simulations in more complex contexts than the ones explored in previous work. We briefly go over how the simulations are implemented and run, and we consider the metrics we will use to study the evolution of agents opinions.

3.3.1 Implementation and Simulations

The full model, and simulations are implemented from scratch using Python 3. All code is publicly available [1], and the results presented here can be easily reproduced by running Jupyter Notebooks which are also made available. Given the complexity of the model being implemented, and the number of different scenarios we will consider, we use the Object Oriented paradigm to improve the reusability of code. Using this paradigm not only ensures consistency within our own work, but also facilitates reproducibility, further extensions, and makes the code more readable. We use the well established library Networkx [39] to deal with complex networks, Matplotlib [45] and Seaborn [77] for visualizations, and Numpy [41] for vector operations.

3.3.2 Disagreement, Diversity and Polarisation

In some cases, an analysis of an opinion formation model can be carried out by plotting agents opinions over time and studying the emerging phenomena. Nevertheless, in more complex scenarios, this method will fail to capture nuances such as the relationship between opinions and the underlying network structure. These more intricate relationships can be studied through metrics defined over the opinion space, which in turn allow for a more quantitative analysis of the model. For our purposes, we will focus on two metrics: i) *Diversity*: Which captures how much agents opinions differ from the mean opinion and ii) *Disagreement*: Which captures the difference in opinion between neighboring agents. Given a vector of opinions μ and a graph $G(V,E)$ representing the network of social interactions, we define *Diversity* $Dv(\mu)$ and *Disagreement* $D_G(\mu)$ as follows:

$$Dv(\mu) = \sum_{i=1}^{|V|} \frac{(\mu_i - \overline{\mu})^2}{|V|} \tag{3.15}$$

$$D_G(\mu) = \sum_{(i,j)\in E} \frac{(\mu_i - \mu_j)^2}{|E|} \tag{3.16}$$

An unnormalised version of these metrics was introduced by Musco et al. [60] and latter used by Chen [22] where *Diversity* was instead labeled *Polarisation*. We find that the term *Diversity* used by Stern [69] is better suited, as the transition from *Consensus* to *Dissensus* leads to an increase in diversity, but shouldn't be considered *Polarising*. Instead, we consider *Polarisation* to be captured by both an increase in *Diversity*, and the fragmentation of the opinion space into two main clusters. We choose to normalize these metrics by the number of vertices (agents) $|V|$ and number of edges $|E|$ respectively, since we will be dealing with networks of different sizes.

Chapter 4

Results

4.1 Single Observation

As we are working with a complex update rule (complex agents), it is useful to consider the model in simple settings before moving to complex social networks. First, we investigate the implications of the update rule by considering the case where agents don't interact, and are repeatedly exposed to the same observation. Given the lack of agent interactions, the *seeking bias* is irrelevant. Instead, we focus on the implications of *assimilation bias* and *evaluation bias* in our proposed update rule.

4.1.1 Impact of varying bias strength

As we have mentioned, we consider the case where instead of observing other agents opinions, all agents repeatedly make the same observation s. Agents take this observation to be related to the parameter θ they are trying to estimate, that is, they take s to be an observation of a random variable $S \sim N(\theta, \sigma_s^2)$. Given this observation, they update their opinions according to Eq. 3.1. While this is a very unrealistic and oversimplified scenario, it is useful to understand the low level implications of the proposed update rule.

We take a set of 30 agents trying to estimate the parameter θ. At each time step, all agents observe the same signal $s = 0.6$, which they take to be an observation of $S \sim N(\theta, \sigma_s^2)$ where $\sigma_s = 0.3$. All agents have the same initial uncertainty given by $\sigma(0) = 0.2$, and mean beliefs of agents are uniformly distributed between -1 and 1. For clarity, we are taking s to have a constant value, instead of drawing s from the distribution $N(0.6, 0.3^2)$. Drawing s from the distribution simply introduces noise, and doesn't affect the emergent behaviour (see Appendix C).

Figure 4.1 shows the evolution of agents mean beliefs, resulting from simulating the repeated update we have described. Different values for the *biased evaluation* δ and *biased assimilation* λ parameters of the update rule are considered. In the trivial case ($\lambda = 0$, $\delta = 0$), all agents eventually adopt the opinion corresponding to the observed signal $s = 0.6$.

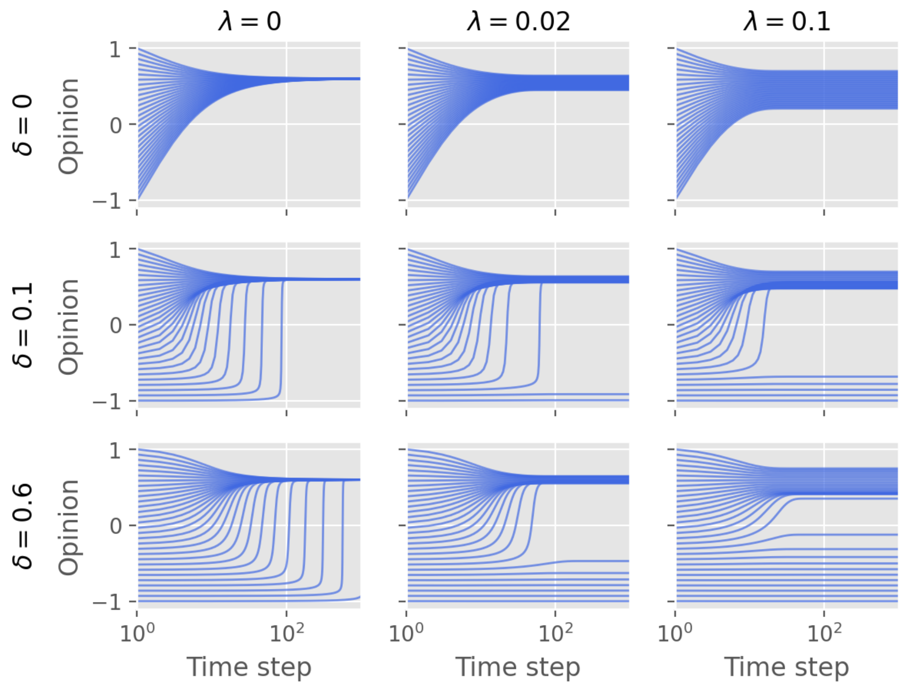


Figure 4.1: Evolution of agents mean beliefs under repeated observation of $s = 0.6$ taken to be sampled for $S \sim N(s, \sigma_S^2 = 0.3^2)$, varying *evaluation bias* δ and *assimilation bias* λ parameters. Agents initial uncertainties are taken to be $\sigma(0) = 0.2$. With increased *biased assimilation* agents beliefs stabilize further from s, with increased *biased evaluation* it takes more iterations for agents to change their beliefs towards s, and when both biases are introduced disagreeing agents will preserve their prior.

The introduction and increase of the *biased evaluation* parameter δ has the effect of delaying agents ability to learn from s based on how distant the prior belief is from the observation. The intuitive interpretation of these results, is that agents discard information as unreliable if it is distant enough from their beliefs. Nevertheless, if the conflicting information is observed enough times, they will eventually recognise its validity. Increasing the evaluation bias parameter δ increases the number of times an agent must be exposed to the conflicting information before changing his beliefs in any substantial way. For $\delta = 0.1$, the agent with initial opinion of -1 changes his beliefs before time step 10^2, while for $\delta = 0.6$, the observation hasn't had a significant effect on the agents beliefs long after time step 10^2.

On the other hand, with the introduction of *biased assimilation* λ, agents change their prior beliefs from the start, but through "squaring" their prior as part of the update, the uncertainty of their beliefs decreases faster. This together with a reduced shift in opinion causes agents opinions to stabilize at a distance from the information being presented to them. We can interpret this phenomena as agents reducing their uncertainty more than they should for each observation, and interpreting information in a biased way. The combination of these elements allows them to hold their beliefs while being exposed to differing information. For higher values of the parameter λ, agents beliefs stabilize sooner and at a greater distance from the value of s.

Finally, when introducing both biases at the same time, we see a third pattern of behaviour arise. Agents that have priors distant enough from the truth, reduce their uncertainty about their opinion even when observing information they consider uninformative. For combinations ($\delta = 0.6, \lambda = 0.02$) and ($\delta = 0.1, \lambda = 0.1$) we see two main groups of agents emerge: agents who eventually learn the truth (albeit with some error), and agents who preserve their prior beliefs. For higher values of the bias parameters ($\delta = 0.6, \lambda = 0.1$), more agents will tend to simply preserve their prior beliefs.

This setting is very similar to the one considered by Sobkowicz [68] and shows a less noisy version of the implications of *biased assimilation*. In the case of *biased evaluation*, the scenario of repeated observations hasn't been considered in previous work and provides a good overview of how the behaviour emerging from the modified likelihood aligns with the intuition behind *biased evaluation* of information. The interactions between both biases leads to a new dynamic where agents opinions are indefinitely unaffected by the observation.

4.1.2 Departure from Bayesian intuitions

We have shown that the behaviour of the model in the setting of repeated observations seems to be consistent with the ideas of *biased evaluation* and *biased assimilation*. Nevertheless, we here point out that the modifications to Bayesian-inference introduced can change the intuitive meaning of the variables involved.

As we have established, by Bayesian intuitions σ_S indicates the reliability or noisiness of the observation s. Thus, for smaller values of σ_S (a more reliable observation), we would expect a greater shift in opinion after observing s. Nevertheless, we see that this is not necessarily the case with the introduction of the modified likelihood used to model *biased evaluation*. Figure 4.2 shows the same setup as in the previous section, but this time, varying σ_S. We see that in the presence of *biased evaluation*, for smaller values of σ_S, it will take more observations for disagreeing agents to change their beliefs. The reason for this, is that by taking the likelihood to be a mixture of $N(s, \sigma_S)$ and a uniform distribution, reducing σ_S will increase the section of the likelihood which is uniform and thus uninformative about θ.

