The echo chamber effect in social networks: theoretical analysis and steering strategies

Antoine Vendeville

The echo chamber effect in social networks: theoretical analysis and steering strategies

Antoine Vendeville

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

February 5, 2024

Supervisors

Benjamin Guedj

Shi Zhou

Examiners

Markus Brede

Hervé Borrion

University College London

Department of Computer Science

Centre for Doctoral Training in Cybersecurity

Centre for Artificial Intelligence

Viva passed on the 11th of October, 2023.

I, Antoine Vendeville, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

I present a mathematical framework to model echo chambers in social networks, and steer their impact. Digital communication is ever so pervasive in our societies, as the online and offline worlds become increasingly entangled. In online social platforms, similar-minded users tend to gather, and form strongly opinionated communities. These so-called *echo chambers* are particularly salient in polarised debates regarding politics, societal issues or conspiracy theories, and tend to foster animosity between opposite sides and fuel reinforcement of pre-existing beliefs. The political and informational landscapes are significantly affected, turning the regulation of echo chambers into a crucial matter of cybersecurity. This calls for a more informed analysis and understanding of this phenomenon.

Whether an echo chamber is actually desirable or not is context-specific: my framework is agnostic and able to accommodate both. In the last few years, echo chambers have become a primary focus of research on opinion dynamics and information diffusion, with a plethora of models amenable to shed light on empirical studies. There is however a lack of principled methods to efficiently steer the echo chamber effect.

In this PhD thesis I develop two mathematical models to describe, quantify and control the echo chamber effect: a macroscopic one based on group-level dynamics, and a microscopic one incorporating user-level features. For each model, I present algorithms which significantly impact the diversity of content users are exposed to, while accounting for individual preferences to avoid backfire effects. The accuracy of the models and the effectiveness of the recommendation algorithms are illustrated through applications on real-world data. This PhD thesis contributes insights to the

benefit of the growing debate on regulation of online social platforms.

Impact Statement

The primary impact of this PhD thesis is the development of mathematical models to describe the evolution of opinions in social networks and online social platforms. Doing so, I contribute to the existing academic literature and advance the scientific knowledge on the hidden laws that govern human interactions, in particular the evolution of opinions in our societies.

The regulation of online social platforms has arisen as a major challenge in the last few years. The spread of conspiracy theories or the January 6^{th} riots in the US are examples of nefarious consequences of these platforms. Therefore, the search for adequate policies to address such issues is of tremendous importance. The secondary impact of this PhD thesis is outside academia, as it provides governments and policymakers with tools to help towards this goal. I do not advocate for specific uses of my methods, but provide examples as proofs of concept.

This research has been assessed by peers and its results published in leading journals and conferences, including *Applied Network Science* (Vendeville et al., 2021), *Complex Networks and their Applications* (Vendeville et al., 2023a, 2022c) and *Conference on Complex Systems* (Vendeville et al., 2022a,b). Other results have been submitted to *Physical Review E* and are currently under peer review (Vendeville et al., 2023b).

Acknowledgements

First, I would like to address a huge thanks to my supervisors, Benjamin Guedj and Shi Zhou, for all their help during these 4 years. You both have always supported me, my work, and encouraged me to believe in my ideas. You have known how to subtly balance between close mentoring when I felt lost, and a liberty of research that helped my creativity thrive at other times.

Benjamin, I don't know how I would have been able to write publishable papers without your thorough proof-reading and correcting. Thank you also for staying up late doing the last corrections on a submission at two in the morning in order to meet the deadline. Your advanced LaTeX skills have been most valuable in these times of need. Shi, you have relentlessly pushed me to reflect on the bigger picture, the meaning and the objectives of my work, before diving into equations and little details as I tend to do. I struggled a lot with that, and your help has been invaluable to overcome this weakness of mine.

Anastasios Giovanidis, you are the reason I undertook this journey in academia. Doing beautiful research has always been your highest priority. Your passion has infected me and I do not see it leaving anytime soon. I would like to take this occasion to thank other researchers for their precious insight and guidance over the years: Jeffrey Howard, Giacomo Livan, Agnieszka Rusinowska, Jean-Philippe Cointet.

I also want to acknowledge the support of my fellow PhD students, from UCL and many other universities in the UK and abroad: Effrosyni Papanastasiou, Maxime Haddouche, Antonin Schrab, Antoine Picard, Fernando Diaz, Arianna Trozze, Antonis Papasavva, Guillermo Moreno and many others. Some of these

encounters have deeply affected me and my research. Many of you have become not only collaborators, but also flatmates, and even close friends.

Lastly, all of that would not have been possible without the continuous support of my partner and my family. You have always been there, in good and bad times, to help me get through this journey. This manuscript is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The advent of Online Social Platforms (OSPs) in the last decade has irremediably changed our societies by allowing us to communicate on an unprecedented scale. This has not come without drawbacks, and concerns are rising about potentially nefarious consequences (Haidt, 2022). Amidst the COVID-19 pandemic, online misinformation related to the disease has spread like wildfire. It is a never-ending stream of far-fetched theories, with millions of voices claiming that the virus is a hoax, is transmitted by 5G towers or that Bill Gates created the vaccine to implant chip in people's bodies (Cuthbertson, 2020; OSoMe, 2020; Zarocostas, 2020). A US engineer even purposely derailed a train over Coronavirus conspiracy theories concerns (Spocchia, 2020). Although difficult to quantify exactly, the large spread of such theories probably had a non-negligible impact on compliance with health safety policies and thus with the virus diffusion. All of this is especially concerning when we know that 68% of adults ever get news on Social Media and 42% take them for largely accurate, at least in the US (Shearer and Matsa, 2018).

OSPs have also significantly impacted democratic processes around the world. In the infamous Cambridge Analytica case, the Brexit and the Trump campaigning teams used illegally harvested personal data in order to increase their vote share via strategic targeting of potential "swing" voters (Kaiser, 2019). In the wake of the 2020 United States presidential elections, the 4chan born and bred QAnon conspiracy

theory has gained such support in the United States (OSoMe, 2020; Sabin, 2020) that Twitter, Facebook and Youtube ended up banning from the platform major groups of users that were spreading it (Collins and Zadrozny, 2020a,b,c). After the results of the election, online groups have been continuously relaying and amplifying claims of election fraud (Beckett and Wong, 2020), to the point where the social platform Parler is currently under investigation for hosting the organisation of the Capitol riot on the 6th of January 2021 (The Guardian, 2021).

These phenomenons may have been facilitated by the so-called *echo chambers*, that have sparked a growing interest in the scientific community. Echo chambers are clusters of like-minded users, that foster a continuous reinforcement of prior beliefs and strongly reject opposing ideas. In turn, their presence hinders democratic debate and provides a fertile breeding ground for extremism and conspiracy theories. Occurrences of this phenomenon have been observed in online discussions surrounding various political and controversial topics (Cinelli et al., 2021; Kirdemir and Agarwal, 2022; Williams et al., 2015), although the amount of users they impact and the extent to which they do so is still up to debate (De Francisci Morales et al., 2021; Dubois and Blank, 2018).

Multiple factors have been advanced as explanations for the existence of echo chambers, such as homophily, confirmation bias, negativity or information overload (Bronner, 2021; Chavalarias, 2022; Hills, 2019; McPherson et al., 2001a). These natural biases are exacerbated by the personalisation algorithms used by the platforms: to sort through the enormous mass of information constantly flowing online, OSPs carefully filter, select and rank only the most relevant content to show to their users, to maximise their engagement. These algorithms tend to hide under the rug anything that supports different views, leading to the entrapment of users into their own personalised *filter bubble* (Pariser, 2011), which perpetually echoes their pre-established opinions.

While they were originally mostly free from legal constraints, the responsibility of OSPs is being increasingly discussed. Governmental bodies have started to step in, envisioning regulatory policies to thwart these nefarious phenomenons. The

European Union has recently decided to take unprecedented action in this regard, with the Digital Service Act (EU, 2022). Amongst other measures, this legislation will force OSPs to make public their personalisation algorithms and allow users to opt out of them. The United Kingdom is also discussing its own Online Safety Bill (UK, 2023). Twitter has already made public their newsfeed algorithm, but it is difficult to fully understand without access to the underlying data that it feeds onto. Even more so, as they have recently restricted access to their API (Weatherbed, 2023).

Major platforms have started taking action against misinformation and the communities that spread it (Bickert, 2020; Blake, 2021; Collins and Zadrozny, 2020a,b,c; Culliford and Paul, 2020). Although effective on the short term, these efforts feel like cutting the head of a Hydra, which will eventually grow back as long as the core body is alive. Indeed the ostracised users can always find new platforms and online groups to welcome them, as we have seen with QAnon conspiracists massively joining the platform Voat after Reddit banned many of their subreddits in 2018 (Monti et al., 2023; Papasavva et al., 2021). Debunking false information is also rarely effective at best (Chan and Albarracín, 2023), and counter-productive at worst (Betsch and Sachse, 2013; Zollo et al., 2017).

Despite the best efforts to improve, it seems like the very way OSPs are built, coupled with human biases, will always entail these nefarious side effects (Bronner, 2021; Chavalarias, 2022; Pariser, 2011; Thaler and Sunstein, 2009; Vosoughi et al., 2018). To support better informed policies regarding the regulation of social platforms, and to gain a deeper understanding of how we communicate as humans, I argue that any long-term attempt to fight these phenomenons shall start with a thorough study of the roots that anchor them into the ground. This is why I am interested in how opinions evolve in social networks, as I now detail my objectives.

1.2 Research objectives

This PhD thesis is two-pronged: it connects fundamental research on opinion dynamics, and the urgent need for an adequate regulation of OSPs. As I extend and improve

pre-existing theoretical models of opinion dynamics, I contribute significantly to the former. Moreover, I demonstrate the real-world applicability of my framework in dealing with the phenomenon of echo chambers. Doing so, I lay fundamental stones to the benefit of the latter, in establishing general leads towards a healthily regulated online environment.

Objective 1. Describe through mathematical models the echo chamber effect and the impact of recommender systems in OSPs. Our first objective is theoretical: to propose novel models of opinion dynamics, particularly adapted to the description and measurement of echo chambers. Recently empowered by the access to data from OSPs—*e.g.* Peralta et al. (2022), research on opinion dynamics is concerned with the study of mathematical frameworks to describe the evolution of opinions, beliefs, views or political leanings in a population. It is particularly adapted to gain a theoretical understanding of the aforementioned phenomenons that take place on OSPs.

For the sake of simplicity, I wish to limit my parameter space to core, essential mechanisms. Thus, my models shall account for peer influence, inner biases, and the presence of personalisation algorithms. Thus, they will provide an accessible, simplified way to evaluate their impact. Formal models are always an approximation of reality, and they will never be perfectly accurate Fernandez-Gracia et al. (2014). However, they can give us valuable insight on the mechanisms at play, and help us build a more precise view of the highly complex world that surrounds us.

Objective 2. Develop efficient algorithmic methods to steer the echo chamber effect via content recommendation. Second, and as an illustration of the effectiveness of the models I developed to pursue the first objective, I want to propose novel algorithms to find optimal content recommendation methods to steer the echo chamber effect. I will solve optimisation problems to maximise the diversity of content that users of an OSP are exposed to, hence opening up the bubbles of congenial information. Importantly, I account for potential backfire effects, where exposure to uncongenial information can have the adverse consequence of reinforcing pre-existing beliefs.

Our framework is agnostic and can accommodate different goals regarding the control of echo chambers, and more broadly the control of opinion and information diffusion online. How and why should the platforms be regulated is a complex issue that I do not delve into in this manuscript. Neither is addressed the questions of whether or not it is desirable that people have access to wide variety of opinions and information. I merely provide **tools** that can be used for a wide variety of goals pertaining to the control of opinion and information diffusion online. I neither encourage nor discourage any specific goal that these tools may be wielded to achieve. However, as a proof of concept, I demonstrate the ability of my methods to decrease the echo chamber effect by increasing the diversity of opinions online users OSP have access to.

1.3 Research contributions

Our contribution is multi-fold. It is mainly theoretical, and the models I develop are useful in a variety of contexts. Their generalisability is the first and foremost strength of my work. The application to the problem of steering the echo chamber effect is thought as a study case to demonstrate the value of the methodology.

Contribution 1. Generalisation of the Voter Model with analysis of discord probabilities. I propose the Enhanced Voter Model (EV Model), that generalises the Voter Model for opinion dynamics to directed, weighted networks with exogenous influence (zealots), any finite number of possible opinions, and individual update rates. I demonstrate how to compute probabilities of discord between agents, and extend the traditional definition of active links density to account for long-range, weighted interactions. I explore the equilibrium states of the model on synthetic networks divided into antagonistic communities, and uncover a rich landscape of varied behaviours. The advantage of the EV Model is its generality, as it is not limited to the analysis of OSP: uses of the model range from offline social networks (Fernandez-Gracia et al., 2014) to particle interactions (Clifford and Sudbury, 1973; Holley and Liggett, 1975).

Contribution 2. Extension of the Newsfeed Model to describe opinion flow. The Newsfeed Model was recently introduced to describe the flow of content throughout an OSP. I propose the Extended Newsfeed Model (EN Model) to incorporate the notion of opinions in the Newsfeed Model. This lets me quantify how political views spread throughout the network. Importantly, I am able to compute the distribution of opinions that users are exposed to on the platforms. I show how the model can be further improved by performing simulations with preferential reposting behaviour. The EN Model is specifically designed to describe OSPs, and this specialisation is its advantage: it allows for an easy modelling of real-life features of OSPs (newsfeeds, walls, likes, mentions, etc.).