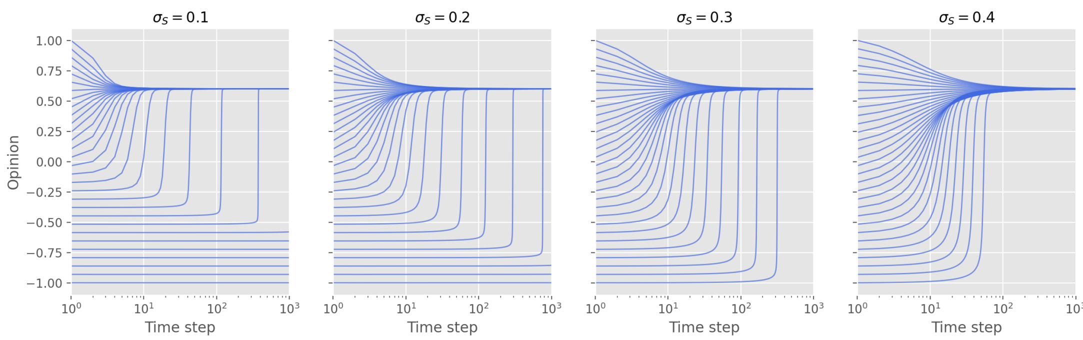


Figure 4.2: Evolution of agents mean beliefs under repeated observation of $s = 0.6$ taken to be sampled from $S \sim N(s, \sigma_S^2)$, with varying values of σ_S and $\delta = 0.6, \lambda = 0$. For smaller σ_S, it takes longer for disagreeing agents to change their beliefs.

A possible interpretation for the role of σ_S in this case would be that smaller values of σ_S indicate stronger observations, that is, observations that are compatible with a smaller set of beliefs (hypotheses). The new interpretation better aligns with the observed behaviour since agents are more likely reject the validity of observations that differ from their beliefs if these observations have stronger implications (i.e. are compatible with a smaller set of beliefs). Although not mentioned in Martins [57] or Adams [7], this deviation from Bayesian intuitions is also present in these models. Consequently the alternative interpretation proposed is also pertinent for their work.

4.2 Fully-Connected Networks

Having a basic understanding of the implications of the update rule, we now move on to the case of interest, where agents observe other agents' opinions. As is often the case in the literature, we are concerned with the transition from a random state of opinions to a stable state. We start by considering agents in a fully connected network where every agent can potentially observe every other agent. Again, in this context, *biased seeking* is irrelevant since rewiring a fully connected network will lead to the same network. We consider the separate and combined effects of *biased evaluation* and *biased assimilation*.

4.2.1 Evaluation bias and fragmentation of opinion

We take a set of 100 agents in a fully connected network, with initial mean opinions uniformly distributed between -1 and 1. Further, take the agents initial uncertainties to be equal and determined by σ_0. To understand the effects of *evaluation bias* on collective opinion formation, we set $\delta = 0.3$. Fig. 4.3 shows the effects of varying the value of agents initial uncertainties in the presence of *biased evaluation*. As agents initial uncertainties get smaller, the final opinion space becomes fragmented into more clusters.

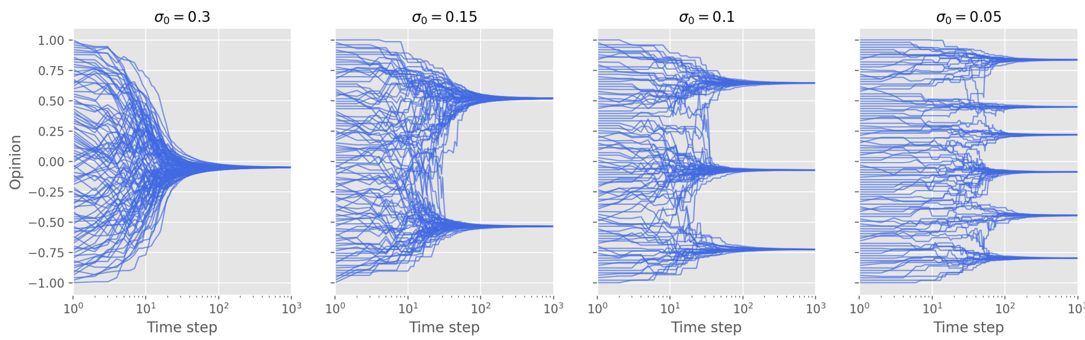
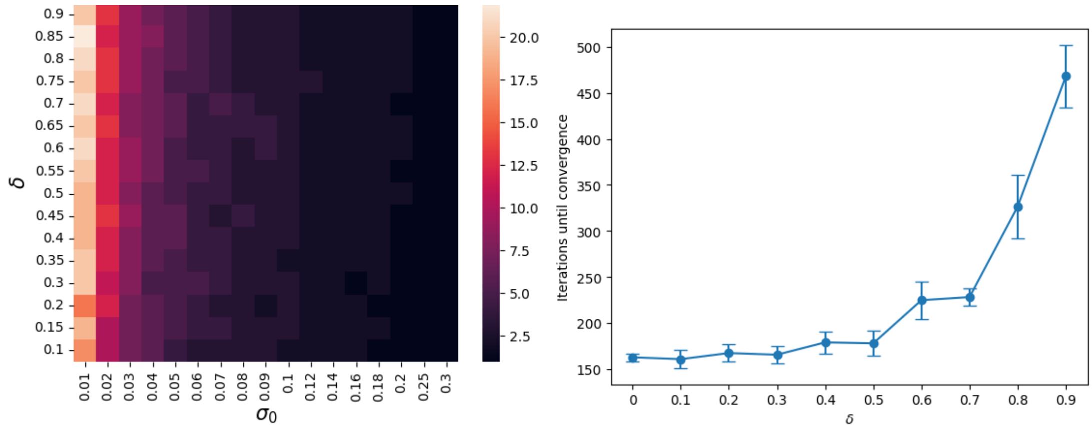


Figure 4.3: Evolution of interacting agents mean beliefs on fully connected network, with varying initial uncertainty and *biased evaluation* $\delta = 0.3$. Opinions become fragmented into more clusters for lower initial uncertainties σ_0.

The core behaviour emerging from *biased evaluation* is the clustering of opinions. This observation is consistent the results observed by Martins [57] where stronger priors also



(a) Number of final opinions

(b) Opinion convergence time

Figure 4.4: Effects of δ and σ_0 on opinion clustering. (a) shows the number of final opinion clusters while varying both parameters. (b) shows the effect of δ on the time it takes for opinions to converge to a stable state (running 5 simulations for each value and plotting respective error bars).

lead to more final clusters of opinion. Note that the ability of our model to replicate such emergent behaviour is not trivial given the differences in update rules and setting: In Martins [57], agents uncertainties are not updated and interactions are sequential (a single agent makes an observation at any given time step).

The level of clustering, or number of emerging opinions is not controlled by the *biased evaluation* parameter δ, but instead by agents initial uncertainties σ_0. As we have previously observed, for higher values of δ, agents will take longer to accept conflicting information, on the other hand, their initial uncertainty σ_0 will determine how distant an observation needs to be for it to be considered as conflicting with their beliefs.

We pursue a stronger investigation into the role of σ_0 and δ in the clustering of opinions, by looking at the final number of opinions (number of clusters) as a function of both parameters. Fig. 4.4a shows the final number of opinion clusters for different combinations of values for δ and σ_0. For each combination of parameters, 3 different simulations are run, and the mean number of clusters is displayed in the heat map. It is clear from Fig. 4.4a the number of opinion clusters is affected mainly by σ_0, revealing that, for fragmentation of opinions, which opinions are considered to be conflicting is more important than how long it takes for a conflicting opinion to be accepted.

Even if the same number of clusters tend to arise with different values of δ, the *biased evaluation* parameter influences how long it takes for these clusters to form. Fig. 4.4b shows the increase in the number of iterations it takes for opinions to converge to a stable state.

4.2.2 Assimilation Bias and Strong diversity

From the previous section, we conclude that *biased evaluation* is responsible for the emergence of a fragmented opinion space. Further, total consensus is reached within each cluster. That is, every agent within a cluster holds the same opinion. This phenomenon is know as *weak diversity* and describes the outcomes of the vast majority of continuous opinion models in strongly connected networks. Nevertheless, the idea of total agreement within clusters which *weak diversity* entails is rarely observed in the real world. Given the limitations of these models, there is a growing effort to capture *strong diversity* [31]. By opposition to *weak diversity*, *strong diversity* is the scenario where there is persistent disagreement even within clusters [12]. We now show that the introduction of *biased assimilation* gives rise to *strong diversity* both in outcomes leading to consensus, and to clustering of opinions.

When considered in isolation, *biased assimilation* leads to diverse consensus: opinions stabilize into a single cluster, but different opinions are held within that cluster. We observe this in Fig. 4.5 were we show the results of simulations run in the same setting as in the previous section, but with $\delta = 0$, and fixed $\sigma_0 = 0.15$. The *assimilation bias* parameter λ controls the diversity of the obtained cluster. For high enough values of λ, agents will tend toward dissensus, where each agent tends to preserve their initial opinion.

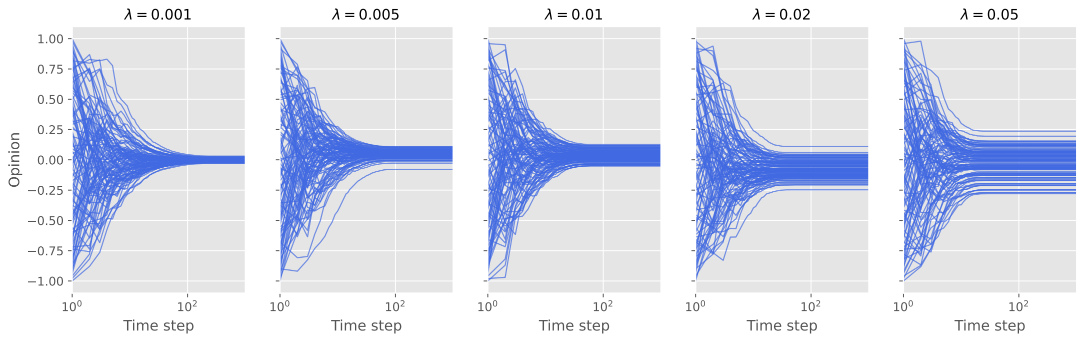


Figure 4.5: Evolution of interacting agents mean beliefs on fully connected network, with varying assimilation bias $\delta = 0$

Strong diversity with opinion clustering emerges when we consider agents with both *assimilation bias*, and *evaluation bias*. This is shown in Fig. 4.6 where we see that again, λ controls the diversity within clusters of opinion.

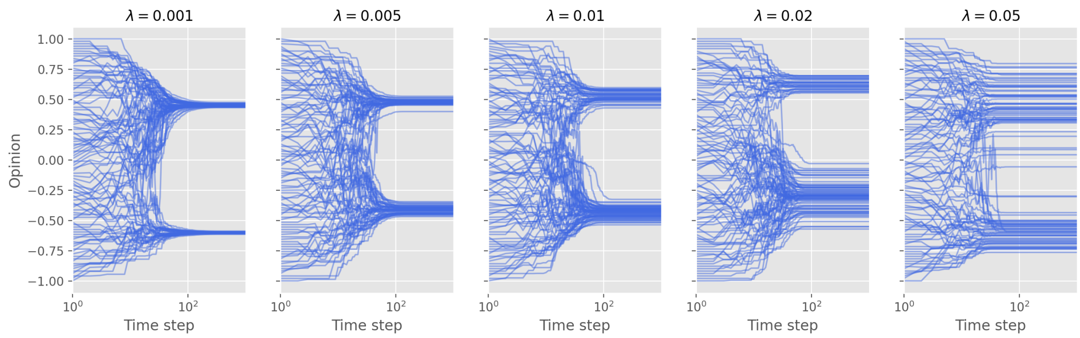


Figure 4.6: Evolution of interacting agents mean beliefs on fully connected network, with varying assimilation bias $\delta = 0.3$.

The presence of *biased evaluation* accelerates the emergence of dissensus with the increase of λ. Further, we observe that the size and diversity of the clusters are not uniform, but are affected by the randomness involved in the simulations.

4.3 Social Networks and the seeking bias

By studying the model in fully connected networks, we have seen that *biased evaluation* leads to a fragmented opinion space and the introduction of *biased assimilation* gives rise to strong diversity, where differing opinions can coexist within opinion clusters. We now move on to the more realistic setting where agents can't interact with every other agent, but instead they are embedded in an interaction network and can only interact with their neighbors. This setting allows us to introduce the *seeking bias* characterized by the parameter ρ. Note that as the *seeking bias* is modeled by rewiring the network, its dynamics can be affected by the structure of the network being rewired. We study the effects of the seeking bias in three different types of artificial networks and consider how different network properties impact these effects.

4.3.1 Dealing with complexity

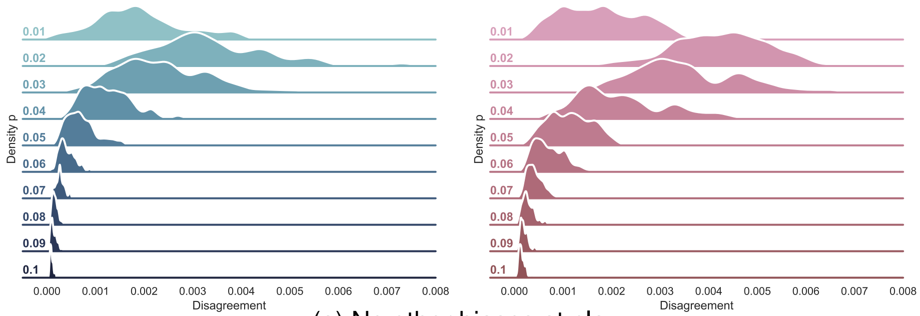
As we increase the complexity of the studied setting the properties of the opinion updating process become harder to analyse by visualising the evolution of agents mean beliefs. It is for this reason that we draw from metrics such as Disagreement, or Diversity presented in the Methods. Since we are interested in how the *seeking bias* is affected by other biases and by network structure, we here focus on Disagreement as it captures both properties about opinions and the topology of the network. The results with regard to Diversity for all the simulations considered in this section can be found in Appendix D.

Since the three networks we will consider are generated through random processes, introducing these complex networks to our simulations adds more randomness to the process of opinion formation. Generating a single network and drawing conclusions from the computed Disagreement would be inappropriate, as the results are likely to be noisy and biased by the specific network constructed. Instead, for each type of network and each set of parameters considered, we generate 100 networks, we run the process of opinion updating on each network, and we look at the distribution of the final disagreement across the 100 generated networks.

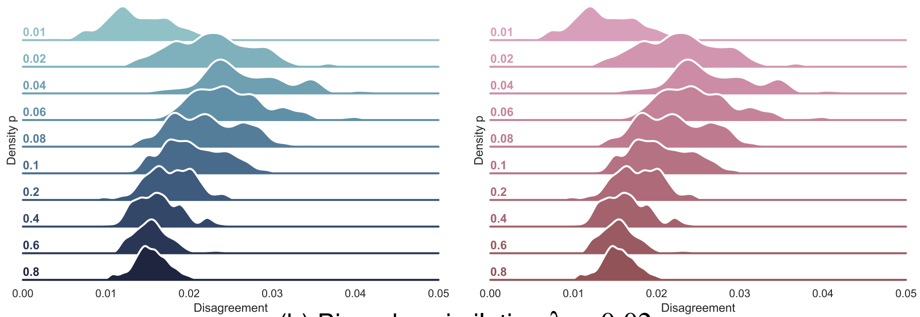
4.3.2 Network Density in Random Netwoks

The first type of artificial network we consider is the Erdös–Rényi random graph [32]. Given a set of nodes, this network is constructed by creating an edge between each pair of nodes with probability p. Although randomly connecting nodes may seem too simplistic, when p is such that each node is connected to at least one other node on average, a giant connected component emerges resembling the random but globally connected nature of social networks. The parameter p is an indicator of the density (also called connectivity) of the network, where the density of a network represents the number of existing edges with respect to the possible number of edges. Real social networks have varying densities, from highly interlinked, to loosely connected individuals [40, 42].

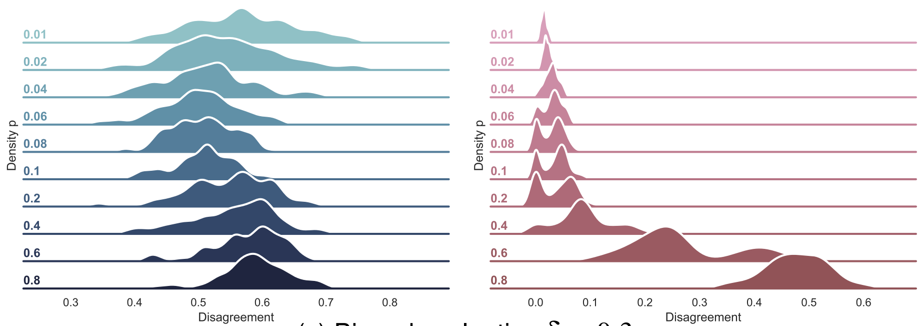
In order to understand the how network diversity impacts final disagreement between agents, and how this impact changes with the *seeking bias*, we look at the distribution of final disagreement for generated Erdös–Rényi random graphs of 100 nodes with different densities. Fig. 4.7 shows how disagreement varies with network density, considering both the case with no *biased seeking* $\rho = 0$ (in blue), and the case with *biased seeking* $\rho = 1$ (in red). Further, we also consider the cases where *biased assimilation* and *biased evaluation* are respectively introduced.



(a) No other biases at play



(b) Biased assimilation $\lambda = 0.02$



(c) Biased evaluation $\delta = 0.3$

Figure 4.7: Final disagreement after 1000 time steps, for varying density p in random networks, with *seeking bias* parameter $\rho = 0$ on the left (in blue), and with $\rho = 1$ on the right (in red). Three settings are considered: (a) where no other biases are introduced ($\delta = 0, \lambda = 0$), (b) with *biased assimilation* ($\delta = 0, \lambda = 0.02$), and (c) with *biased evaluation* ($\delta = 0.3, \lambda = 0$).

When no other biases are at play (Fig. 4.7a), we observe that final disagreement decreases as we increase the density of the network, trending towards total consensus for higher values of p. We find an anomaly in this trend for disagreement at $p = 0.01$

which decreases with respect to $p = 0.02$. This change in the trend is a product of the network becoming fragmented for low enough densities. We are surprised to find, that in this context, when introducing the rewiring for *biased seeking*, we see little effect on the resulting disagreement, and the same trend is preserved.

When considering the case with *assimilation bias* (Fig. 4.7b), we see a general increase in disagreement with regards to the unbiased case (by a factor of 10), but similarly, the introduction of rewiring has little effect with regards to final disagreement.

The more interesting result appears when we introduce *biased evaluation*. As shown in Fig. 4.7c, we see that the introduction of *biased evaluation* leads to disagreement that seems fairly invariant to network density without *biased seeking*. Nevertheless, with the introduction of *biased seeking*, final disagreement decreases considerably as the network becomes more sparse.

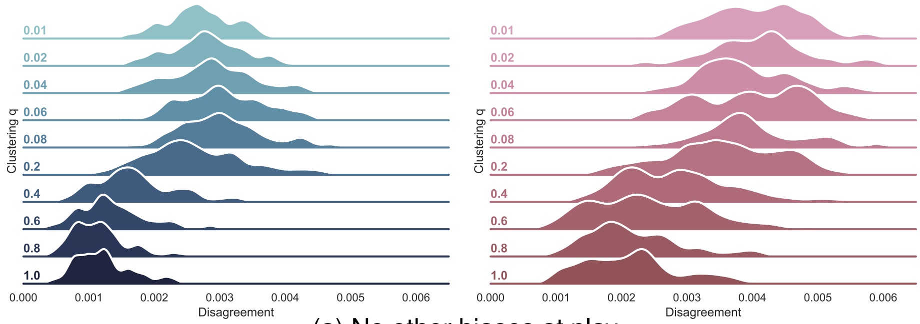
We conclude from this analysis that in order for the *seeking bias* to have a noticeable impact on disagreement in random networks, in must be introduced together with *biased evaluation*. In this case, *biased seeking* will lead to lower disagreement between agents. Further, the effects of the *seeking bias* are amplified as the diversity of the network decreases.