Contribution 3. Introduction of metrics to measure the echo chamber effect and the diversity of content on OSPs. I propose to measure the echo chamber effect by the proportion of congruent opinions that users of an OSP are exposed to. I develop formulas to compute this proportion in the EV Model and the EN Model. I also demonstrate formulas to calculate the diversity of opinions that users are exposed to in the EV Model and the EN Model. The two metrics are closely related, and in some cases equivalent. I calculate these values and analyse the results on synthetic networks (EV Model) and a Twitter dataset (EN Model).

Contribution 4. Empirical evaluation of the models. I compare the theoretical equilibrium states predicted by the EV Model and the EN Model to real-world data. The effectiveness of the EV Model is evaluated through its ability to forecast the outcome of democratic elections in the US and the UK. The effectiveness of the EN Model is evaluated through its ability to predict the distribution of content that users are exposed to, in a dataset from Twitter.

Contribution 5. Introduction of methods to compute optimal recommendation rates to steer the echo chamber effect. I show how the EN Model and the EV Model can be used to find optimal content recommendation rates, that maximise the diversity of content that users are exposed to in an OSP. Doing so, I am able to effectively steer the echo chamber effect. I adopt both a macroscopical and a microscopical point of view, via respectively the EV Model and the EN Model.

I account for possible backfire effects, as I assume that users are susceptible to reinforce their pre-existing beliefs when exposed to incongruent opinions.

1.4 Thesis outline

Chapter 2 lays the foundations for my research. I define online social platforms, and echo chambers within. I also introduce the basics of scientific research on social networks, specifically graph theory and opinion dynamics. In Chapter 3 I present recent advances on the theoretical and empirical empirical analyses of echo chambers, the control of opinions in social networks, and the mitigation of opinion polarisation and echo chamber effects. The method of my research is presented in Chapter 4, where I introduce my models, highlight the dual macroscopical-microscopical approach, and present the real-life datasets I use in my applications. Chapter 5 and Chapter 6 are devoted to the mathematical analysis of the EV Model and the EN Model, respectively. The methods for steering the echo chamber effect and the obtained results are found in Chapter 7. Finally, a general discussion of my results, as well as limitations and avenues for future research, are provided in Chapter 8.

All programs were coded with Python. Tables of notations are available in Appendix A.

Chapter 2

Background

In this chapter, I introduce the main objects of my study. I define the notions of social networks and online social platforms. I briefly summarise the history of the latter, as well as their main characteristics. I introduce the concept of echo chambers, and the underlying mechanisms that are thought to foster the phenomenon. Finally, I introduce the bases of the mathematical study of social networks and opinion dynamics. Traditional models and fundamental results that stemmed the field are presented.

2.1 Social networks and online social platforms

According to the Oxford English Dictionary[1], a social network is either *(i) a network of social interactions and personal relationships*, or *(ii) a dedicated website or other application which enables users to communicate with each other by posting information, comments, messages, images, etc*. While the first meaning dates back to the origin of mankind, the second is way more recent. Both will be used throughout this manuscript, and to avoid any confusion I reserve the term "social network" to refer to the first one. The second will be designated by "Online Social Platform", or OSP for short. As an OSP often relies on an underlying social network, the two terms can often become entangled.

[1] https://www.oed.com/

2.1.1 History

The first mainstream OSP is commonly agreed to be Six Degrees, launched in 1997. The platform reached a few million users at its peak, but its popularity was limited by the scarcity of internet access at the time (Ngak, 2011). The name of the platform is a reference to a well-known, and somewhat surprising, fact about social networks—I come back to it later. With the growing use of internet, multiple different websites have waxed and waned over the years. Nowadays, there exists a plethora of very popular OSPs with each their own particularities. Some focus on short text communication (Twitter, Sina Weibo), others on pictures (Instagram, Pinterest), others on video (TikTok, YouTube), etc. The most popular ones total billions of monthly active users. The top spot is thrusted by Facebook, boasting nearly 3 billion monthly active users in 2023[2].

2.1.2 Structure and functionalities

Most OSPs share a common basis of structural features and functionalities. The fundamental utility of these platforms is to provide a way for people to exchange online with others. Users are able to post content—*e.g.* messages, pictures, videos, links to external websites—and engage with content posted by others: *liking* a post to signal approval, *reposting* to spread the piece of content further, *commenting* to discuss, express an emotion or opinion. To received updates about content posted by another, one must *follow* them. Doing so, they become a *follower* of the other user, who becomes their *leader*. The *newsfeed* (equivalently, *timeline*) of a user contains content created or propagated by their leaders. There exists a variety of policies about how a user is allowed to create a connection with another: Twitter for example allows anyone to instantly connect with anyone else, while Facebook requires mutual consent. As I will mostly work with data from Twitter (now *X*), I will often refer to the action of creating a new post as *tweeting*, and the reaction of reposting as *retweeting*.

A significant characteristics of OSPs is their ability to personalise the user experience. Namely, the newsfeeds are usually not a simple chronological list of

[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

content. Rather, platforms have developed advanced *personalisation algorithms* to propose the most relevant content possible to the users at any time. What items are presented or hidden, in which order do they appear, in what form exactly: all of this and many more details are carefully curated to ensure that users are most likely to engage with the content they are presented with. In addition, the newsfeeds often contain advertisement, and recommended content: items not necessarily created by the leaders of the user, but that they may be interested in. Most platforms also recommend users to one another ("You may know this person...").

2.1.3 The phenomenon of echo chambers

There exists a concordance of social, psychological and algorithmic mechanisms that has fostered the emergence of hermetic clusters of users in OSPs: the so-called *echo chambers*. This phenomenon has attracted more and more attention in recent years.

An echo chamber is a segregated community of like-minded users, wherein shared prior beliefs are continuously reinforced while opposite opinions are vigorously rejected. Content spread within almost exclusively fits a single narrative, and any opposing voice is promptly swept away. As members of such a group become more and more entrenched in their beliefs, users of different mindsets slowly drift further and further away from one another. This entails the *polarisation* of opinions, and is especially concerning given that OSPs are becoming primary platforms for political discussion. Despite a lack of consensus on a formal characterisation of echo chambers, groups of users showing such behaviour have been repeatedly observed in a variety of OSPs, covering a wide range of topics—see amongst others, Bakshy et al. (2015); Grömping (2014); Hosseinmardi et al. (2020); Phadke et al. (2020); Williams et al. (2015).

Phenomena underlying the formation of echo chambers are not precisely known, and several mechanisms have been proposed as potential enablers throughout the years. First and foremost, human beings have always had a tendency to homophily, that is associating with others of similar views and beliefs (McPherson et al., 2001b). It is thus natural that online, we prefer to communicate with like-minded others and avoid those of different opinion. This effect is further reinforced by what Pariser

(2011) coined as the *filter bubble* in his celebrated book. OSPs usually perform algorithmic filtering in order to suggest the most bespoke content to its users and hide what is less relevant. They offer to each of their members a personalised list of curated content often in the form of their newsfeed. Because as humans we can only spend a finite amount of time online, it seems natural to resort to automatic methods that can quickly filter for us the gigantic amount of new information being continuously produced. And because we are naturally more likely to click on congenial content (Bakshy et al., 2015; Nikolov et al., 2015), it makes sense to filter out cross-cutting posts. However by selecting content that fits our profile the best, these recommendation algorithms tend to hide under the rug anything that support different views, thus promoting confirmation bias (Lord et al., 1979). Hence they have the adverse consequence of reinforcing prior beliefs and inhibit the evolution of ideas (Dandekar et al., 2013; Spohr, 2017).

As put by Bogost (2021), *there is no reason to believe that everyone should have immediate and constant access to everyone else in the world at all times*. The possibility to reach anyone in the world, coupled with the immediate availability of massive amount of information might have turned out to be a bane rather than a boon (Bogost, 2021; Hills, 2019). Under the constant-raging storm of new content, psychological factors affect our absorption of information in a subjectively biased way, for example favouring more alarming and extreme news or skewing their meaning to fit our pre-existing views (Bronner, 2021; Chavalarias, 2022; Liao and Fu, 2013). The context surrounding a tweet as well as the intentions of its poster are often unknown, while they are essential to fully grasp the meaning of the conveyed message.

Propagation of information also tends to alter its meaning, often amplifying perceived risk and negativity (Moussaïd et al., 2015). As radical rhetoric employed by extremist parties often relies on threatening information to draw in more voters, it was observed that extremism increases with online engagement (Bessi et al., 2016; Hosseinmardi et al., 2020; Vaccari et al., 2016; Wojcieszak, 2010). Consequently, the most active people online are often the most radical (Arugute et al., 2023; Weber

et al., 2020), creating a vicious circle where a rise in radicalism in inherent from the ever-growing use of OSPs. Extremist parties have indeed shown the biggest increase in online activity in recent years (Hong, 2013; Hosseinmardi et al., 2020). They have adopted the web as a platform of expression quicker than moderate people have, and occupy a proportionally bigger share of online than offline space (Hong and Kim, 2016; Takikawa and Nagayoshi, 2017). Moreover, they also often act as breeding pools for misinformation and conspiracist content, which tends to propagate more easily and more widely than verified news (Osho et al., 2020; Zhang et al., 2020).

So how does this phenomenon persist? Because confirmation bias affects our interpretation of content, neutral pieces of information are easily understood as a support for pre-existing beliefs (Nickerson, 1998). A very telling example of this was given by Lord et al. (1979), in an experiment where participants against or in favour of death penalty became more convinced of their position after all reading the same essay on the matter. The unilateral rejection of opposite views makes it even harder to make people reevaluate their opinion. Even more so than rejection, exposure to opposite views often entails the so-called *backfire effect* of reinforcing the prior opinion (Betsch and Sachse, 2013; Nyhan and Reifler, 2010). Thorough scrutiny should then be applied to content exposing polarised users to cross-cutting views in order not to exacerbate the situation even further (Cook and Lewandowsky, 2012; Schaewitz and Krämer, 2020).

2.2 The mathematics of social networks

OSPs are closely related to social networks. The study of the latter has not waited for the former to thrive, and I now present the basic mathematical framework scholars rely upon: graph theory.

2.2.1 Basic definitions and properties

A convenient way of representing a social network is via a *graph*. Graphs have been studied since the XVIII$^{\text{th}}$ century and have a wide variety of applications, from protein interactions to internet topology and routing systems (Barabasi, 2016; Newman, 2010). A graph is a set of *nodes* linked to one another by the means of

edges. A graph is said to be *complete* if each node is connected to all others. In our context, the nodes are users and edges represent follower-leader relationships. I will use without distinction the words node, user or agent, and edge, link, or connection.

A graph is said to be undirected if all of its edges are reciprocal, otherwise it is called directed. If one wants to attribute different levels of importance to different edges, it is convenient to assign a *weight* to each of them. In a network of acquaintances for example, the edge between my best friend and I might carry a heavier weight than the one linking me with a distant relative. The absence of connection is signified by a weight of zero. Weights are usually positive real numbers, but there is a whole theory dedicated to the study of *signed* graphs that incorporate negative connections, usually to represent enmity—see the seminal papers of Harary (1953); Heider (1946). For the purpose of my work I restrain myself to positive weights, but negative ones can be a powerful tool for the study of social networks. In the literature survey I mention promising research based on this notion.

Given a graph $\mathcal{G}$, let $\mathcal{N} = \{1, \dots, N\}$ denote the set of its nodes. N is the total number of nodes. $\mathcal{E}$ will be the set of edges and E their number. The density of a graph is the ratio of E over the total possible number of edges, which is $N(N-1)/2$ for undirected graphs and $N(N-1)$ for directed ones. The structure of a graph is encoded in an *adjacency matrix* that I denote by W. Its $(i,j)^{\text{th}}$ coordinate w_{ij} is the weight of the edge $j \to i$. The adjacency matrix of an undirected graph is symmetric. In an unweighted graph, w_{ij} is either 0 (absence of edge) or 1 (presence of an edge).

Let $\mathcal{L}_i = \{j \in \mathcal{N} : w_{ij} > 0\}$ denote the set of leader of user i. In an undirected graph, leaders and followers of i are the same, and often referred to as *neighbours* of i. The degree (*resp.* in-degree, out-degree) of a node is its number of neighbours (*resp.* leaders, followers). The distribution of degrees in a graph gives important insight on its topology.

2.2.2 Communities

A path from user j to user i is a succession of nodes where each one is a leader of the next. Hence, it gives a way to reach i starting from j. If there exists a path from j to i and one from i to k, then there is a path from j to k. In an undirected

graph, if there is a path from j to i then there is a path from i to j. Thus, we can group together nodes that are able to reach each other. Doing so, we obtain the *connected components* of the graph. An undirected graph is said to be connected if it has only one connected component, meaning every node can reach every other. For directed graphs, a *strongly connected component* contains nodes that can reach each other both ways, and a *weakly connected component* contains nodes that can reach each other if we discard the directionality of links—*i.e.* considering the graph as undirected.

The notion of *community* is central to the study of social networks. Although no precise definition is universally admitted, it designates a group of users with many connections amongst them, and fewer connections with the outside. A complete, isolated component of a graph is a perfect community. In practice however, the topology is rarely that extreme. There are many ways to quantify intermediate levels of community structure, the most widely used being perhaps the notion of *modularity* (Newman, 2006). Given knowledge of the graph topology and community memberships, the modularity is calculated by comparing the number of edges within communities, with their expected number were all edges of the graph rewired at random. Formally, the modularity of an undirected, unweighted graph $\mathcal{G}$ with two communities and adjacency matrix A is defined as follows.

$$Q = \frac{1}{4m} \sum_{i,j} A_{ij} - \frac{d_i d_j}{2m} \varepsilon_{ij}. \tag{2.1}$$

Here, $\varepsilon_{ij} = 1$ if i and j are in the same community, and -1 otherwise. m is the total number of edges in the graph, and d_i the degree of node i. Thus, the term $A_{ij} - d_i d_j / 2m$ measures the difference between the effective presence of an edge, and the probability of its existence were all edges rewired at random. When Q is high, there are both *(i)* more edges within communities, and *(ii)* less edges between communities, than expected by chance.