4.3.3 Clustering in Small-World Networks

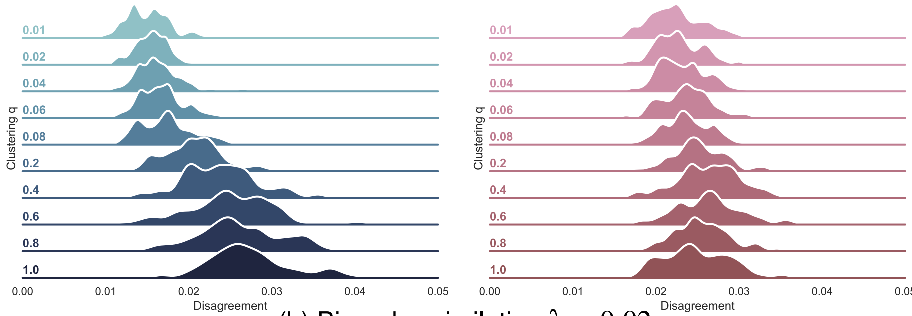
Purely random graphs as studied in the previous section, exhibit short path lengths, that is, any two nodes are a small number of connections away. The emergence of this phenomenon in real social networks is perhaps best know through the idea of six degrees of separation, which states that all people are at most six social connections away from each-other. This concept was first introduced by Frigyes Karinthy in a short story [49], and has lead to several empirical studies which confirm that many real networks are characterised by short path lengths [75, 78, 24]. Nevertheless, social connections are not made completely at random, instead social groups form representing clusters in the network of social interactions. Formally, a network exhibits clustering when two nodes are more likely to be connected if they have a common neighbor.

Small-World networks have both short path lengths and high clustering, arguably making them more representative of real social interactions. The Watts–Strogatz algorithm allows us to randomly generate such networks [79]. The algorithm arranges a set of nodes in a circle and connects each node to its k nearest neighbors ($k/2$ on each side). With the obtained network, the edges of each node are rewired with probability q. That is, the connection between that node and its neighbor is broken, and a new random connection is made. Here, k controls the density of the network, and q controls the level of clustering: for low q, agents are mostly connected to their nearest neighbors (high clustering), for higher q, clusters are broken by randomly rewiring connections (low clustering). As we have already studied the effects of density in the previous section, we now focus on clustering controlled by q. Thus, for the rest of thus section, we will set a fairly sparse density $k = 4$.

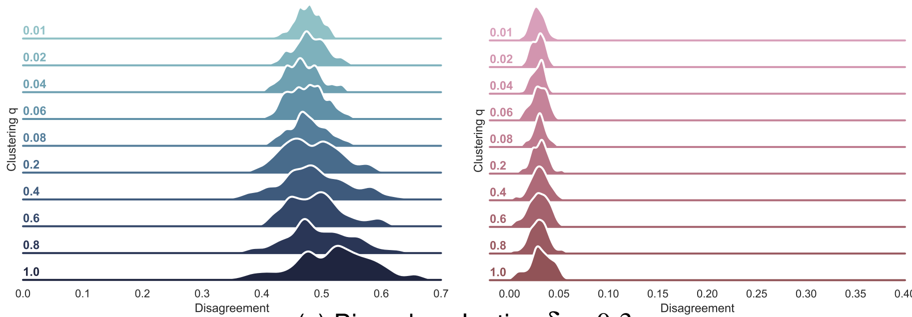
As for random graphs, we look at the distribution of final disagreement for generated Watts–Strogatz networks of 100 nodes, with varying clustering q. Fig. 4.8 shows how final disagreement varies with network clustering q with and without *seeking bias*, in the same three cases considered for random networks.



(a) No other biases at play



(b) Biased assimilation $\lambda = 0.02$



(c) Biased evaluation $\delta = 0.3$

Figure 4.8: Final disagreement after 1000 time steps, for varying clustering q in Watts–Strogatz networks, with *seeking bias* parameter $\rho = 0$ on the left (in blue), and with $\rho = 1$ on the right (in red). Three settings are considered: (a) where no other biases are introduced ($\delta = 0, \lambda = 0$), (b) with *biased assimilation* ($\delta = 0, \lambda = 0.02$), and (c) with *biased evaluation* ($\delta = 0.3, \lambda = 0$).

In the case where no other biases are involved (Fig. 4.8a) and without *biased seeking*, increasing clustering (decreasing q) up to $q = 0.08$ increases final disagreement. After that point, disagreement seems to slightly decrease for higher clustering. Note that with high clustering, agents in the same cluster are semi-isolated from the rest of agents, leading to low disagreement within clusters. Nevertheless, clustering also affects the ability of opinions to spread to the whole population, leading to higher disagreement between clusters. We interpret this trade-off to be captured by the observed trend in disagreement. With the introduction of *biased seeking*, we see an overall increase in disagreement, especially for higher clustering. By rewiring the connections, clustering decreases through time allowing opinions to spread more easily to the whole population. Nevertheless, as the results suggest, it doesn't decrease quickly enough for the overall disagreement to decrease in a way that compensates for the loss of clustering.

In the case with *biased assimilation* (Fig. 4.8b), the opposite pattern arises, where increasing clustering leads to a decrease in disagreement. As the introduction of *biased assimilation* makes agents opinions stabilize after a given number of time steps, even with low clustering, opinions won't spread for the whole population to reach a consensus. As such, low disagreement within clusters is more important in reducing disagreement, and it is obtained with high clustering. We see again that the introduction of rewiring leads to higher disagreement, with a more noticeable effect for high clustering.

Finally, in the case with *biased evaluation*, we observe a great decrease in disagreement with the introduction of *biased seeking*, but this effect seems to be indifferent to the level of clustering.

We here conclude that, surprisingly, *biased seeking* can also result in an increase in disagreement which is most noticeable for high clustering and in the presence of *biased assimilation*. Nevertheless, we find that again, the strongest effect of *biased seeking* is a reduction in disagreement which appears only in the presence of *biased evaluation*.

4.3.4 Degree heterogeneity in Scale-Free Networks

In the two networks we have studied so far, each node has a similar number of neighbors (Degree). Nevertheless, when it comes to social interactions and social influence, it is often the case that some people are more influential or connected than others. This is especially apparent in social media, where few people accumulate the vast majority of views and followers of the platform. A stronger claim is often made in social network analysis, which is that the degree of nodes in social networks follows a power-law distribution [59, 18]. Networks with this property are called Scale-Free, and the most common algorithm for generating such networks is the Barabási-Albert model [15].

In the Barabási-Albert algorithm, we start with a set of m_0 nodes, and we add new nodes to the network one at a time until the we reach the desired size. Each added node is connected to $m < m_0$ existing nodes, where the probability of the new node being connected to a node i is

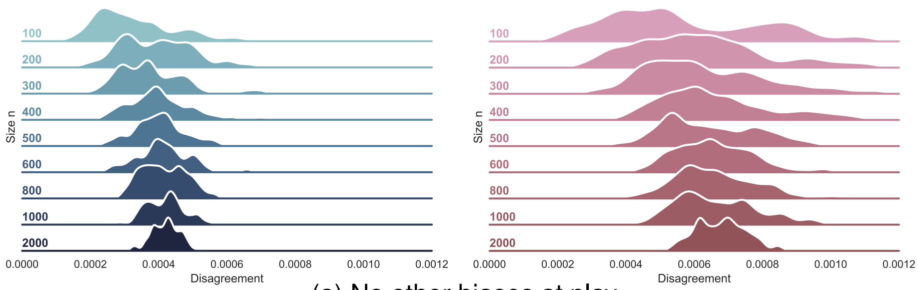
$$p_i = \frac{k_i}{K} \tag{4.1}$$

where k_i is the degree of i and K is the sum of the degrees of all nodes in the network. In this way, it is easy to see how nodes with more connections will accreete the majority of the connections as the network grows. The parameter m controls the density of the network. We will fix $m = 3$ as we have already covered the effects of network density. Instead, we will focus on the effect of the size of the network when generated using the Barabasi-Albert model.

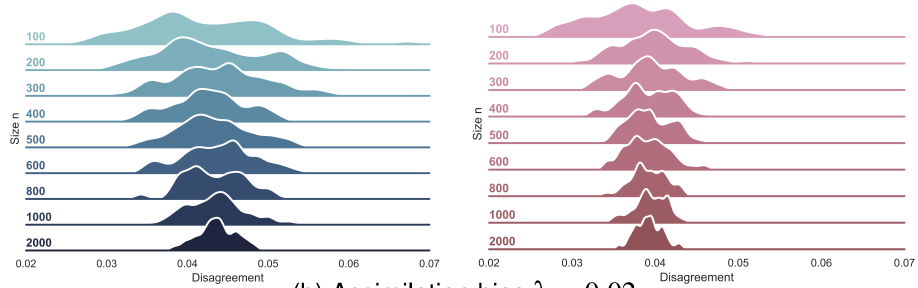
We look at the distribution of final disagreement for generated Barabasi-Albert networks of varying sizes. Fig. 4.9 shows how final disagreement varies with network network size, with and without *seeking bias*, in the same three cases considered previously.

When no other biases are introduced (Fig. 4.9a), it is worth noting that disagreement is much lower (by a factor of 10) than for random or small-world networks. Given that only a few agents accumulate the vast majority of social connections, many agents will simply be exposed to the opinions of the more influential agents. By repeatedly observing the same few agents, their opinions will quickly converge to match that of the more connected agents, leading to specially low disagreement. We observe that introducing *biased seeking* increases disagreement by a small amount, perhaps more noticeable for larger networks.

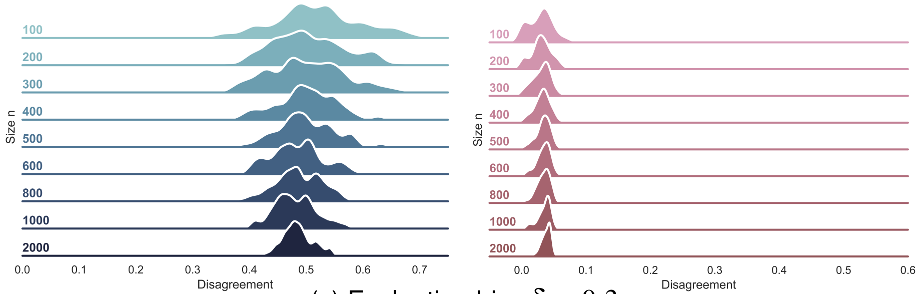
In the case with *biased assimilation* (Fig. 4.9b), introducing *biased seeking* leads to a slight decrease in disagreement, and in the case with *biased avaluation*, we again observe a strong decrease in disagreement with the introduction of *biased seeking*.



(a) No other biases at play



(b) Assimilation bias $\lambda = 0.02$



(c) Evaluation bias $\delta = 0.3$

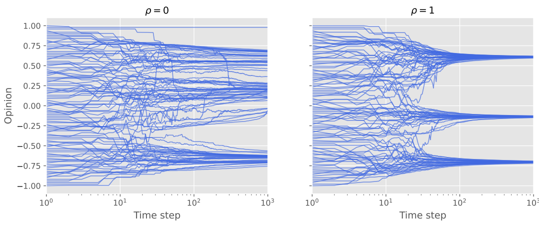
Figure 4.9: Disagreement for varying network sizes in scale-free network (a) with *seeking bias* parameter $\rho = 0$, and (b) with *seeking bias* parameter $\rho = 1$ with *assimilation bias* $\lambda = 0.02$. Plots on the left (in blue) show the case where there is no *seeking bias* ($\rho = 0$), plots on the right (in red) show the case where there is strong *seeking bias* ($\rho = 1$).

Although the effects found when no other biases are at play and with *biased assimilation* are not too significant, we see that for scale-free networks, the decrease in disagreement introduced by *biased seeking* is also present when *biased evaluation* is introduced.

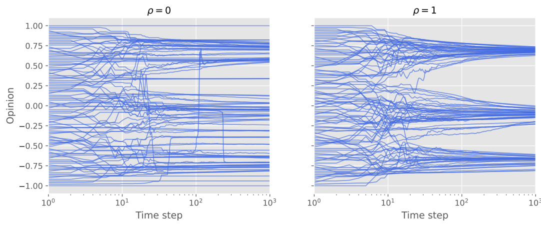
4.3.5 Qualitative implications of biased seeking

We have shown that the effects of *biased seeking* on disagreement between agents are very dependent on network structure and on the other biases present. Surprisingly, preferential attachment doesn't always lead to higher homophily, since it can increase disagreement as shown for Small-World and Scale-Free networks. Nonetheless, *biased seeking* has the highest impact when *biased evaluation* is also present. In this case, we see a great decrease in disagreement which becomes more pronounced in sparse networks.

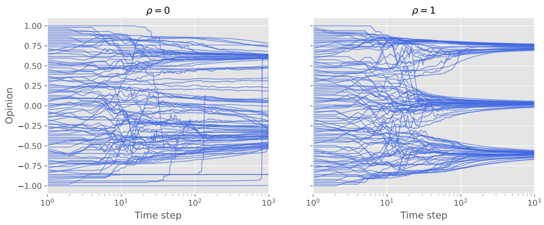
Other than understanding the impact of *biased seeking* on the metric of *disagreement*, to get a more complete picture of the evolution of the opinion space, we consider the evolution of agents mean opinions in the case with *biased evaluation* where *biased seeking* is most influential.



(a) Random Graph



(b) Small World



(c) Scale Free

Figure 4.10: Effect of *biased seeking* ρ on the evolution of agents mean beliefs in (a) Erdös–Rényi random graph ($p = 0.01$), (b) Watts–Strogatz network ($k = 4, q = 0.8$), and (c) Barabasi-Albert network ($m = 3$) with 100 agents ($\lambda = 0$, $\delta = 0.3$, $\sigma_0 = 0.1$). The distribution of agents opinions becomes more diffuse in complex networks, nevertheless, with the introduction of *biased seeking* clear opinion clusters emerge.

It is clear, in comparison with our simulations in fully connected networks, that embedding agents in a complex network structure leads to much more diverse opinion distributions (in contrast with the clear clusters emerging in fully connected networks). The structure of the network affects the ability of opinions to spread, leading to higher diversity and less clear cut clusters. The introduction of *biased seeking* $\rho = 1$, has the effect of better differentiating the emerging clusters, and reducing diversity within clusters. Note that in these networks, strong diversity emerges even without *biased assimilation*.

4.4 Extreme Agents and Polarisation

We have provided an overview of the effects of each bias, and their combined presence in fully connected networks and in more realistic and complex networks. We now focus on polarisation, which is perhaps the most relevant phenomenon in the process of opinion formation given its prevalence in the many political spheres. Given our model based on confirmation biases, we study which biased behaviour has a greater effect in generating polarised communities. We consider an opinion dynamics process to be polarising, if it leads to an increase in the diversity metric presented in the methods while generating two main clusters of opinion.

4.4.1 Communities and the Stochastic-Bloch model

To understand what conditions lead to the emergence of two polarised communities, a common choice for the underlying network structure is the Stochastic Block Model [3, 81, 22]. This model generates a network with communities defined by high connectivity within communities and low connectivity between communities. We construct two communities with 100 nodes each where the probability of creating an edge between two nodes is 0.2 if they are in the same community, and 0.05 if they are in different communities. This choice of parameters is derived from the study of polarising dynamics in Chen et al [22] which we will compare our results to (See Sec. 4.4.4).

4.4.2 A non polarising model for uniform uncertainties

If we consider agents prior uncertainties to be unrelated to their opinions, such as taking all agents to have the same prior uncertainties as we have done until now, the proposed model is not polarising under any combination of bias parameters. Although the introduction of the *evaluation bias* parameter can lead to two different clusters of opinion, the overall diversity is reduced from the initial random setting and as such we don't consider this process to be polarising.

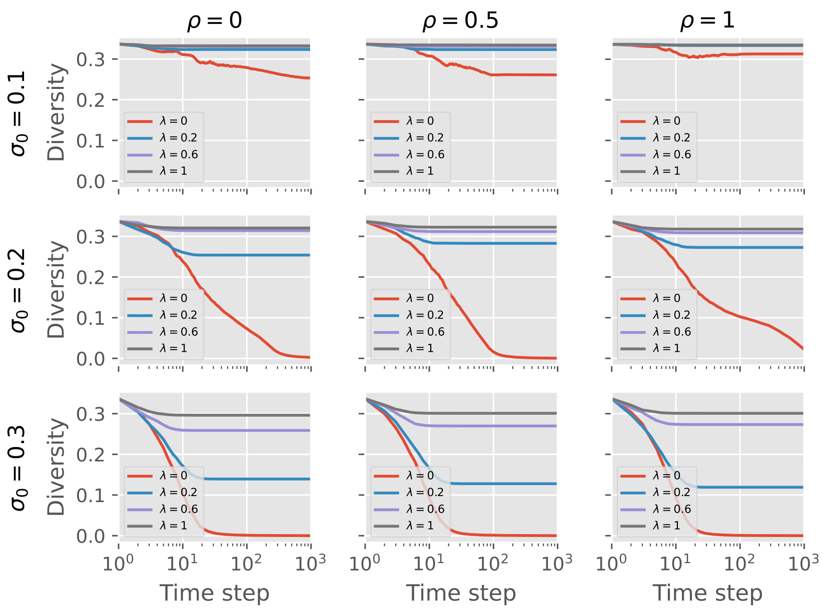


Figure 4.11: Evolution of diversity for simulations run with different combinations of the bias parameters in stochastic-block model with two communities of 100 agents each. No increase in diversity is observed for any combination of biases.

Fig. 4.11 shows the evolution of *Diversity* over time for different combinations of the three introduced biased behaviours. It is clear that although stronger biases prevent consensus, *Diversity* will always decrease, or at most remain constant.

4.4.3 Polarisation emerging from the introduction of extreme agents

A recurring theme in polarising models is that of extreme agents, derived from the observation that people with extreme opinions (modeled as -1 and 1) tend to be more reluctant to change their mind. Extreme agents are often defined as agents with an unchanging opinions of -1 or 1. Given that we are working with probabilistic opinions, we relax this assumption and simply take extreme agents to have a smaller prior uncertainty σ_{EX}. We denote η the number of extreme agents introduced for each extreme opinion. For $\eta = 5$, we would have 10 extreme agents (5 with prior mean of -1 and 5 with prior mean of 1). We set normal agents initial uncertainty to $\sigma_0 = 0.2$, and extreme agent's to $\sigma_{EX} = 0.02$.

We first consider the effect of each individual bias and look at which biased behaviour produces a potentially polarising outcome (an increase in diversity). Fig. 4.12 shows the evolution in disagreement and diversity over time for varying numbers of extreme agents η, and individually introducing each of the relevant biases. The main observation to be made, is that in the absence of *evaluation bias*, both *Diversity* and *Disagreement* decrease over time for any number of extreme agents leading to some form of consensus. On the other hand in the presence of *evaluation bias*, we observe that with the introduction of more than 10 extreme agents, the *Diversity* and *Disagreement* increase over time.

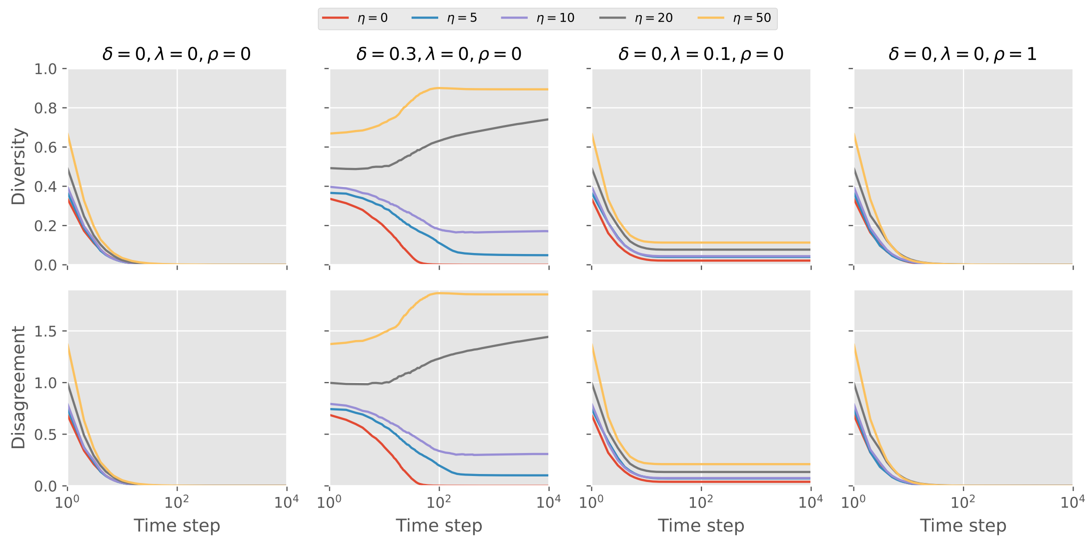


Figure 4.12: Evolution of Diversity and Disagreement with varying numbers of extreme agents, and different parameter settings for the three biases.