Another way of measuring community structure is the *clustering coefficient*. It quantifies how much nodes in a network tend to form triangles, which is social networks for example means that “the friend of my friend is my friend”. The local

clustering coefficient at node i is defined by

$$C_i = \frac{2e_i}{d_i(d_i-1)}, \tag{2.2}$$

where e_i is the number of edges that exist between neighbours of i. Hence, C_i is the ratio of the number of edges between neighbours of i, and their maximum possible number. If two of them j,k are also connected, the summand is 1 and C_i increases. This means that the three nodes form a triangle. The average local clustering coefficient is then computed as $\sum_i C_i/N$. There exists another definition of the clustering coefficient at the global level, often referred to as *transitivity ratio*:

$$C = \frac{\sum_{i,j,k} A_{ij}A_{ik}A_{jk}}{\sum_i d_i(d_i-1)}. \tag{2.3}$$

In that case, C measures the proportion of triangles that exist in the graph. While closely related, C and $\sum_i C_i/N$ can give close but also widely different results depending on the considered graph. I refer the interested reader to Schank and Wagner (2005) for an in-depth analysis and comparison of the two definitions.

It has been found that the network of follow-leader connections is not necessarily the best way to describe relationships in OSPs (Huberman et al., 2009; Leskovec and Horvitz, 2008). Indeed users tend to follow a lot of people they never interact with, and it might be more realistic to consider an interaction network, where relations are based on actual communication. I call *retweet network* a graph where there is an edge from j to i if i has reposted content from j, and *follow network* a graph where there is an edge from j to i if i follows j. I will also evoke *mention networks*, with an edge from j to i if i has mentioned j in a post—done via the @ symbol on Twitter or Instagram, for example.

2.2.3 Graph models for social networks

Many types of graphs have interested researchers over the years. In the second half of the XX$^{\text{th}}$ century, a particular one has started to spark interest in the scientific community. Driven by the seminal and famous work of Erdös and Rényi (1959), graphs with random topology have become ubiquitous. An undirected Erdös-Rényi

(ER) random graph with N nodes is built as follows: given N isolated nodes and a link probability $p \in [0,1]$, connect each pair of nodes with probability p. The graph is almost surely connected as soon as $p > \ln N/N$. ER graphs also have the advantage to exhibit the *small-world* property: the path between any two nodes is rather short, mirroring empirical findings in social networks. This was observed as soon as in 1967 in a real-life experiment conducted by Stanley Milgram (Milgram, 1967). He observed that the median number of hops needed to go from any American person to any other is six, which was lower than expected. This property has been frequently verified in OSPs (Kurka et al., 2016).

Another typical property of social networks is *clustering*: the friend of my friend is likely to be my friend (Kurka et al., 2016; Mislove et al., 2007). The Watts-Strogatz (WS) model considers a ring of nodes, each connected to its k nearest neighbours. Then each link is rewired randomly with probability p. For $p = 1$, we recover the ER graph. For low p, this process generates graphs with high a number of triangles, meaning that two friends of a same third person are often connected with each other.

The ER and WS model generate graphs with binomial degree distributions, which for social networks does not corresponds to empirical observations. Rather, we often observe a power-law distribution: the probability of having degree k is proportional to $k^{-\alpha}$ with $\alpha > 1$, meaning that most nodes have a small number connections, and a select few have them in large numbers (Huberman et al., 2009; Mislove et al., 2007). Those are celebrities, political figures, media outlets... This *scale-free* property is reproduced by the Barabasi-Albert (BA) model: starting with a small number of connected nodes, add new nodes one at the time. Each new node connects to a fixed number m of others, the probability of connection to i being proportional to the degree of i. This "rich-get-richer" process accentuates inequalities in degrees, resulting in the aforementioned power-law distribution.

None of the ER, WS or BA graphs exhibit significant community structure. A remedy is to use the Stochastic Block Model (SBM): attribute a community membership to each node, and chose a connection probability p_{ab} for any two

communities a, b (not necessarily distinct). Then, draw an edge between any two nodes from communities (a, b) with probability p_{ab}. By tuning the probabilities, we can obtain more or less significant community structures. However, this also creates networks with binomial degree distributions. Extensions that generate power-law distributed SBM graphs have been studied (Karrer and Newman, 2011; Qiao et al., 2019).

2.3 Opinion dynamics

The graph theory laid above is a simple yet powerful tool to describe the structure of social networks. I have, however, not said anything about the dynamics of interactions between users. Throughout the years, many models have been built to try and explain the dynamics of opinions on social networks. Perhaps the earliest well-known works in this area are from French (1956) and DeGroot (1974), who studied how a society of individuals may or may not come to agreement on some given topic. Assuming people repeatedly update their belief by taking weighted averages of those of their acquaintances, they showed that if the society is not divided in isolated components, *consensus* is reached. That is, everyone will eventually agree—provided the process runs for a sufficiently long time. This is a fundamental concept in the study of opinion dynamics, although consensus may not happen that often in practice. We shall see that others have built on the works of French (1956) and DeGroot (1974) to study the fragmentation, divergence, and polarisation of opinions.

2.3.1 General concepts

Most models of opinion dynamics share a common basis of features and principles. Consider a social network of N agents, each with opinions $x_1, \ldots, x_N$ in some *opinion space* $\mathcal{S}$. For a binary divide between two stances on a topic, *e.g.* Labour versus Conservative, take $\mathcal{S} = \{0, 1\}$ or $\{-1, +1\}$. For a continuous, more nuanced spectrum of leanings, it is often $\mathcal{S} = [0, 1]$ or $[-1, +1]$. Opinions may also be vectors, either to represent multi-dimensional opinions, or to describe a probability distribution over multiple possible views on a single subject. The state of the system at time t is

described by

$$x(t) = (x_1(t), \ldots, x_N(t)) \in \mathcal{S}^N. \tag{2.4}$$

Unless stated otherwise, I assume edge weights to be normalised, so that $\sum_{j \in \mathcal{N}} w_{ij} = 1$ for all $i \in \mathcal{N}$. The matrix W is then said to be row-stochastic. I will denote the $N \times N$ identity matrix, with ones on the diagonal and zeros everywhere else. I assume the user graph to be static, but we shall see that some works incorporate dynamical changes of connections.

Opinions are bound to evolve due to multiple factors. Most models choose an update rule, through which agents repeatedly re-evaluate their opinions. Updates may be continuous or discrete in time, and synchronous (everyone re-evaluates their beliefs simultaneously) or asynchronous. Almost all of the models have in common one feature as part of their update rule: influence of peers. If most of my acquaintances hold a certain opinion, it is more likely that I adopt it as well. Other features influencing opinions have been studied over the years, such as inner biases, exogenous influences, or external shocks—*i.e.* real-world events that may dramatically affect the opinion landscape. Remark that recommender systems present in OSPs can be seen as a form of exogenous influence. This is a point of view I will adopt in my methodology.

Zealots An important concept here is that of *zealots*. I call zealot, any one-sided source of influence. This includes inner biases, some external sources of influence, and stubborn agents who never change opinion—that may be politicians, lobbyists, or journalists. These are widely different concepts, however for modeling purposes they are described the same way. Indeed, zealots impact the system without being impacted by it: their influence is exogenous to the system. Masuda (2015) and Moreno et al. (2021) think of zealots as pinning controllers, tunable quantities through which some objective function (in their case, the average opinion of agents) can be optimised. The notion of zealots is thus amenable to model many different phenomenons. I call s-zealot a zealot defending opinion $s \in \mathcal{S}$. If a user is subject to an inner bias towards opinion s, I say that i is influenced by the s-zealot.

The study of opinion dynamics is mainly concerned with the long-term proper-

ties of the system. Whether or not the system eventually reaches an *equilibrium*, or *steady state*, is a central question. I say that the system is in equilibrium after $t > 0$ if the distribution of $x(t')$ is the same for all $t' > t$.

2.3.2 The French-DeGroot model

The seminal works of French (1956) and DeGroot (1974) (FD) make no hypothesis about $\mathcal{S}$ other than being a vector space. For DeGroot (1974), x_i is a distribution over all possible value of some unobserved parameter θ. The process described in Equation 2.5 corresponds to users pooling their knowledge to try and discover the true value of θ. Incidentally, this idea is at the root of social learning models—see Section 2.3.5. Opinions are repeatedly updated via

$$x(t+1) = Wx(t). \tag{2.5}$$

This is often referred to as *linear consensus dynamics*. According to Eq. 2.5, each user changes their opinion by taking a weighted average of others' opinions—possibly including their own. The value w_{ij} quantifies the influence of user j on user i. If $w_{ij} = 0$ then i does not take j's opinion into account, if $w_{ij} = 1$ then j's opinion is the only that matters to i. Diagonal entries measure the self-confidence of the users. The larger w_{ii}, the lower w_{ij} for $j \neq i$ and thus the lower influence from others on i. If $w_{ii} = 1$ then i puts no weight on others' beliefs and will always keep the same opinion.

We have the following fundamental convergence result.

Theorem 1 (DeGroot, 1974)**.** *The dynamics described by Eq. 2.5 are those of a Markov chain with transition matrix W^T. Thus, the system converges if and only if 1 is a right eigenvalue of W. Any eigenvector of W corresponding to eigenvalue 1 is then an equilibrium solution. There is a unique solution if and only if W is aperiodic. Consensus is reached if and only if there exists a user who can reach every other.*

Confer Section 4.3.1 in the appendix for a brief introduction to Markov chains. The requirement of a "super-user" who can reach every other as a necessary condition for consensus is recurrent in opinion dynamics. This is the case in particular when

de network is connected, *i.e.* each agent can reach every other. If the condition of aperiodicity is violated, we find ourselves in a situation where individual opinions are cyclically repeated but never settle to a particular value. For example, consider a directed graph connecting two users $\{1,2\}$ with initial opinions $x(0) = (x_1, x_2)$ and connected by edges with weights $w_{1,2} = w_{2,1} = 1, w_{1,1} = w_{2,2} = 0$. In this setting, users 1 and 2 simply switch opinions at each step, so that $x(t)$ indefinitely oscillates between (x_1, x_2) and (x_2, x_1).

An example realisation leading to consensus is shown in Figure 2.1. I generated an ER network (*cf.* Section 2.2.3) with $N = 1,000$ agents and edge probability $p = 0.008$. Because $\ln N/N \simeq 0.007$, the generated ER network is connected with probability one. I simulated the DeGroot dynamics, starting from random uniform values in $[0,1]$. We observe convergence to a consensus, which is explained by the fact that the network is connected.

In the case where there is convergence but no consensus, there exists pairs of users whose opinions are completely independent. Consensus might be reached in each of the strongly connected components of the graph taken individually, but nothing guarantees that limiting opinions will be the same across all of them.

2.3.3 The Friedkin-Johnsen model

In their seminal paper, Friedkin and Johnsen (1990) (FJ) extend the FD paradigm to introduce inner biases. Let y be a vector of size N that lies in the same space as x and whose i-th entry quantifies the inner bias of user i. Let $\beta \in [0,1]$ be a tunable parameter. Opinions are updated via

$$x(t+1) = \beta W x(t) + (1-\beta) y. \tag{2.6}$$

Hence β quantifies the susceptibility of users, that is the relative importance of opinion pooling in the presence of innate preferences. If $\beta = 1$ then those are unaccounted for and Equation 2.6 reduces to the FD update (2.5). If $\beta = 0$ on the other hand, everyone is stubborn and opinions are set in stone, with values y.

We have the following fundamental convergence result.

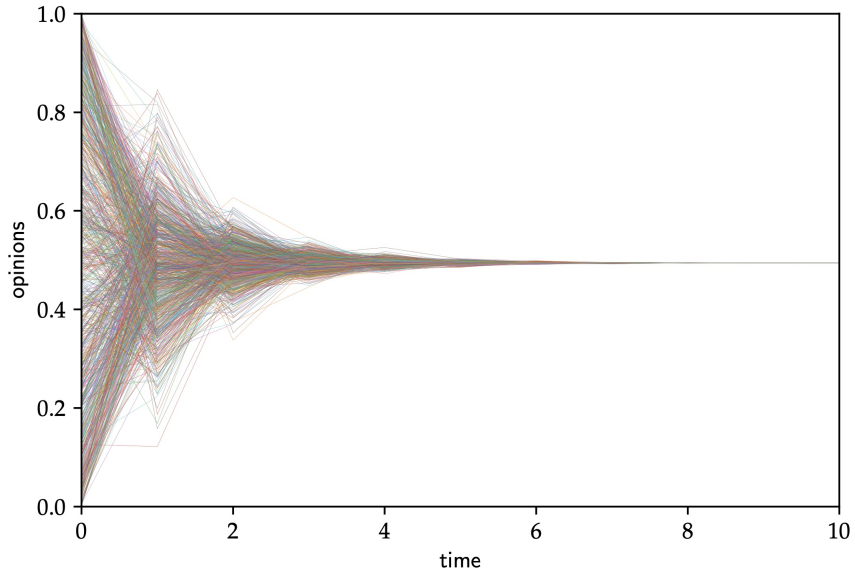

Figure 2.1: Example realisation of the FD model with convergence to consensus. Each line represents the evolution of the opinion of a single user. The social network was generated under the Erdös-Rényi model with $N = 1,000$ nodes and connection probability $p = 0.008$. The opinion space is $\mathcal{S} = [0,1]$. Initial opinions were drawn uniformly at random in $\mathcal{S}$.