As we have shown that, in the proposed model, only *biased evaluation* leads to an increase in diversity, we take $\delta = 0.3$ and consider how the interactions with other biases affect the polarising dynamics. These interactions with other biases can be observed in Fig. 4.13. With the introduction of *biased assimilation* the polarising effects are preserved, but become more moderate. Introducing *seeking bias* accelerates the convergence time of overall diversity, but mainly has the effect of reducing disagreement to 0 where there is complete consensus within the polarised communities giving rise to echo chambers (communities where all agents hold the same opinions).

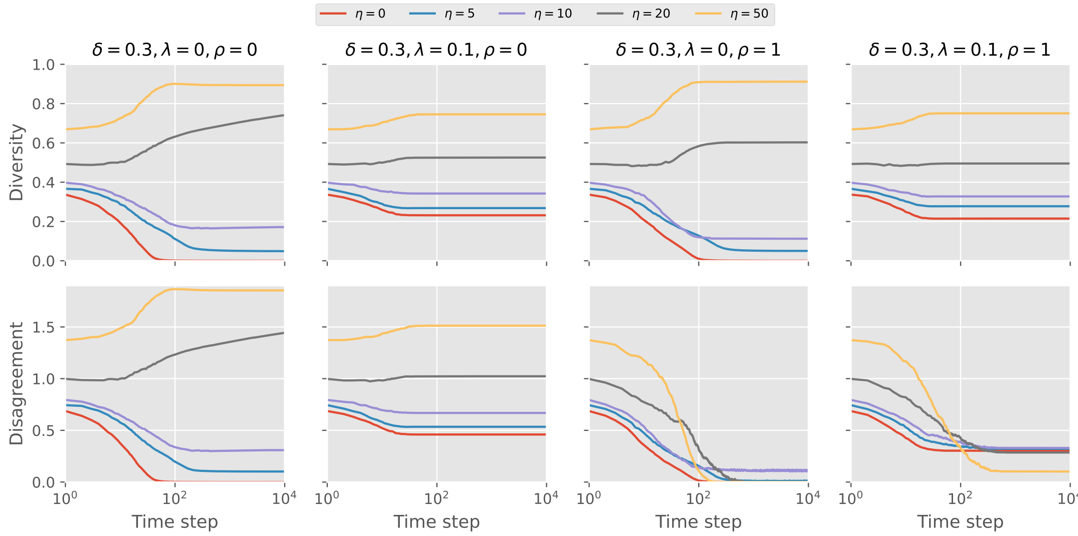


Figure 4.13: Evolution of Diversity and Disagreement with varying numbers of extreme agents, and different parameter settings for the three biases.

Finally, when all biases are introduced, we observe a similar pattern of increasing *Diversity* (polarisation) and decreasing *Disagreement*, but where disagreement doesn't converge to absolute consensus within groups, but instead stabilizes before that point, creating diverse echo chambers (communities where agents hold similar beliefs but with some level of disagreement).

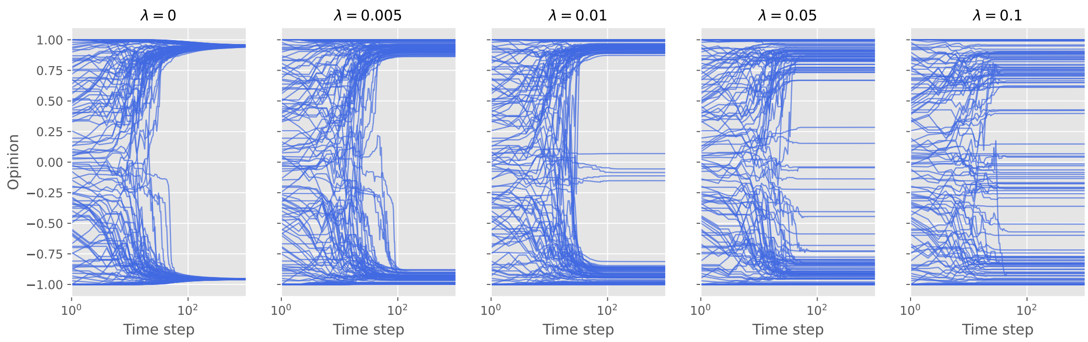
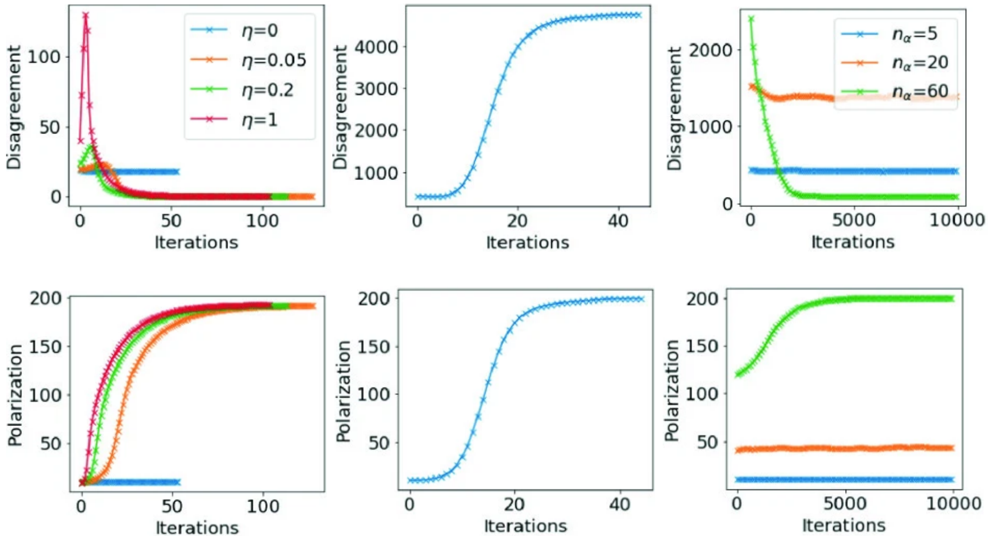


Figure 4.14: Evolution of agents beliefs with extreme agents ($\eta = 50$) and with all biased behaviours ($\delta = 0.3$, $\rho = 1$, and varying λ). Two polarised communities emerge, and disagreement within communities arises with the introduction of *biased assimilation*.

As we discussed, polarisation is described not only by the increase in diversity, but also by the fragmentation of opinions into two main clusters. Fig. 4.14 shows that the increase in diversity is also accompanied by opinions tending towards the extremes, verifying that the increase in diversity is in fact polarising. Further, we observe that increasing *biased assimilation* leads to higher disagreement within the polarised communities, tending towards dissensus for higher values of the parameter.

4.4.4 Other polarising models

To put our results in context, we draw from the work of Chen et al. [22], and compare our results to the commonly observed behaviours in other polarising models. Fig. 4.15 shows the results obtained by Chen et al. for three different polarising models FJCB, BOF and ECHO. These results are obtained using the same setting we have considered. That is, agents are embeded in a network generated through the stockastoc block model, with to comunities of 100 agents and probability of connection within community of 0.5 and between communities of 0.02.



(a) FJCB (b) BOF (c) ECHO

Figure 4.15: Evolution of Polarisation (Diversity) and Disagreement measures over time for three different polarising models: FJCB, BOF and ECHO. In FJCB social connections are altered and η controls the strength of these alterations. In ECHO, η_α controls the number of extreme agents. The observed polarising dynamics lead to high Polarisation (Diversity) and low disagreement,

Note that the *Disagreement* and *Polarisation* metrics shown in the figures, are unnormalised versions of the *Disagreement* and *Diversity* metrics we are working with. We observe two different polarising dynamics. For the FJCB and ECHO models, when the models are polarising, *Diversity* (*Polarisation*) increases, and *Disagreement* converges to 0. For BOF, we see that both *Diversity* and *Disagreement* increase, thus opinions become polarised and agents maintain strong disagreement with their neighbors. The novelty of our model is that it generates a third pattern, where we see an increase in *Diversity* (*Polarisation*), and a decrease in *Disagreement* which stabilizes instead of converging to 0, thus preserving some diversity within polarised groups.

4.5 Validation

Validation of opinion dynamics models is challenging and not commonly observed in the literature. This is in part due to the fact that opinions are complex and multidimensional, and as such it is hard to capture them through a numerical value. Further all possible outcomes of opinion formation are observed in real life within different contexts, making it hard to determine what outcomes a model should be able to predict.

In an attempt to provide some form of validation for our model, we draw from ideas introduced by Devia and Giordano [29] with regard to validating opinion dynamics model. In their work, they propose that since a range of behaviours is observed in the evolution of opinions: *consensus*, *dissensus*, *clustering*, and *polarisation*, the validity of a model should be assessed by the models ability to reproduce these states, and trasitions between different states. They go on to propose a framework for validation through histograms and survey data, which we won't apply in the scope of this work.

Instead, to asses the qualitative validity of the model, we look at its ability to generate *Consensus*, *Clustering*, *Dissensus* and *Polarisation* in real world networks. We use four different datasets of online social communities. Table 4.1 summarises the structural properties of these networks computed using networkx [39].

Network name	Edge meaning	N	E	$\langle k \rangle$	C
Advogato	Trust	6,539	40,548	12.4	0.18
Flickr	Similarity	105,938	2,316,948	43.74	0.089
Facebook	Friendship	4,039	88,234	43.69	0.606
Twitter	Friendship	63,731	1342296	33.02	0.565

Table 4.1: Structural properties of the studied networks. N is the number of nodes, E is the number of edges, $\langle k \rangle$ is the average degree, C is the clustering coefficient.