Theorem 2 (Friedkin and Johnsen, 1990)**.** *If β is not an eigenvalue of W then the dynamics described by Eq. 2.6 converge and we have*

$$x(t) \underset{t\to\infty}{\longrightarrow} (1-\beta)(I-\beta W)^{-1}y. \tag{2.7}$$

Thus, each user ends up settling on one constant opinion, but there is not necessarily consensus as this limiting opinion can vary from one user to another. It is calculated as an average of everyone's inner biases, weighted by the matrix $(1-\beta)(I-\beta W)^{-1}$. The condition on β assures that $(I-\beta W)$ is invertible. Note that it is not a necessary condition, as in the FD model $\beta = 1$ is an eigenvalue of W.

Everyone might not necessarily attach the same importance to their innate preferences versus pooling the group's opinions. Parsegov et al. (2017) generalise the FJ model by replacing β with a $N \times N$ diagonal matrix B with all entries smaller than 1. The update equation is then given by

$$x(t+1) = BWx(t) + (I-B)y. \tag{2.8}$$

Thus user i puts weight $B_{ij}w_{ij}$ on the opinion of user j and $1-B_{ii}$ on their innate preferences. The authors prove the following convergence result.

Theorem 3 (Parsegov et al., 2017)**.** *If W is irreducible and $B \neq I$ then the model is convergent and we have*

$$x(t) \underset{t\to\infty}{\longrightarrow} (I-BW)^{-1}(I-B)y. \tag{2.9}$$

The hypothesis on B guarantees the invertibility of $(I-BW)$. Once again, note that this theorem gives a sufficient condition for convergence that is not necessary—in the FD paradigm $B = I$ and yet opinions converge in irreducible, aperiodic settings. The more general following theorem provides a sufficient and necessary condition for convergence that accounts for such cases.

Theorem 4 (Parsegov et al., 2017)**.** *The model is convergent if and only if*

$$U^\star := \lim_{k\to\infty} (BW)^k \tag{2.10}$$

exists, and in that case we have

$$x(t) \underset{t\to\infty}{\longrightarrow} U^\star y + \sum_{k=0}^{\infty} (BW)^k (I-B)y. \tag{2.11}$$

2.3.4 Bounded confidence

Early works on opinion dynamics were particularly concerned with the question of *consensus*, *i.e.* whether or not the population would eventually agree. With the advent of opinion polarisation and the emergence of echo chambers online, the focus of research has shifted towards models able to explain these phenomena rather than produce the rarely-observed consensus. Perhaps an important milestone in this direction is the apparition of the so-called "bounded confidence" models, that incorporate homophily in their dynamics. The fundamental idea is that users only interact with others holding opinions close enough to theirs.

In their seminal paper, Deffuant et al. (2000) propose the following process: at each time step, two agents are selected uniformly at random. If they are too far apart

ideologically, constructed debate cannot take place and their opinions stay the same. But if they are sufficiently like-minded, they influence each other so that the gap between their opinions tightens. Formally, say agents i and j are randomly selected at time $t+1$. Then if $|x_i(t) - x_j(t)| < \varepsilon$, their opinions are updated via

$$\begin{aligned} x_i(t+1) &= x_i(t) + \mu(x_j(t) - x_i(t)) \\ x_j(t+1) &= x_j(t) + \mu(x_i(t) - x_j(t)). \end{aligned} \tag{2.12}$$

If $|x_i(t) - x_j(t)| \geq \varepsilon$ however, nothing happens. The model parameters ε and μ respectively quantify the strength of the homophily effect and the susceptibility of agents to other's opinions. In the case $\mu = 1/2$, both new opinions are the same and lie in the exact middle of the two previous open.

The model is analytically intractable in most cases. Numerical simulations reveal the possibility of both consensus and clustering of opinions depending on the parameters. It is believed that $1/2\varepsilon$ is a good approximation for the limiting number of clusters, so that consensus is reached when $\varepsilon > \varepsilon_c = 1/2$ (Castellano et al., 2009).

Hegselmann and Krause (2002) incorporate ideas from both the Deffuant and the FD model. At each time step, every agent updates their opinions to the average amongst all their like-minded neighbours:

$$x_i(t+1) = \frac{\sum_{j,|x_i(t)-x_j(t)|<\varepsilon} w_{ij} x_j(t)}{\sum_{j,|x_i(t)-x_j(t)|<\varepsilon} w_{ij}} \tag{2.13}$$

where w_{ij} is the weight of edge $j \to i$. Here the long-term behaviour depends not only on ε but also on the average degree of the underlying graph model. If the average degree is constant in the limit $N \to \infty$ (*e.g.* Erdös-Rényi random graph) then the threshold for the emergence of clusters is $\varepsilon_c = 1/2$; otherwise if the average degree diverges as $N \to \infty$ (*e.g.* complete graph) then $\varepsilon_c \approx 0.2$ (Castellano et al., 2009).

These works were perhaps inspired by the celebrated paper of Axelrod (1997), who studies a multi-dimensional, discretised counterpart of Deffuant et al. (2000). Users hold opinions on F different features, with q possible traits, or values, for

each. Think of the features as different society issues—*e.g.* immigration, climate change—and the opinions as the various positions one may hold on each issue—*e.g.* pro or against, sceptic or not. At each time step, a user i and one of its leaders j are selected at random. With probability proportional to the number of features on which they agree, i adopts the opinion of j on a random feature on which they disagree.

Depending on q and F, the system may either end up in consensus or in a configuration where different stable clusters coexist, within each all users share the same trait on every feature. Most of the literature devoted to the model studied it on regular lattices. Castellano et al. (2000) found that the system exhibits a phase transition at $F = 2$, so that:

- the relative size of the largest cluster depends only on q for $F = 2$, but also on the size of the lattice for $F > 2$;
- the distribution of the size of clusters follows a power-law with exponent ≈ 1.6 for $F = 2$, and a power-law with exponent ≈ 2.6 for any $F > 2$.

The relative size of the largest cluster goes to one as q goes to zero, meaning the smaller the number of traits, the closer to consensus the system gets. For $(F,q) = (15,10)$ for example, consensus is reached (Axelrod, 1997).

2.3.5 Social learning

Models of social learning introduce the idea of rationality in the behaviour of agents by proposing Bayesian updates of opinions (Acemoglu et al., 2011; Banerjee and Fudenberg, 2004; Bikhchandani et al., 1992). Consider a group of agents trying to guess the true value $\theta^\star$ of some underlying *state of the world* θ. For example, they might be trying to decide if vaccination is safe or not. We assume they proceed one at a time, each making a single guess. To inform their decision, they are aware of previous guesses made by their peers, and in addition they receive an independent private signal from the external world. These signals may represent any outside information independent of the social network, such as newspaper articles. Their distribution is conditioned by θ and known to the users. Agents then form their

guesses via Bayesian inference based on this signal and previous guesses made by others.

In term, the population may or may not correctly learn the true value of θ. The interested reader may refer to Acemoglu et al. (2011) for a very complete description and analysis of such a model. I mention the following convergence result.

Theorem 5 (Acemoglu et al., 2011)**.** *Consider an infinite network of agents. Under mild conditions on the distribution of signals, guesses almost surely converge to* $\theta^\star$ *when* $\bigcup_{i\in\mathbb{N}} \mathcal{L}_i$ *is of infinite size.*

The condition on the union of leaders sets prevents the presence of a finite set of agents upon whom everyone else relies to make their guess. Such agents can easily hinder convergence to the truth when promoting a wrong value of θ.

An interesting variant on this idea is from Jadbabaie (2012): agents repeatedly receive independent signals, and update their beliefs by combining a Bayesian update based on the latest signal, with a FD-style pooling of their leaders' beliefs. Instead of taking once-in-a-lifetime guesses, all users shape their beliefs over time. The belief of agent i at time t is described by a private probability distribution $\mu_{i,t}$. At each time step, every agent i receives a new, independent private signal $s_{i,t+1}$. That signal is generated by a time-independent, conditional distribution $\ell_i(\cdot|\theta)$ known by user i—and only them. Beliefs are then updated via

$$\mu_{i,t+1} = w_{ii}\mathrm{BU}(\mu_{i,t}|s_{i,t+1}) + \sum_{j\in v_i} w_{ij}\mu_{j,t}. \tag{2.14}$$

Here again, w_{ij} quantifies the influence of agent j on agent i. Hence the second term is a weighted average of neighbours' beliefs, reminiscent of the FD and FJ paradigms. $\mathrm{BU}(\mu_{i,t}|s_{i,t+1})$ is a Bayesian update based on the observed signal $s_{i,t+1}$ at time $t+1$. Formally:

$$\mathrm{BU}(\mu_{i,t}|s_{i,t+1}) = \frac{\ell_i(s_{i,t+1}|\theta)\mu_{i,t}(\theta)}{\int_{\theta\in\Theta}\ell_i(s_{i,t+1}|\theta)d\mu_{i,t}(\theta)}. \tag{2.15}$$

The following convergence result holds.

Theorem 6 (Jadbabaie, 2012)**.** *Assume that*

1. *the network is strongly-connected,*
2. $a_{ii} > 0$ *for all* i*,*
3. *there exists* i *such that* $\mu_{i,0}(\theta^\star) > 0$*,*
4. *there is no* $\tilde{\theta} \in \Theta$ *such that* $\ell_i(\cdot|\tilde{\theta}) = \ell_i(\cdot|\theta^\star)$ *for all* i*.*

Then all agents eventually learn the truth, i.e. *almost surely* $\mu_{i,t}(\theta^\star) \underset{t\to\infty}{\longrightarrow} 1$ *for all* i*.*

With assumption 3. we require that at least one user initially assigns a positive probability to the true state: the population cannot get convinced of something that is initially deemed rigorously impossible by everyone. If assumption 4. does not hold, then signals cannot help distinguish between states $\tilde{\theta}$ and $\theta^\star$. Users might then be lead to wrong conclusions without any way of knowing it.

2.3.6 The Voter Model

I am particularly interested in the Voter Model (VM), introduced by Clifford and Sudbury (1973) and Holley and Liggett (1975) in the context of particles interaction. It is one of the most widely studied models of binary opinion dynamics. Consider a network with users holding opinions in $\{0,1\}$. Given an initial distribution of states, each agent is endowed with an independent exponential clock of parameter 1: whenever it rings, the user selects a leader uniformly at random and adopts their opinion. Equivalently, at the times of a Poisson process of parameter N, an agent is drawn uniformly at random and samples a new opinion under the distribution of their leaders'. We have the following consensus result.

Theorem 7. *Consensus is reached if there is a user who can reach everyone else.*

The interested reader may refer to Yildiz et al. (2010, Theorem 2.2) for a proof using Markov Chain modelling. The intuitive idea is that no matter the current number of 0 and 1 nodes, there exists a succession of individual state changes with strictly positive probability that results in everyone holding the same opinion. I refer the interested reader to the study of Vazquez and Eguíluz (2008) for a very complete

description of the model on uncorrelated networks. Many variants of the model have been proposed over the years—see the review from Redner (2019).

A particularly interesting one is the addition of zealots into the graph, which finds its source in a celebrated work from (Mobilia, 2003). In this study, zealots are stubborn agents who never change opinion. If zealots all defend the same opinion, then everyone will eventually adopt it. In the presence of zealots with different opinions however, consensus is usually not reachable (Mobilia, 2003; Sood et al., 2008). Instead, opinions converge to a steady-state in which they fluctuate indefinitely (Mobilia et al., 2007; Yildiz et al., 2013). This is illustrated in Figure 2.2.

2.3.6.1 Voter model on general networks

Most works on the VM only analysed certain chosen topologies (complete graph, ER graph, BA graph...), and few have proven results valid for any network. The works of Yildiz et al. (2013) and Masuda (2015) are important milestones in this regard, as they provide a way to calculate individual opinion distributions at equilibrium on any weighted, directed network with stubborn agents. Yildiz et al. (2013) proved that the probability of i holding opinion 0 at equilibrium was equal to the probability that a backward random walk starting in i reaches a stubborn agent supporting opinion 0 before one supporting opinion 1. They did not give an exact formula for this value, but it was later proven by Masuda (2015) that this is expressed by

$$x_i = \frac{\sum_{j\in\mathcal{N}} w_{ij}x_j + z_i^{(0)}}{\sum_{j\in\mathcal{N}} w_{ij} + z_i^{(0)} + z_i^{(1)}}. \tag{2.16}$$

Here, x_i is the probability for user i to have opinion 0 and $z_i^{(s)}$ is the influence of the s-zealot on user i.

2.3.6.2 Active links density

To study the evolution of a system under voting dynamics, researchers have mostly used the *magnetisation*, that is the average opinion of the population at equilibrium. Another quantity has attracted attention in the last few years, the *active links density*. A link is said to be active if the two nodes it joins have different opinions. Links

may switch between active and inactive over time, and a recurrent question is to find the average proportion of such links in the network (Avena et al., 2022; Caridi et al., 2019; Suchecki et al., 2005; Vazquez and Eguíluz, 2008). Usually, this proportion decreases exponentially with time, until it reaches zero in a state of consensus. Ramirez et al. (2022) initiated the study of active links density in the VM with zealots, through simulations, for complete and Erdös-Rényi networks.

In Chapter 5, I demonstrate a general formula to compute the probability to find two nodes holding different opinions for any directed, weighted network, with zealots, any (finite) number of possible opinions, and different update rates across agents. This probabilities can then be used to calculate the *generalised active links density*, which I introduce to extend the active links density to account for long-range, weighted interactions.

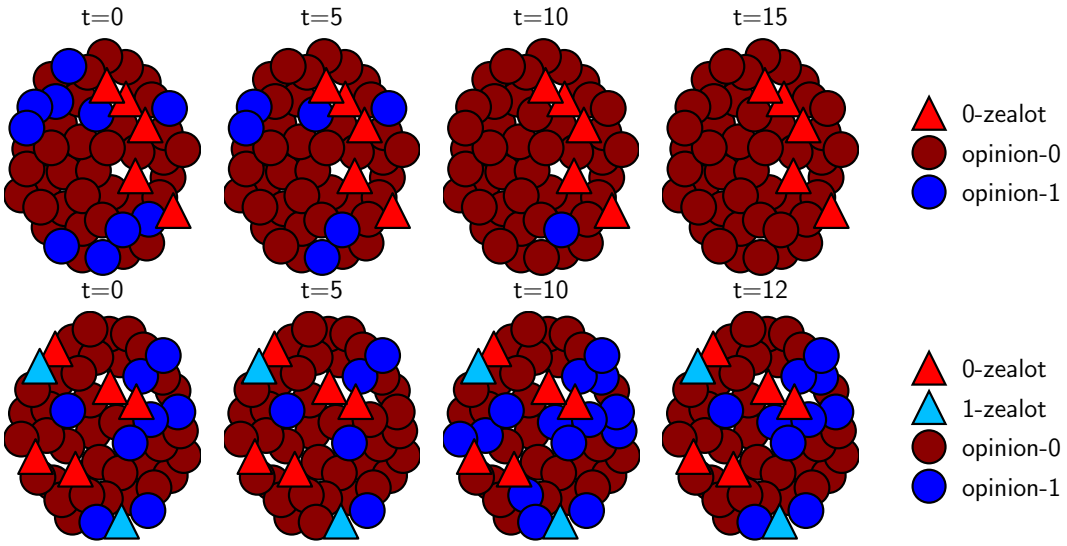

Figure 2.2: Example realisation of the Voter Model on a complete graph at different times with $N = 50$ nodes. **Left:** zealots only defend opinion 0, and everyone eventually adopts opinion 0. **Right:** with zealots in both sides, the system reaches a state of equilibrium where no opinion takes over.

2.3.7 Simple and complex contagion

In the Voter Model, agents change their opinions through a single contact with a neighbour of them. This is also the case in the Deffuant Model for example. These two models fall into a broader category: *Simple Contagion* models (Pastor-Satorras

et al., 2015). These are mostly used in epidemic spreading, but can also be applied to describe opinion dynamics and information diffusion. To model the diffusion of social behaviours in general however, models of *Complex Contagion* are often preferred: they assume that multiple exposures are required for an individual to become infected or to adopt a certain opinion (Centola, 2018). They were originally introduced by Granovetter (1978) via the threshold model, in which individuals change their opinion or behaviour when a high enough number of acquaintances of them have done so.

It is generally agreed upon that Complex Contagion models are more effective to describe the diffusion of social behaviour or opinions (Cencetti et al., 2023; Lerman, 2016; Mønsted et al., 2017; Notarmuzi et al., 2022; Sprague and House, 2017). However, these models rely on nonlinear dynamics, which makes their theoretical analysis difficult. I choose to mostly focus on Simple Contagion models in this thesis. First, to take greatest advantage of my background in the analytical study of mathematical models. Second, and most importantly, this allows me to propose efficient optimisation methods for the steering of the echo chamber effect. Having access to analytical formulas lets me wield optimisation algorithms to obtain quickly exact solutions for a wide range of model parameters. Without such formulas, I would have had to rely on heuristics such as greedy optimisation, which are not only much slower but also much less likely to find global optimums.

The trade-off between analytical tractability of Simple Contagion and accuracy of Complex Contagion is well illustrated by the Extended Newsfeed Model, that I also study in this thesis. The basic model corresponds to a Simple Contagion process, and I find a good correspondence between the opinion distributions predicted by the (tractable) model equations, and observations in a real-life OSP. I then propose a preferential reposting mechanism that introduces nonlinearities in the models equations, and corresponds to a Complex Contagion process. As expected, it exhibit a better fit with the data but can only be studied in simulations. Additionally, I find that optimisation results obtained for the base model can still be applied and perform well in the nonlinear version with preferential reposting. This is a good illustration

of the advantages and drawbacks of both categories of models.

Chapter 3

Literature Review

3.1 Empirical studies of echo chambers

The exact nature of echo chambers, and the extent to which they foster opinion polarisation, have attracted a lot of attention from the scientific community. Perhaps the earliest well-known work in that area is that of Adamic and Glance (2005), who found out that US political blogs mostly referred to others of similar leaning during the presidential election of 2004. A similar observation was made by Conover et al. (2011), who found that the retweet patterns within the US Twittersphere shows very high levels of political homophily. That is, users retweet others with similar political views way more often than opposite-minded ones. This result has been replicated multiple times since then in various contexts of North American politics, highlighting the Democrat-Republican divide (Barberá, 2015; Barberá et al., 2015; Garimella et al., 2017a; Garimella and Weber, 2017; Halberstam and Knight, 2016; Himelboim et al., 2013; Liu and Weber, 2014). For example Barberá et al. (2015) observed cross-cutting retweet rates two to five times lower than in-group ones, and during the election of House representatives in 2012, more than 90% of retweets of candidates' tweets were emitted by users of corresponding leaning (Halberstam and Knight, 2016).

Homophily is however not limited to retweeting behaviour nor to the United States, and has also been observed in the retweet and follow patterns of various other platforms and countries (Bakshy et al., 2015; Bright, 2018; Cinelli et al., 2020;

Cota et al., 2019; Garimella and Weber, 2017; Gruzd and Roy, 2014; Grömping, 2014; Halberstam and Knight, 2016; Himelboim et al., 2013; Hosseinmardi et al., 2020). In the Facebook friendship network studied by Bakshy et al. (2015), cross-cutting friendships only accounted for 20% of all connections. Bright (2018) did a transnational study covering all major European political parties, showing homophily even across borders when looking at the retweeting behaviour of partisans. Global controversial topics such as climate change and conspiracy theories have also been found to exhibit homophily across borders, in both their retweet and follow patterns (Bessi et al., 2016; Del Vicario et al., 2016; Weber et al., 2020; Williams et al., 2015). Halberstam and Knight (2016) also explored the relation between group size and homophily. They found that homophily increases with group size, and so does exposure to congenial content. Finally, the production and consumption behaviours often show homophilic patterns, in that most users create content aligned with what they consume (Garimella et al., 2018; Himelboim et al., 2013).

Echo chambers are characterised not only by a concentration of like-minded users, but also by a rejection of opposite ideas. On Twitter, this second feature is best expressed in the mention network. Halberstam and Knight (2016); Liu and Weber (2014); Williams et al. (2015) highlighted the existence of a "fighting bridge" between different communities, as a significant part of cross-cutting mentions were found to be aggressive and insulting messages. In the contexts of both North American politics and the Israel-Palestine debate, Liu and Weber (2014) found that more than 40% of cross-cutting mentions disagreements, and between a third and a half of the others were insults. The effect was most pronounced in the work of Williams et al. (2015), who studied the debate surrounding climate change. Almost 100% of cross-cutting mentions carried a negative sentiment, while this number ranges between 20 and 35% for in-group mentions. Negative relations might be more prevalent than we believe for a lot of users, especially journalists (Tacchi et al., 2022).

Because information spreads further amongst like-minded users (Del Vicario et al., 2016), the homophilic nature of the Twitter network favourises partisan content

when it comes to virality (Xia et al., 2020). This echoes the findings of Garimella et al. (2018), where the authors make an interesting distinction between different types of users. Partisans are the most polarised, well-embedded into their community, producing and consuming content aligned with their ideology. Gatekeepers only produce one-sided content but might consume from both groups. Bi-partisans, akin to what are often referred to as "moderate" users or "centrists", may produce and consume content from both sides: they are not "trapped" in an echo chamber. A major result from this work is that partisans benefit from the highest social reward, as they are significantly more popular, central, and embedded in their community than others, especially bi-partisans. Thus it seems like the "purest" users are the most popular and influential within their sphere, encouraging further the creation and diffusion of partisan content. Using bots to probe the US political Twittersphere, Chen et al. (2021) established further evidence of this phenomenon and found that partisan accounts received more followers than others.

Social reward mechanisms might also incentivise users to join conspiracy communities. In a study of Reddit, Phadke et al. (2020) found that people joining conspiracy communities were beforehand ostracized by mainstream users and communities. They often had low "karma", the platform's public measurement of one's popularity based on appreciation shown by others on your posts and comments. Rejected by the majority of the platform,[1] they are lured in conspiracy communities by their users and once in, they have most of their contact with other members of these groups.

Interestingly, this goes hand in hand with the idea that online activity and polarisation are positively correlated. There is strong evidence that more a user is active on an OSP, the more polarised they get and vice-versa (Bessi et al., 2016; Hosseinmardi et al., 2020; Vaccari et al., 2016; Weber et al., 2020; Wojcieszak, 2010). When posting and sharing more and more polarised content keeps increases your popularity more and more, there is no incentive to stop doing so. The very existence of feedback from your peers precludes any willingness for concessions with regards

[1] Reddit has been found to harbour mostly left-leaning and non-conspiracist users (Cinelli et al., 2020).

to diverging beliefs, unveiling a vicious circle of polarisation reinforcement and increase in activity. The more polarised a user is, the less they will interact with different-minded others, even on their political side — Bright (2018) found that moderate users were more likely to interact with other moderates across the left-right divide than with extremists on their side. This finding is supported by Conover et al. (2011) who showed that the more moderate a user is, the more they will mention and be mentioned by ideologically opposite users. However, several studies underline that centrist users might constitute but a minority of the online population, or perhaps a silent majority — especially when it comes to conspiracist content (Bessi et al., 2016; Cota et al., 2019; Del Vicario et al., 2017b; Zollo et al., 2017).

Further supporting this link between activity and polarisation, Garimella et al. (2017a) found that polarisation is at its strongest during heated debates. Through a long-term analysis of the US political Twittersphere, Garimella and Weber (2017) showed that polarisation is at its highest (resp. lowest) right before an election (resp. right after). Controversy thus acts as a driver for polarisation, and topics that are exempt from it such as sports show remarkably little polarisation (Barberá et al., 2015; Garimella et al., 2017a, 2018; Liu and Weber, 2014).

High polarisation on the US political Twittersphere is not that surprising, as several studies point out that partisanship has been steadily increasing in the country for longer than OSPs have existed (Andris et al., 2015; Dimock et al., 2014; Gentzkow et al., 2016; Lelkes, 2016). Because it is known that partisanship is correlated with attitude towards science (Dunlap et al., 2016; Funk and Tyson, 2020; Gauchat, 2012; Shi et al., 2017), it should be no surprise that conspiracy theories—which already existed before OSPs—are thriving as well. However, not only are these phenomenons not limited to the US, they also have exploded in intensity through social media (Peralta et al., 2023). They have reached a larger part of the world population and have been put to the forefront of today's news, for example affecting in a negative way our response to the global pandemic of CoVID-19.

Finally, several works provide evidence of the *backfire effect*. Debunking false information is not only ineffective on average Chan and Albarracín (2023),

it can often lead to an increase in the misplaced belief Betsch and Sachse (2013). On Facebook, Zollo et al. (2017) found that members of conspiracy groups often became more active within their community right after being exposed to debunking content. In a study by Bail et al. (2018), US Republican and Democrat partisans were (willingly) to cross-cutting content and were even more polarised afterwards. Similar results were obtained by Nyhan and Reifler (2010). Thus, any attempt to mitigate the echo chamber effect, or lessen polarisation, should take that into account and be careful when exposing users to uncongenial information.

3.2 Recent advances in opinion dynamics

With the advent of OSPs, there has been a surge of research on opinion dynamics. Mathematical models have proven a fertile ground to gain theoretical insight on online phenomenons such as echo chambers and opinion polarisation. These models have also allowed for an exploration of the consequences entailed by recommender systems. I expose some recent advances on these subjects. There exists a plethora of opinion dynamics models in the literature, and for the sake of brevity I only present a couple major axes of recent research on the topic.

As empirical works have highlighted, there is a wide variety of mechanisms that contribute to the emergence of echo chambers and opinion polarisation. Many novel models that incorporate these various social and psychological features have emerged in the past few years. For example, in the French-Degroot paradigm, simply pooling opinions of others may seem quite unrealistic. Dandekar et al. (2013) showed that averaging dynamics of the sort never yield polarised structures. Hence, they are invalid when it comes to an accurate description of opinions dynamics. It is not too surprising, as we know that people are subject to various other forces such as homophily or confirmation bias, amongst others. Dandekar et al. (2013) found that if the FD model is modified to account for confirmation bias, polarisation can occur.

3.2.1 Effect of stubborn agents

We have seen in Section 2.3 that inner biases (in the FJ model for example) and bounded confidence are simple mechanisms precluding consensus. They do not

necessarily imply the existence of echo chambers of opinion polarisation however, as a population of disagreeing agents may be well mixed and their opinions well spread over the opinion space. When most users are completely permeable to the opinions of others, the presence of stubborn agents favourising one opinion over the others is also determinant in the convergence process and the possibility of reaching a consensus. In the Voter Model for example, if there are stubborn agents promoting different opinions, consensus is not reachable (Mobilia et al., 2007; Yildiz et al., 2013).