For each different network, we consider initially uniform opinions between -1 and 1, and we vary *biased assimilation* λ, *biased evalutaion* δ and the number of extreme agents in order to achieve the desired outcome. Note that we don't introduce *biased seeking* as we want to preserve the network structure obtained from the datasets. Fig. 4.16 shows the evolution of opinion densities for the four different networks and for different combinations of parameters. We see that our model is able to reproduce all outcomes, although polarisation appears to be harder to reproduce for larger networks (Flickr, Twitter). Further, the number of extreme agents introduced in order to achieve a polarising outcome is a fifth of the network size, which is a very high number of extreme agents, and perhaps unrealistic.

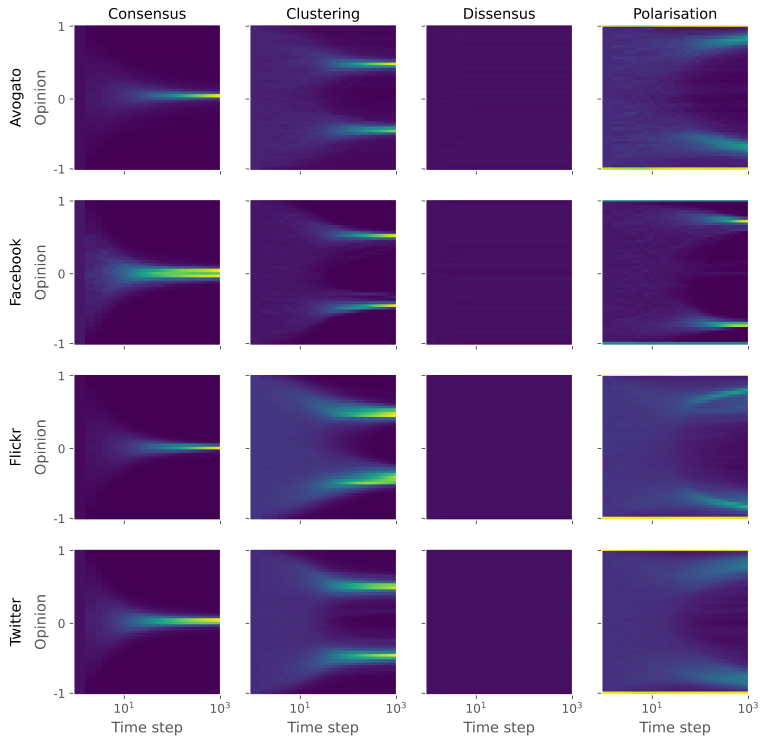


Figure 4.16: Opinion densities over time for 4 different social media networks replicating each qualitative outcome of opinion formation. Parameters are chosen as follows for each distributions: Consensus ($\lambda = 0,\ \delta = 0, \eta = 0$), Clustering ($\lambda = 0,\ \delta = 0.3, \eta = 0$), Dissensus ($\lambda = 1,\ \delta = 0.3, \eta = 0$), Polarisation ($\lambda = 0,\ \delta = 0.3, \eta = \frac{n}{10}$)

Chapter 5

Conclusions

5.1 Summary of findings

In this project, we propose a model of opinion dynamics based on confirmation biases. We draw from existing assumptions for confirmation bias in opinion dynamics, and categorise them into three behaviours: *biased seeking*, *biased evaluation*, and *biased assimilation* and *belief perseverance* which are modelled as one. Given the probabilistic and complex nature of some of these assumptions, we propose simplifications that allow us to run simulations in more complex settings such as large social networks.

We first validate the low level dynamics of the opinion update rule by considering the case where agents repeatedly make a single observation. We observe that *biased evaluation* makes agents unresponsive to evidence differing with their opinions until it is presented to them enough times. At that point, they will finally update their opinion to match the information presented to them. On the other hand with *biased assimilation*, agents are influenced by evidence presented to them from the beginning, but the influence of conflicting evidence diminishes, and agents beliefs are able to stabilize at a distance of the evidence being presented to them.

We then study the effects of *biased evaluation* and *biased assimilation* with agent interactions in fully connected networks, finding that *biased evaluation* leads to opinion clustering which increases with lower initial uncertainties for agents. The introduction of *biased assimilation* leads to the emergence of diverse final opinions. When both biases are present, clustering still emerges, but we observe diversity within clusters (strong diversity) which is a rare yet desirable result in strongly connected networks.

In order to consider the effects of *biased seeking*, social networks more complex than fully connected networks need to be considered. Thus, we study the effects of *biased seeking* in Random graphs, Small-World networks and Scale-Free networks. Interestingly, we find that preferential attachment (*biased seeking*) doesn't always lead to higher homophily. For instance, in Small-World networks, *biased seeking* increases final disagreement between agents especially when the networks have high initial clustering. We find that the intuitive effect of *biased seeking*, which is a decrease in disagreement between neighboring agents, only arises when *biased seeking* is also at play, and is more noticeable for sparse networks.

When studying the effects of biases on polarisation, we find that no combination of the introduced biases make the model polarising in itself. Instead, it is through the introduction of extreme agents that we observe the emergence of polarisation, but only when *biased evaluation* is present. Other than the fact that biased evaluation is at the core of polarising dynamics, we find that the introduction of *biased assimilation* and *biased seeking* leads to a novel polarising dynamic, where agents are polarised into two distinct communities, but some disagreement is preserved within communities.

Finally, we find that our model based on confirmation biases can reproduce four observed distributions of opinions: *consensus*, *clustering*, *dissensus* and *polarisation* in real networks from social media platforms. Nevertheless, the number of extreme agents required for the model to become polarising is very high and perhaps unrealistic.

5.2 Limitations

This project highlights interesting and desirable behaviours that can arise from extending probabilistic models of opinion dynamics, and by building on top of previous assumptions. Nevertheless, there are many limitations that arise from the models extended, and from the assumptions needed to combine them. Furthermore, although some qualitative validation of the model is proposed, this project suffers from a lack of rigorous validation with data.

5.2.1 Deviations from Bayesian principles

The first limitation of this project is briefly presented in Section 4.1.2, and it is the fact that the deviations from the Bayesian inference have undesired implications for the meaning of the variables involved. Further, note that many assumptions are made in order to extend the models to complex setting, such as: agents can observe other agents actual opinion, agents take other agents opinions to be an observation of the underlying variable, or agents take the reliability of their observations to be equal to their initial uncertainty. Many of these assumptions can easily be contested. Thus, as we extend the model to model complex setting and introduce more assumptions, it starts to resemble Non-Bayesian models more and more, where the update rule becomes more arbitrary and far removed from the baseline of rationality provided by Bayesian Inference.

5.2.2 Model complexity

Another notable limitation is that by introducing several different parameters for each confirmationally biased behaviour, we end up with a model that is very complex. Decisions have to be made about the values of the parameters $\lambda, \delta, \rho, \sigma_0, \eta$ on top of other things such as network structure. Thus, the wide range of behaviors displayed by the model could simply be a product of its complexity, providing less insight into human behavior.

5.2.3 Lack of validation

Finally, the model is only validated through its ability to reproduce a range of qualitative behaviours in real networks. A comparison of model predictions with data would be important in order to compare and improve the model. Although this issue is common to most of the filed of opinion dynamics, more and more attempts are being made for introducing ways of validating and analyzing opinion formation models.

5.3 Suggestions for future work

Bibliography

[1] Code and data. https://github.com/Martin-Prieto/Honours-Project.

[2] Real network datasets. http://konect.cc/networks/.

[3] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

[4] Daron Acemoglu, Victor Chernozhukov, and Muhamet Yildiz. Learning and disagreement in an uncertain world, 2006.

[5] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[6] Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1:3–49, 2011.

[7] Johnathan A Adams, Gentry White, and Robyn P Araujo. The role of mistrust in the modelling of opinion adoption. *Journal of Artificial Societies and Social Simulation (JASSS)*, 24(4):Article–number, 2021.

[8] Armen E Allahverdyan and Aram Galstyan. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557, 2014.

[9] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[10] Claudio Altafini. Consensus problems on networks with antagonistic interactions. *IEEE transactions on automatic control*, 58(4):935–946, 2012.

[11] Claudio Altafini. Dynamics of opinion forming in structurally balanced social networks. *PloS one*, 7(6):e38135, 2012.

[12] Brian DO Anderson and Mengbin Ye. Recent advances in the modelling and analysis of opinion dynamics on influence networks. *International Journal of Automation and Computing*, 16(2):129–149, 2019.

[13] Craig A Anderson, Mark R Lepper, and Lee Ross. Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of personality and social psychology*, 39(6):1037, 1980.

[14] Omid AskariSichani and Mahdi Jalili. Influence maximization of informed agents in social networks. *Applied Mathematics and Computation*, 254:229–239, 2015.

[15] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[16] Vincent D Blondel, Julien M Hendrickx, and John N Tsitsiklis. Continuous-time average-preserving opinion dynamics with opinion-dependent communications. *SIAM Journal on Control and Optimization*, 48(8):5214–5240, 2010.

[17] John G Bullock. Partisan bias and the bayesian ideal in the study of public opinion. *The Journal of Politics*, 71(3):1109–1124, 2009.

[18] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[19] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

[20] Francesca Ceragioli and Paolo Frasca. Continuous and discontinuous opinion dynamics with bounded confidence. *Nonlinear Analysis: Real World Applications*, 13(3):1239–1251, 2012.

[21] Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.

[22] Tianyi Chen, Xu Wang, and Charalampos E Tsourakakis. Polarizing opinion dynamics with confirmation bias. In *Social Informatics: 13th International Conference, SocInfo 2022, Glasgow, UK, October 19–21, 2022, Proceedings*, pages 144–158. Springer, 2022.

[23] Peter Clifford and Aidan Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[24] Eman Yasser Daraghmi and Shyan-Ming Yuan. We are so close, less than 4 degrees separating you and me! *Computers in Human Behavior*, 30:273–285, 2014.

[25] Guillaume Deffuant and Sylvie Huet. Collective increase of first impression bias. *Complexity*, 15(5):25–33, 2010.

[26] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

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

[28] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific reports*, 7(1):40391, 2017.

[29] Carlos Andres Devia and Giulia Giordano. A framework to analyze opinion formation models. *Scientific Reports*, 12(1):13441, 2022.

[30] Yucheng Dong, Min Zhan, Gang Kou, Zhaogang Ding, and Haiming Liang. A survey on the fusion process in opinion dynamics. *Information Fusion*, 43:57–65, 2018.

[31] Peter Duggins. A psychologically-motivated model of opinion change with applications to american politics. *arXiv preprint arXiv:1406.7770*, 2014.

[32] Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[33] Aili Fang, Kehua Yuan, Jinhua Geng, and Xinjiang Wei. Opinion dynamics with bayesian learning. *Complexity*, 2020:1–5, 2020.

[34] Marcos Fernandes. Confirmation bias in social networks. *arXiv preprint arXiv:2207.12594*, 2022.

[35] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[36] Serge Galam. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. *Journal of Mathematical Psychology*, 30(4):426–434, 1986.

[37] Alan Gerber and Donald Green. Misperceptions about perceptual bias. *Annual review of political science*, 2(1):189–210, 1999.

[38] Michel Grabisch and Agnieszka Rusinowska. A survey on nonstrategic models of opinion dynamics. *Games*, 11(4):65, 2020.

[39] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[40] Derek Hansen, Ben Shneiderman, and Marc A Smith. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.

[41] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[42] Itai Himelboim, Marc A Smith, Lee Rainie, Ben Shneiderman, and Camila Espina. Classifying twitter topic-networks using social network analysis. *Social media+ society*, 3(1):2056305117691545, 2017.

[43] Richard A Holley and Thomas M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.

[44] Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

[45] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[46] Robert A Jacobs and John K Kruschke. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011.

[47] Emily M Jin, Michelle Girvan, and Mark EJ Newman. Structure of growing social networks. *Physical review E*, 64(4):046132, 2001.

[48] Unchitta Kan, Michelle Feng, and Mason A Porter. An adaptive bounded-confidence model of opinion dynamics on networks. *Journal of Complex Networks*, 11(1):415–444, 2023.

[49] Frigyes Karinthy. Chain-links. *Everything is different*, pages 21–26, 1929.

[50] Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

[51] Ezra Klein. *Why we're polarized*. Simon and Schuster, 2020.

[52] Jonathan J Koehler. The influence of prior beliefs on scientific judgments of evidence quality. *Organizational behavior and human decision processes*, 56(1):28–55, 1993.

[53] U Krause. A discrete nonlinear and non-autonomous model of consensus. In *Communications in Difference Equations: Proceedings of the Fourth International Conference on Difference Equations*, page 227. CRC Press, 2000.

[54] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

[55] Charles G Lord, Lee Ross, and Mark 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.

[56] Yanbing Mao, Sadegh Bolouki, and Emrah Akyol. Spread of information with confirmation bias in cyber-social networks. *IEEE transactions on network science and engineering*, 7(2):688–700, 2018.

[57] André CR Martins. Bayesian updating rules in continuous opinion dynamics models. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(02):P02017, 2009.

[58] Jean-Denis Mathias, Sylvie Huet, and Guillaume Deffuant. Bounded confidence model with fixed uncertainties and extremists: The opinions can keep fluctuating indefinitely. *Journal of Artificial Societies and Social Simulation*, 19(1):6, 2016.

[59] Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3(1):1783, 2013.

[60] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*, pages 369–378, 2018.

[61] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[62] Ryosuke Nishi and Naoki Masuda. Collective opinion formation model under bayesian updating and confirmation bias. *Physical Review E*, 87(6):062123, 2013.

[63] Hossein Noorazar, Kevin R Vixie, Arghavan Talebanpour, and Yunfeng Hu. From classical to modern opinion dynamics. *International Journal of Modern Physics C*, 31(07):2050101, 2020.

[64] Matthew Rabin and Joel L Schrag. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics*, 114(1):37–82, 1999.

[65] Hegselmann Rainer and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. 2002.

[66] David P Redlawsk. Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *Journal of Politics*, 64(4):1021–1044, 2002.

[67] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402, 2021.

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

[69] Samuel Stern and Giacomo Livan. The impact of noise and topology on opinion dynamics in social networks. *Royal Society open science*, 8(4):201943, 2021.

[70] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. Inferring causal networks from observations and interventions. *Cognitive science*, 27(3):453–489, 2003.

[71] Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

[72] Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

[73] Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[74] Riitta Toivonen, Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, and Kimmo Kaski. A model for social networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):851–860, 2006.

[75] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social networks*, pages 179–197. Elsevier, 1977.

[76] Federico Vazquez. Opinion dynamics on coevolving networks. *Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems*, pages 89–107, 2013.

[77] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

[78] Duncan J Watts, Peter Sheridan Dodds, and Mark EJ Newman. Identity and search in social networks. *science*, 296(5571):1302–1305, 2002.

[79] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[80] Gérard Weisbuch, Guillaume Deffuant, Frédéric Amblard, and Jean-Pierre Nadal. Meet, discuss, and segregate! *Complexity*, 7(3):55–63, 2002.

[81] Yu Xing and Karl H Johansson. Transient behavior of gossip opinion dynamics with community structure. *arXiv preprint arXiv:2205.14784*, 2022.

[82] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation (TEAC)*, 1(4):1–30, 2013.

[83] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

Appendix A

Problems with arrays as distributions

A.1

Any appendices, including any required ethics information, should be included after the references.

Markers do not have to consider appendices. Make sure that your contributions are made clear in the main body of the dissertation (within the page limit).

Appendix B

Analytical solution for normal posterior

If you had human participants, include key information that they were given in an appendix, and point to it from the ethics declaration.

Appendix C

Evolution of beliefs while sampling from the distribution

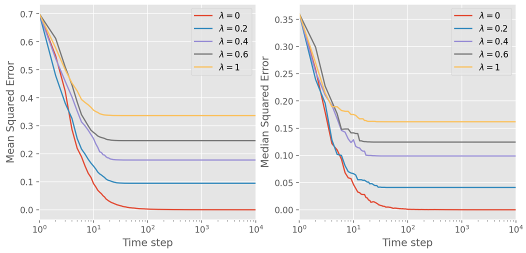
If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration. This information is often a copy of a consent form.

C.1 Interactions with truth related observations

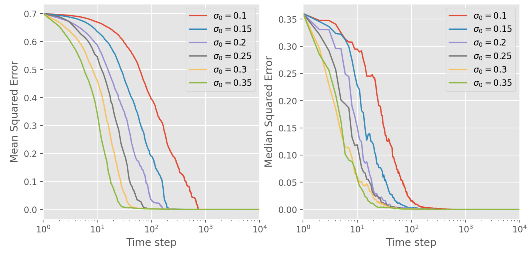
Within the Bayesian framework of social learning, agents take other agents opinions to be truth related. Thus, they assume agents are exposed to some signal related to the truth and their shared opinion reflects these observations. Until this point, we have only looked at agent interactions on their own without a source of information. The emerging dynamics are simply a result of the nature of agents priors and the update rule. This analysis is helpful to better understand the nature of the model and is widely used in the field. Nevertheless, we now look at the more realistic scenario where agents update their opinion based on both interactions with other agents and observations of some truth related signal. More specifically, we are interested in which component of confirmation bias has a greater effect on the ability of interacting agents to learn the truth from a noisy signal.

At each time step, an agent will interact with another agent with probability of 0.5 and make an observation of S otherwise. Although there are certainly differences between how people learn from others opinions and from observations (news, articles, experiences), we will take both processes to be biased. Thus, agents will use the biased update rule both when they are interacting with other agents and when they are making an observation of S. Given the biased update with respect to S, we also consider a scenario similar to the one studied in 4.1.1 where agents don't interact and only receive signals from S, but this time, they observe the signal with probability 0.5. We will use this scenario as a baseline, to determine whether the ability to learn S is impacted only by the biased observation of S or also by interacting with other agents in a biased manner.

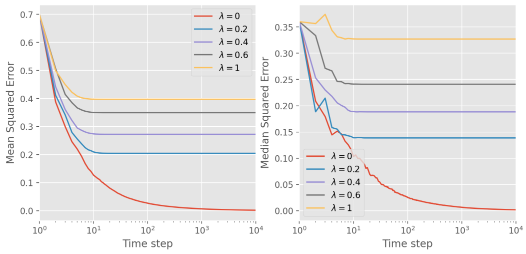
The ability of agents to learn the truth can be captured by error metrics with respect to the true source of information. We will use the mean squared error and the median squared error as they give different insights into the distribution of agents errors. Figure C.1 shows the evolution of these metrics for increasing values of the assimilation bais parameter λ, and the evaluation bias parameter δ respectively. Unsurprisingly, for both assimilation and evaluation, an increase in the bias leads to



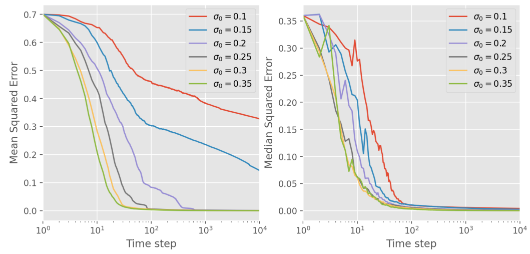
(a) $\delta = 0$, no agent interactions



(b) $\lambda = 0$, no agent interactions



(c) $\delta = 0$, with agent interactions



(d) $\lambda = 0$, with agent interactions

Figure C.1: Evolution of errors between agents opinions and information source with varying bias strengths, without (a, b) and with (c, d) agent interactions.

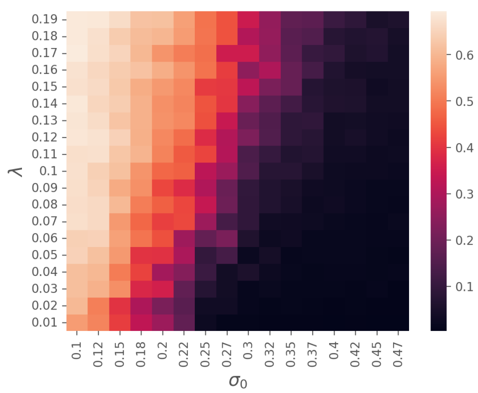


Figure C.2: Evolution of interacting agents mean beliefs on fully connected network, with varying assimilation bias $\delta = 0.3$

Appendix D

Effects of network structure on diversity