Recently, Sikder et al. (2020) studied the impact of stubborn agents on opinion polarisation and echo chamber effect in social learning frameworks. They assumed that stubborn agents replace incongruent signals by congruent ones with a certain probability q. As the authors say, this models both *active* bias (deliberately choosing not to believe the information) and *passive* bias induced by the recommender system (proposing congenial content to the user in order to generate clicks/likes). The authors then explore the relation between quantity of stubborn agents and diversity of opinions, polarisation, and echo chambers. The authors notably find that the more connected the network is, the more it can absorb confirmation bias without affecting accuracy. Finally the authors make empirical verification with survey data, finding correlations between access to internet and belief in climate change conspiracies.

3.2.2 Private and public opinions

The difference between *private* and *public* opinion has been studied as a driver of polarisation. Indeed, expressed opinions may often diverge from genuine inner ones, as social context might push individuals to alter views they share with others. Duggins (2017) proposed a model that includes several social and psychological features such as conformity (*i.e.* tendency to seek positive feedback), homophily, and (in)tolerance. Depending on the exact balance between those effects, the author observed a wide array of situation. For example, in absence of psychological forces other than intolerance and homophily, the society either converges to total consensus or total polarisation. In settings otherwise favourable to polarisation (*resp.* consensus), conformity entails consensus (*resp.* distinctiveness entails polarisation).

The results were validated with empirical data: the author took opinions survey responses in the US, inferred the most likely model parameters, and showed that the output distributions matched those observed from the data.

Banisch and Olbrich (2019) also studied the idea that opinion polarisation stems from social feedback. Their model uses reinforcement learning and does not account for any other effect such as homophily, confirmation bias or backfire effect. Users express the opinions publicly and receive a reward in the form approval or disapproval from their peers. In this context, agents adjust what they express based on the feedback they are expecting. This feedback then leads to a re-evaluation that will affect future behaviour of the user. The authors show that in some networks with sufficiently high modularity, this will inevitably lead to polarisation. Densely connected groups with intial inclination for an opinion will drift more and more towards more extreme views.

Those results do not hold for Erdös-Rényi graphs however, which do not exhibit community structures. This shows that such structures are needed for polarisation to emerge. A phenomenon of *gatekeeping* is also observed, where beliefs are unable to spread across bridges between different communities. Indeed, picture a node at the fringe of a community, who communicates with other adjacent communities via their own fringe nodes. Despite connections outside the cluster, the considered agent's payoff is still slightly higher for him to try and adopt the opinion of the group to which it belongs. Because such nodes are the only bridge between communities, when they can't change opinions we immediately see that clusters will remain indefinitely entrenched in their views.

3.2.3 Negative influence

In order to account for antagonism, and potential backfire effects, Keuchenius et al. (2021) argue that it is necessary to consider negative ties when studying polarisation in social networks. Negative edge weights imply that interactions with foes are repulsive, and tend to widen the gap between discordant opinions. Thus, they are a convenient and simple way of modeling the backfire effect. Some works have extended models to incorporate negative ties, such as Li et al. (2013) for the

Voter Model, and Shi et al. (2016) for the DeGroot model. The works of Altafini (2012, 2013) have also explored the problem of consensus with negative connections. Interestingly, the author proves that dynamics on balanced networks[2] are exactly the same whether or not the network is a single all-positive cluster or two antagonistic clusters, if we consider the absolute value of opinions.

The idea of negative influence is also at the core of the model proposed by Hazla et al. (2020). In this framework, opinions are reinforced by both interactions with congruent and incongruent information. The authors consider multi-dimensional opinions in $\mathbb{R}^d$ describing users' views on several different topics, which are affected by scalar product with other vectors. Those can model contact with external influences such as advertising campaigns, newspaper articles, political debates on television... A positive (*resp.* negative) scalar product indicates positive (*resp.* negative) influence, nudging the users' opinion in the direction of (*resp.* opposite of) the applied vector. Interestingly, the authors show that if $d \geq 2$, polarisation occurs with probability one. Thus, it is the unavoidable fate of any system of agents who re-evaluate their opinions solely through the lenses of positive and negative influence. This antagonisation of opposite-minded individuals in the form of a backfire effect has been a recurring explanation for the existence of echo chambers, however studies like the one from Takács et al. (2016) showed that it is not necessarily observed in practice. Many models are amenable to reproduce polarisation of opinions without negative influence—*e.g.* Axelrod (1997); Friedkin and Johnsen (1990); Mäs and Flache (2013) amongst others.

Another appeal of the work of Hazla et al. (2020) is that they consider multi-dimensional opinions, while most work in the field consider uni-dimensional debates—*e.g.* left-wing vs. right-wing. It might however often be the case that opinions are better represented in higher dimensions. The left-right divide for example does not sufficiently explain the ideological differences between French political parties (Ramaciotti Morales et al., 2022). There has been an increase of research in multi-dimensional contexts (Baumann et al., 2021; Bizyaeva et al., 2023; Macy

[2] A signed network is said to be balanced if either all links are positive, or it is split into antagonistic clusters so that all intra-cluster edges are positive, and all inter-cluster edges are negative.

et al., 2021; Parsegov et al., 2017; Ye et al., 2020; Zafeiris, 2022).

3.2.4 Polarisation as a runaway process

Recently, Axelrod et al. (2021) raised the crucial question of the existence of levels of polarisation at which, no return is possible and polarisation will inexorably increase. They proposed a very exhaustive model, which includes many features: social influence, backfire effect, tolerance, exposure, self-interest and external shocks. They do not use negative links, but via an extended form of bounded confidence dynamics, they incorporate the possibility for interactions between different-minded agents to be repulsive. The authors find that intolerance is the key parameter. For example with highly intolerant populations the increase of exposure only leads to more polarisation. This is perhaps part of the reason why polarisation has increased with globalisation and OSPs, as a consequence of free, long-range, borderless communication—as put by Bogost (2021). Contrary to the results of the FJ model or the VM with zealots, Axelrod et al. (2021) find that even a small quantity of self-interest—*i.e.* attraction to a static innate opinion, is very effective at preventing polarisation.

The fact that polarisation can become a runaway, unstoppable process beyond a certain threshold also posits the questions of identifying this tipping point. This is at the core of the study from Macy et al. (2021), who provide an extensive study of the *hysteresis* of a multi-dimensional opinion dynamics model that includes influence, homophily, (in)tolerance and exogenous shocks. Hysteresis is a physical phenomenon, when changes in a system induced by the evolution of a parameter, are not symmetrically cancelled when tuning the parameter back to its initial value. Thus, passed a certain point, it might be very difficult to go back to previous states. The authors observe hysteresis on most of the parameter of the system, reinforcing the conclusions of Axelrod et al. (2021).

3.2.5 Evolving networks

All the works I have evoked so far are only concerned with the evolution of beliefs on a static social network. However, some may argue that the evolution of connections is almost as important, if not more. Taking into account dynamical networks, where

edges appear and fade, is of particular interest. Holme and Newman (2006) explored the idea of co-evolution between network structure and opinions evolution. They proposed a model in which *(i)* individuals form their opinions based on those of their neighbours and *(ii)* individuals with similar beliefs connect with each other. Both are repeatedly applied, with relative frequency controlled by a tunable parameter.

It appears that there exists a transition phase depending on the value of this parameter. When the parameter favourises *(i)* then users will form communities of pre-existing acquaintances with similar opinions. When it favourises *(ii)*, as it is the case with recommendation algorithms, users will connect with similar-minded others and break ties with pre-existing acquaintances with different opinions. This leads to the formation of echo chambers. Extensions of classical opinion models to include evolution of links have been studied (Del Vicario et al., 2017a; Grabisch et al., 2023; Kan et al., 2023; Klamser et al., 2017; Minh Pham et al., 2020; Proskurnikov et al., 2014; Sasahara et al., 2020). Often, connection rewiring based on homophily fosters the emergence of echo chambers.

Individuals may form and break connections based on their preferences, but these connections are also often suggested by the OSP's personalisation algorithm. Moreover, as the platform's algorithm controls what appears or not on the newsfeeds, it might reduce or augment the visibility of certain leaders, thus practicality adjusting the weights of connections, or even rewiring them. It is also common for the algorithms to incorporate features recommending users to one another ("People you may know"). I now present recent findings on the impact of such recommendations.

3.2.6 Impact of recommender systems

Rewiring connections based on similarity, be it of opinion or common acquaintances, tends to foster the emergence of echo chambers (Cinus et al., 2021; Ferraz de Arruda et al., 2022; Santos et al., 2021). Notably, Cinus et al. (2021) propose a framework to study the impact of recommending either social connections or content on any opinion dynamics model. Encouraging homophilic connections, as well as presenting too much congenial information to users, often leads to a feedback loop as both opinions and recommendations get more extreme (Rossi et al., 2021; Yi

and Patterson, 2019). Perra and Rocha (2019) find that the presence of triangles in the graph is a key driver of the emergence of echo chambers under various content recommendation policies. However note that under certain conditions, the OSP algorithm may actually reduce polarisation (Cinus et al., 2021; Ferraz de Arruda et al., 2022; Ramaciotti Morales and Cointet, 2021; Santos et al., 2021).

An early work by Daly et al. (2010) proposed an empirical study of impact of various types of friends recommendation algorithms on a social network's structure. The considered algorithms were: content-based (similarity of interests), topology-based (friend-of-a-friend), and mix of both. They apply these algorithms on IBM's SocialBlue, an intranet social networking platform dedicated to the company. Users were split into different groups, and each group was assigned a different recommendation algorithm. All algorithms were found to reduce the modularity of the network. The topology based (friend-of-friend) algorithm yielded a "rich get richer" effect. Content matching helped connect distant communities, but at the price of a lower acceptance rate. This goes against the idea that recommendation creates isolated communities—although there is no edge removal involved, and the data does not involve discussions on a controversial topic.

The work of Ferraz de Arruda et al. (2022) examines the influence of social network algorithms on echo chambers and polarisation, via a theoretical approach compared with real-world data. The author develop a model to analyse the dissemination of content within a social platform, and investigate potential algorithmic biases in the evolution of opinions. The findings indicate that the rewiring of friendships contributes to the formation of echo chambers. In some circumstances however, recommender system may lead to consensus. This nuance is also found in the works of Ramaciotti Morales and Cointet (2021); Santos et al. (2021). The former model the impact of different state-of-the-art recommender algorithms on the FD dynamics, applied to a Twitter network of users discussing French politics. The latter study the impact of structural link recommendation on polarisation. Namely, users are recommended others who share many neighbours with them—ideology of users and content is not taken into account. Although this type of recommendation is amenable

to polarisation—quantified as the deviation of opinions from the mean, sometimes recommending others with not many common friends moderates polarisation, even in the presence of a backfire effect.

Cinus et al. (2021) develop an interesting framework to study the effect of recommending connections. It allows for the use of any opinion dynamics model and any recommendation algorithm—the authors use 4 different ones, based either on topology or similarity of opinions. They consider various initial networks, with varying levels of modularity and opinion homophily. Results are then compared with a control experiment, where users are not subject to recommendations. The presence of an echo chamber effect is assess with the NCI and the polarisation of opinions with RWC—*cf.* Section 3.3 for definitions of these metrics. The authors find that recommender systems accentuate initial biases in the echo chamber effect: if homophily is initially high then NCI will increase, if it's low then then NCI will decrease. RWC almost always increases. However if modularity is too high, NCI and RWC decrease in most cases. Results are robust across the two different opinion dynamics models. Finally, the authors also suggests intervention strategies to reduce NCI and RWC.

Perra and Rocha (2019) propose a model similar to the one I develop in Chapter 6. Users have the ability to post their opinions on the newsfeeds of their followers, which have a limited size. To update their opinions, users randomly select from their newsfeeds at regular intervals. The opinions in the newsfeeds are subjected to various eviction policies, including random, oldest, most recent, preferential (removing diverging opinions), and nudging (prioritizing a selected opinion). The main findings of their study suggest that high clustering, characterized by a substantial number of triangles in the network, is the primary factor leading to the formation of echo chambers, regardless of the eviction policy employed. The preferential eviction policy demonstrates a higher tendency to induce echo chambers compared to other eviction policies. Additionally, the authors observe that the random eviction policy, which involves no filtering, is the least resilient against attempts to control opinions.

3.2.7 Empirical evaluations

While research on opinion dynamics is mainly theoretical, more and more efforts have been devoted to confronting the models with real-life data. A convenient way to do so is to use public records of votes or opinions, to avoid the problem of estimating political leanings from online communication networks. Liu et al. (2010) shows that the zero-temperature Ising model is a good predictor of opinion dynamics in the US congress. This is a powerful result, as this model is solely based on positive and negative peer influence without any additional psychological or social mechanisms. This demonstrates how peer influence can explain real-life dynamics of opinions. Leonard et al. (2021) study the polarisation of political elites in the US. They propose a nonlinear model of opinion polarisation that accounts for the mood of the population. They find that the model is able to explain precisely the level of polarisation amongst Republican Elites and Democratic Elites since 1960. Notably, they highlight a threshold effect: Republicans have passed a "non-return point", beyond which polarisation becomes a self-feeding mechanism and is thus very difficult to reduce. Duggins (2017) argue that total consensus or extreme polarisation are not to be seen in the real world and that opinions distributions are actually smoother and continuous, with pockets of extreme opinions to be observed within. The authors take opinions survey responses in the US, infer the most likely model parameters, and show that the output distributions match those observed from the data.

As mentioned, using data from OSPs is more difficult, in that it requires some knowledge of users opinions or political leanings, which are rarely accessible directly. Scholars have developed algorithms to infer them—*e.g.* Martin-Gutierrez et al. (2023); Pougué-Biyong et al. (2023); Ramaciotti Morales et al. (2022). A recent overview of available methods can be found in Aldayel and Magdy (2021). I mention a few interesting works on empirical validation of opinion dynamics in OSPs below, and refer the interested reader to Peralta et al. (2022) for an in-depth survey of the domain.

Drawing from the analysis of US political data extracted from Twitter, Halber-

stam and Knight (2016) develop a formula that quantifies homophily as a function of group size. As they observe that larger groups present higher degrees of homophily and higher exposure to like-minded content, the authors argue that "homophily generates a built-in advantage in knowledge for voters belonging to the majority group". Application on real data yields encouraging results, close with the theory. Sasahara et al. (2020) study the emergence of polarised communities through the lenses of influence and unfriending. They develop a new model taking these into account and show that even with very low value for the parameters, polarisation happens. They notably suggest that discouraging the formation of triadic closure with the recommendation algorithms could help against polarisation. Most importantly, they also validate their model's predictions against Twitter data from Conover et al. (2011). Sikder et al. (2020) show, via a Bayesian learning model, that access to internet impacts the belief in climate change conspiracies as it increases polarisation around the topic. Fernandez-Gracia et al. (2014) introduce a *noisy* voter model, with mobility of agents, to study the results of presidential elections in the United States. They find that model is able to capture statistical fluctuations, and long-range correlations, in the vote shares per county, for presidential elections in the United States.

Finally, the Newsfeed Model, that I extend later in this thesis, was tested against empirical data in a work of mine prior to this PhD research (Giovanidis et al., 2021). The model equations were able to capture the top influencers in both a Twitter and a Weibo dataset with a good accuracy, when compared to empirically evaluated values of influence.

3.3 Measuring the echo chamber effect

There is no consensus on how to measure the echo chamber effect, or polarisation, in social networks. A straightforward method is to use traditional measures of communities in networks, such as the modularity—*cf.* Section 2.2.2. This metric is purely topological however, and does not account for dynamics taking place on the graph, be it the evolution of opinions or the diffusion of information.

A metric more adapted in this regard was proposed by Garimella et al. (2016). They define the *Random Walk Controversy* (RWC) as the probability than a random walk initiated in one community ends up in another. Thus, while still based on topological features, it has a dynamical interpretation. The random walk could represent a piece of content spreading throughout a network, and thus a low probability of reaching a community when emitted from another could signify the presence of a echo chamber effect. Similar to this is the metric proposed by Diaz-Diaz et al. (2022), who use the the probability that information diffused from a community reaches another one, assuming it spreads according to a hybrid contagion model. Other topology-based metrics include Grabisch et al. (2023); Guerra et al. (2013). The former focuses on boundary nodes between communities, and the latter on the disappearance of connections between communities, in a model of opinion dynamics with evolving links.

A polarised distribution of opinions is generally defined as a two-peak distribution, with its mean roughly in between. In that vein, Musco et al. (2018) calculate polarisation as the sum of squared, mean-centered, opinions. Thus, it is higher as opinions are further from the mean. Martin-Gutierrez et al. (2023) propose a similar approach to treat multi-dimensional opinions, with a general measure of polarisation based on total variance of opinions. Interestingly, they show how to quantify how topics or groups of topics contribute to polarisation, via the eigen-decomposition of the covariance matrix. Sikder et al. (2020) define polarisation as the proportion of agents with the minority opinion, and echo chambers as sets of agents with connections only to like-minded others or to stubborn agents. The latter form the *boundary* of the chamber and have the power to filter the information that goes in and out, reminiscent of the definition of gatekeepers from Garimella et al. (2018). In the same vein, neighbourhoods with heavily skewed opinions are used by Perra and Rocha (2019) as a measure of echo chambers, and Chitra and Musco (2020) considers neighbourhoods with no disagreement between users.

The diversity of opinions, or content, that users are exposed to is also a good way to quantify the echo chamber effect. It is often expressed as a Gini coefficient

(Ramaciotti Morales and Cointet, 2021) or an Shannon entropy (Mackin and Patterson, 2019). I will use a similar approach. In a binary setting, Garimella et al. (2017c) are interested in the number of users reached by only one of two opinions. To quantify the diversity of information that user i is exposed to, Matakos et al. (2020) use

$$\sum_{(i,j)\in\mathcal{E}} w_{ij}(s_i - s_j)^2, \tag{3.1}$$

where s_i is the average leaning of information that i is exposed to. Thus, it is a weighted average of the distance between the opinion of a user and that of their leaders. A similar idea was proposed by Cinus et al. (2021), who defined the Neighbours Correlation Index (NCI) for a user i as the Pearson correlation coefficient between their opinion and the average one of their neighbours.

3.4 Control of opinion dynamics

The problem of controlling spreading processes, in particular information diffusion or opinion dynamics, has received a lot of attention since the early 2000s. Our objective to steer the echo chamber effect inscribes itself in the wake of this research.

3.4.1 Influence maximisation

Early works were concerned with *influence maximisation*, that is finding optimal sets of nodes who exert the most total influence over a network. In the example of a social network, we would expect that convincing these actors to promote a certain product, or to share a certain piece of content, will reach the most users and therefore yield the highest return on investment.

3.4.1.1 Seminal works

The work of Kempe et al. (2003) is foundational in this regard. They consider the following setting: start at t_0 with a set of initially *active* nodes, called the *seeds*. Others are *inactive*. At each time step, some inactive nodes may become active, and if none do, the process stops. The central question the authors try to answer is: how to find the smallest possible seed set so that a maximal number of nodes are active at the end of the process?

They propose two possible processes to model the activation of nodes, the Independent Cascade (IC) and Linear Threshold (LT) models. At each step of the process in the IC, each active node tries to activate each of his inactive neighbors independently and succeeds with a certain probability, fixed beforehand and that may depend on the considered nodes. For the LTM, a node will become active if a sufficient proportion of their neighbours are active—the threshold is fixed beforehand and may vary between nodes.

Kempe et al. (2003) found that although finding optimal seed sets is NP-hard, greedy algorithms can yield very good results. Both the IC and the LT models have inspired many subsequential works on the topic, *e.g.* Chen et al. (2010); Hu et al. (2014); Tong et al. (2017); Wang and Wang (2023) amongst others. I refer the interested reader to Banerjee et al. (2020); Zareie and Sakellariou (2023) for two recent reviews of the domain.

3.4.1.2 Recent advances

Yi et al. (2021) provide algorithms for selecting an optimal sets of stubborn nodes in order to push opinions in a chosen direction in the FD model. The work of Goyal et al. (2019) is similar but concerned with the FJ model, over a finite time horizon. Still for the FJ model, Abebe et al. (2021) seek to control opinions by acting on the exogenous influence received by users. The effectiveness of their method is demonstrated on several real-life graphs.

Liu et al. (2010) introduce a greedy algorithm to select seed nodes in order to maximise the magnetisation (*i.e.* share of one opinion) in the zero-temperature Ising model[3], on any graph. The optimisation procedure comes down to a min-cut problem, and outperforms the baseline that consists in choosing seed nodes according to degree. A similar problem is studied by Lynn and Lee (2016), who search for optimal external fields to apply on nodes in the Ising model in order to maximise the magnetisation. These fields can represent the recommender system for example. If the temperature is low, then the magnetisation is maximised by focusing the external

[3]The Ising model is ubiquitous in statistical mechanics. See for example https://www.damtp.cam.ac.uk/user/tong/statphys.html for an introduction.

field on the nodes with highest out-degree. If the temperature is high and the network strongly connected, the magnetisation is maximised when focusing on low-in-degree nodes.

Recently, Hazla et al. (2020) gave a necessary and sufficient condition to be able to make any number of multi-dimensional opinions converge to a chosen vector and provide a precise strategy to do so. They also discussed strategies for maximising proximity of opinions with a chosen vector using a limited number of interventions.

While most early works focused on IC and LT, the problem of influence maximisation is easily adapted to other models. One is of particular interest to us, the Voter Model.

3.4.1.3 In the Voter model

Even-Dar and Shapira (2007) study the problem of selecting seed nodes so as to maximise the share of opinion 0 at a finite time t. They found that the heuristic of selecting highest-degree nodes yielded optimal results. Yildiz et al. (2013) and Masuda (2015) push the idea further, integrating exogenous influence in the form of two external controllers (zealots) A and B. They develop a greedy algorithm to solve the following problem: knowing what nodes are influenced by B, what nodes should A target in order to hold a maximum influence on the network? Yildiz et al. (2013) provide a closed-form expression for the optimal solution in terms in the case where both zealots control a single node. Otherwise, they propose a greedy algorithm. Masuda (2015) looks at two different cases. In the first, each zealot controls a single node with the same budget. In the second case, it is ten nodes and A optimises the set of nodes to control, via stochastic hill-climbing algorithm. In undirected networks, the nodes selected are close to the highest-degree ones but on directed networks, this correlation vanishes.

This problem was studied further by Moreno et al. (2021), from both a continuous and discrete optimisation perspective. Namely, B's targets being given, A has to optimise theirs either by continuously distributing their budget, or by targeting a certain number of nodes each with the same intensity. The authors show that continuous optimisation performs better than discrete. Indeed, there is a small number of nodes

that are heavily targeted, a very small number that are not targeted at all, and the vast majority that are mildly targeted, with similar values. The authors also compare the results with several heuristics, namely degree-based allocation, shadowing (targeting the same nodes as B), shielding (targeting direct neighbours of B's targets), and random allocation. In the discrete regime, degree-based choice of nodes performs the best of all the heuristics. For the continuous case it is a combination of shadowing and shielding. Finally, the authors also show that if both A and B can react to each other's move, the Nash equilibrium of the game consists of equal allocation amongst all nodes for both controllers.

Other works with the voter model include that of Li et al. (2013), who demonstrate the utility of taking edge signs into account for the purpose influence maximisation. The voter model is modified so that some edges are negative and others positive. When a node copies their neighbour, if the edge is negative, they adopt the opposite opinion of theirs. The authors find that, greedy algorithms for finding optimal seed nodes to maximise the spread of one opinion, always perform better when edge signs are taken into account.

3.4.2 Mitigation of echo chamber effects and polarisation

With the increasing importance of social networks in the political debate and information diffusion, there has been a recent surge in research on methods to disrupt echo chambers and reduce polarisation. Thus, rather than trying to maximise the spread of an opinion, scholars have gained interest in the problem of seeking consensus (reduction of polarisation), or a maximal diversity of opinions in the network (reduction of the echo chamber effect). I am mainly interested in the latter, but mention some interesting works regarding the former.

Part of the research I evoke in this section is concerned with link rewiring. While the platforms do not alter the leader set of a user in theory, the personalisation algorithm does so in practice, as it may hinder or promote visibility for certain leaders. This will often have an impact on echo chambers, as platforms usually favourise homophilic connections. Thus, studying link rewiring policies is a fundamental aspect of research on the topic.

Musco et al. (2018) seek to minimise polarisation in the FJ model, while at the same time minimising disagreement between neighbours. Doing so, users are not exposed to too much cross-cutting content, reducing the risk of backfire effects. They define polarisation as the variance of equilibrium opinions, and disagreement as the average distance between the opinion of two neighbours. The authors consider *(i)* a method via control of link weights, and *(ii)* a method via control of users exogenous influence. These can be thought of as the action of a recommender system, *(i)* adjusting the exposition to content based on its source, and *(ii)* adjusting the exposition to content based on the opinion it conveys.

The work of Chitra and Musco (2020) is in the same vein. They consider the problem of minimising disagreement by acting on links, on both the FD and FJ models. Interestingly, the authors show that the FJ update rule is equivalent to users looking to minimize disagreement and internal conflict—defined as the distance between a user's innate preferences and their opinion. Through both theoretical developments and real-world data analysis from Reddit and Twitter, the authors give evidence that changes in edge weights induced by the OSP algorithm can lead to the formation of echo chambers—*i.e.* neighbourhoods with no disagreement between users. To try and correct this, they propose to add an penalization term to the rewiring dynamics orchestrated by the platform's algorithm, incentivising the platform to make many small modifications instead a few important ones. Under this new rule, the echo-chamber effect is mitigated while the platform's objective of reducing disagreement between neighbors is still fairly respected. Similarly, Yi et al. (2021) and Mackin and Patterson (2019) provide algorithms for selecting an optimal sets of stubborn nodes in order to minimise polarisation and disagreement (resp. diversity of opinions) in the FJ and FD models.

Recently, Grabisch et al. (2023) studied a FJ-like model with evolving connections. They assume that opinions in $[-1,+1]$ evolve under

$$x(t+1) = \left[\beta^t W(t) + (1-\beta^t)I\right] x(t), \tag{3.2}$$

where $\beta \in [0,1]$ quantifies the persistence of pre-existing opinions. Connections

are described by a time-evolving adjacency matrix $W(t)$: at each step, agents with opinions closer than a threshold σ create a connection—if not already existing, and agents with opinions further apart than a threshold τ break their connection—if it exists. At some point, β^t becomes too small for opinions to keep evolving, and they are crystallised in place.

The authors find that the diameter of the final distribution of opinions decreases with β, meaning that higher speed of crystallisation entails less variability in opinions. The risk of polarisation, quantified by the probability for the final network to become disconnected into several components, also increases with β. The authors then study optimal strategies for a social planner who would be able to select β, in order to minimise both the variability of opinions and the probability of polarisation. They also study strategies for two adverse political campaigns that seek to maximise their respective share of the electorate—one tries to maximise the number of negative opinions, and the other the number of positive ones.

Reminiscent of works on influence maximisation, Garimella et al. (2017c) and Matakos et al. (2020) study the problem of minimising the echo chamber effect by trying to find seed nodes that maximise the diversity of information users are exposed to. Both prove to be NP-hard, and the authors present greedy algorithms to find suboptimal solutions. Garimella et al. (2017c) considers the IC model with two competing campaigns, each modeled by a single propagation episode. They look for choices of seed nodes that minimise the number of nodes reached by only one campaign.

Close to what I present in Section 7.2, Matakos et al. (2020) seek to maximise the average weighted difference between opinions of neighbours—similar to disagreement in Chitra and Musco (2020); Musco et al. (2018); Yi and Patterson (2019). Interestingly however, they do not assume any underlying dynamics, but rather hypothesise that exposure vectors are already known. Thus, it is applicable to any model of opinion dynamics or information spreading.

Garimella et al. (2017b) propose a method to reduce polarisation as defined by the RWC (*cf.* Section 3.3) through addition of edges in the network. The focus is put

on which nodes to connect in order to get the best reduction in polarisation, while being sure that the edge is "accepted". Doing so, they avoid potential backfire effects. Four important points are taken into account: *(i)* it is possible to nudge people by recommending content from an opposing side, *(ii)* extreme recommendations might not work (backfire effect), *(iii)* moderate users are easier to convince, *(iv)* expert users and hubs are often less biased and can play a role in convincing others. The third point recalls findings from Banisch and Olbrich (2019) about gatekeeping users on the border of communities.

Perhaps the most interesting in the work of Garimella et al. (2017b) is the edge-acceptance mechanism: the authors consider that even though the recommendation algorithm may show contents from a specific node to another, we do not know for sure that this newly created link will properly function and not become a canal of antagonisation between users. The acceptance probability is computed as a function of the extremism of the concerned nodes, measured using the RWC. The authors show that their method is able to significantly reduce controversy on real-life Twitter datasets, and works best when connecting high-degree nodes with each other.

An interesting perspective brought by Cen and Shah (2020) highlight the fact that there are several actors at play, between the platform administrators, influencers and advertisers. This makes it delicate to regulate algorithmic personalisation without affecting the interests of these stakeholders, and any proposed solution to the problem should take their interests into account. The authors model the opinion of a user as a random variable conditioned by the content appearing on their personalised newsfeed, and are interested in detecting *learning divergence*: the emergence of a significant difference between this variable and the opinion conditioned on a neutral newsfeed without algorithmic personalisation. They propose a data-driven procedure to moderate learning divergence, and importantly show that this can be done even without knowledge of the process through which opinions are derived from the newsfeed.

Finally, it may be useful to inject some randomness into the recommender system. Rossi et al. (2021) model the interaction between a user's opinion and

the personalised recommendation, and show that it often leads to a feedback loop and more extreme opinions and recommendations. They find that adding some randomness into the recommender can help mitigate the phenomenon, while still preserving a minimal amount of relevance for the content shown to users.

Empirical attempts Finally, I mention two interesting empirical works. A heuristic study from Grevet et al. (2014), based on survey answers, echoes some of the results presented above. They suggest that online political polarisation could be reduced by, *(i)* slightly increasing exposure to weak connections and reducing exposure to strong connections—*cf.* Chitra and Musco (2020); Musco et al. (2018); Yi and Patterson (2019), *(ii)* striving to highlight common interests that may exist between opposite communities to bring them closer—*cf.* Garimella et al. (2017b).

Using bots, Yang et al. (2022) introduce a promising method to break open echo chambers in the presence of a backfire effect. In the context of anti-immigration discourse in European Twitter, they deploy three bots. The first one doesn't post and doesn't interact, the second one applies the *arguing* method: just posting pro-immigration content, the third one applies *pacing and leading*: starts by posting anti-immigration content then slowly shifts towards pro-immigration. Pacing and leading outperforms arguing, and contact with users, *i.e.* interaction, makes it even better. Interacting with users via the arguing bot however just makes it even worse. Thus, it is possible to slowly pull people out of echo chambers.

3.5 Limitations of existing works

Research on the Voter Model have mostly focused on undirected networks described by specific degree distributions, such as classical random graph topologies (Sood et al., 2008; Vazquez and Eguíluz, 2008; Yildiz et al., 2010). Few works have derived results valid for any given topology (Masuda, 2015; Yildiz et al., 2013). Moreover, while some have included additional features such as zealots, or larger opinion spaces, no general model that can account for many different features has been proposed. The evolution of active links in particular, an order parameter of great interest (Caridi et al., 2019; Ramirez et al., 2022), has not been yet generalised

to general settings. This is why I propose the Enhanced Voter Model, valid on any given directed, weighted network with zealots, multiple opinions and individual update rates.

The field of opinion dynamics has also mostly focused on social networks as a general concept, but often do not incorporate the mechanisms features by OSPs. The modelling of walls and newsfeeds for example, and how information flows between, has been scarcely studied (Cen and Shah, 2020; Perra and Rocha, 2019). The Newsfeed model, a work of ours started before this PhD project (Giovanidis et al., 2019), tackled this problem. I extend it further to account for opinion labelling, to describe the echo chamber effect, and to improve the correspondence between its theoretical predictions and empirical observations.

Most of the works on opinion control are heuristic or suboptimal algorithms due to the NP-hardness of the problem at hand. The principle of seed node selection is somewhat difficult to apply in practice: it means convincing some users of the networks to defend one and only one opinion, all the time. The practicalities of such endeavour, and most importantly its ethical implications, seem obstacles difficult to overcome if one wishes to put some of these methods into application. The vast majority of existing works focuses on edge recommendation, or link adjusting. The problem of *content recommendation*, that is what opinions should the content recommended to each user support, has barely been touched. Moreover, besides Garimella et al. (2017b); Musco et al. (2018); Yang et al. (2022), existing works do not account for potential backfire effects. Finally, most works consider uni-dimensional opinions, while it is not always an accurate reflection of reality (Ramaciotti Morales et al., 2022).

I address these limitations by proposing a global framework that benefits from analytical tractability, is effective at both the macroscopical and microscopical levels, can treat multi-dimensional opinions, accounts for backfire effects, is evaluated on real-life data, and is adaptable to various others problems such as reinforcing echo chambers or favourising one opinion versus the others. Indeed echo chambers, or consensus on a single opinion, might in some context be a phenomenon to wish for—

e.g. consensus on the reality of climate change, or the effectiveness of vaccines. As methods based on link recommendations have been widely treated in the literature, I focus on content recommendation, but my framework can easily accommodate other methods. The optimisation problems I treat also benefit from having computable exact solutions—as opposed to heuristic ones, or those based on greedy algorithms that usually return suboptimal solutions.

Chapter 4

Methodology

I now detail the methodology that will be used in the following chapters, containing the results obtained during the PhD.

4.1 Theoretical models and metrics of interest

In Chapter 5 and Chapter 6, I develop and analyse two mathematical models adequate for the description of echo chambers. The Enhanced Voter Model (EV Model) describes the evolution of opinions in a population subject to social influence. The Extended Newsfeed Model (EN Model) describes diffusion of content in an OSP, and includes labelling of items with the opinion they support. Notably, the EV Model is applicable to social networks in general, while the latter is specifically dedicated to the study of OSPs. To gain confidence in their viability, the models are confronted with empirical data. The datasets are described at the end of this chapter.

Towards my application to the steering of the echo chamber effect I demonstrate how to compute, for each model, two closely related quantities:

echo chamber effect (ECE) the proportion of congruent opinions users are exposed to,

accessible opinion diversity (AOD) the variance of opinions users are exposed to.

The ECE experienced by user n is denoted by Γ_n, and the average ECE over the whole population considered is $\langle\Gamma\rangle$. For the AOD I use Φ_n and $\langle\Phi\rangle$.

4.2 Steering the echo chamber effect at two levels

In Chapter 7 I demonstrate the applicability of my models to the benefit of steering the echo chamber effect. More precisely, I demonstrate how to compute optimal recommendation rates to effectively shift the average AOD. I deliberately choose not to act directly on the ECE. Indeed if a user supports opinion 0 and is entrapped in an echo chamber where all they see is opinion 0, the ECE can be minimised by simply recommending only opinion 1 to the user. But doing so, we risk to create a novel echo chambers—simply with opinion 1 instead of 0. This is why I choose to act on diversity, and as we will see, it positively impacts the echo chamber effect as wanted. I account for users preferences so as to avoid any backfire effect.

I propose two different, complementary approaches to the steering problem: a macroscopical one and a microscopical one. Each takes advantage of one of the two models I develop. In both cases, I consider a finite, discrete number of possible opinions and assume a known, fixed social network of interacting agents subject to inner biases. Importantly, I assume no pre-existing recommendation algorithm.

Not that, in the application, I use global averages when computing the ECE and the AOD over the whole network. This is not optimal in scale-free graphs for example, where nodes with thousands of connections are more important than those with just a few. Thus, when using the methods I develop, depending on the context it might be useful to consider weighted averages to reflect the different relative importances of nodes.

4.2.1 Macroscopical approach

The first approach is based on the EV Model, and adequate for a global perspective. In particular, it is applicable to social networks in general, beyond the precise settings of OSPs. I make the mean-field assumption of a fully-connected population of like-minded users, with the same inner biases and no individual features differentiating one from the others. This is particularly relevant for homogeneous groups of users sharing the same beliefs, where everyone is exposed to the opinion of all others. Facebook groups or Reddit subs fit this description quite well, as all the content posted therein is presented to everyone, with no regards for the potential

heterogeneity of connections between users. For that I leverage the EV Model, which extends the traditional voter model. Belonging in the field of statistical physics, it is well suited for a macroscopical point of view.

4.2.2 Microscopical approach

The second approach is based on the EN Model. It incorporates user-level features, befitting a local perspective. The model describes the propagation of content throughout the newsfeeds of users in an OSP, and lets me quantify precisely the distribution of content, and thus opinions, on the newsfeeds of users. Importantly, it features high analytical tractability, meaning I can propose fine-grained optimisation problems to steer the echo chamber effect at the user-level. This approach allows for individually targeted action, provided availability of low-level information regarding users characteristics and connections between them. However, one does not always have access to such information—perhaps because of the limitation of the data at hand, or the difficulty to infer from it, and the uncertainty that would result from such inference process. Hence the necessity for the higher level, macroscopical approach, that only requires some broad knowledge about the system of interest.

4.3 Mathematical tools

I briefly introduce some mathematical tools and results that will be useful in the analysis or my models.

4.3.1 Markov chains

A discrete, homogeneous Markov Chain is a infinite sequence $X_1, X_2, \ldots$ of random variables. They are all valued in some common discrete set called the *state space*, its elements called the *states* of the system. They satisfy the Markov property, that is each one only depends on the precedent:

$$\mathbf{P}(X_k = x_k | X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) = \mathbf{P}(X_k = x_k | X_1 = x_1). \tag{4.1}$$

These probabilities are supposed invariant over time: they do not depend on k. They are encoded in a *transition matrix* P, whose $(i,j)^{\text{th}}$ element is defined by

$$P_{ij} = \mathbf{P}(X_k = j | X_{k-1} = i), \quad \forall k \geq 1. \tag{4.2}$$

If π_k is a horizontal vector encoding the distribution of X_k, it holds that $\pi_k = \pi_{k-1}P$. The matrix transpose P^T can be seeen as the adjacency matrix of a graph $\mathcal{G}$. The chain is said to be *irreducible* is $\mathcal{G}$ is strongly connected, and *aperiodic* if there is no integer $k > 1$ so that $P^k = P$. An irreducible, aperiodic Markov chain is said to be *ergodic*.

Steady-state Eventually, the chain may reach a state of equilibrium, or steady-state. Any left eigenvector of P characterises a state of equilibrium. If the chain is ergodic, there exists such a state, and it is unique—up to multiplication by a constant. If not, there may exist no steady-state, or there may exist multiple ones. A normalised left eigenvector of P is called a *stationary distribution* of the chain, and gives the probabilities of finding the system in each state when making a punctual observation at equilibrium.

4.3.2 Simulations

To simulate equilibrium distributions, or estimate them empirically, I will need to perform time averages. Indeed, ensemble averages are *(i)* unavailable for datasets that only cover a single realisation of a process, and *(ii)* too costly to compute with sufficient precision. The ergodic theorem assures us that time averages and ensemble averages correspond, under mild conditions that are always verified in this work.

Theorem 8 (Ergodic theorem)**.** *Let $(X_t)_{t\geq 0}$ be a continuous-time Markov chain over some state-space I. Assume X is irreducible, positive recurrent and denote by π its invariant distribution. Then for any bounded function $f : I \mapsto \mathbb{R}$ we have almost surely*

$$\frac{1}{T}\int_0^T f(X_t)\,dt \longrightarrow \sum_{k\in I} \pi_k f(k), \quad \text{as } T \to \infty. \tag{4.3}$$

I refer the interested reader to Norris (1997) for a proof. This theorem can

be seen as the law of large numbers for Markov chains; it expresses that means computed over time converge towards expectations under the invariant distribution.

Consider a simulation of X with jumps at times $t_1, \ldots, t_m$. The process is constant between them. Consider also a dataset with a finite number of datapoints, corresponding to observations of X at different times, that I also denote by $t_1, \ldots, t_m$ for simplicity. The system might evolve in-between these points, but we would not know, and I assume that it did not. The left-hand side of Eq. 4.3 becomes

$$\frac{(t_2 - t_1) f(X_1) + \ldots + (t_n - t_{n-1}) f(X_n)}{t_n - t_1}. \tag{4.4}$$

This is the formula I use to compute time averages for simulations and empirical evaluations.

4.4 Datasets

I now introduce and briefly describe the various datasets I am going to use in my analysis.

4.4.1 UK and US elections

This is used in Section 5.5 to evaluate the performance of the EV Model. I use the official database of the United Kingdom general elections results, published by the House of Commons (Audickas et al., 2020), as well as results for presidential elections in the United States manually collected from Wikipedia[1]. Each time I am interested in the percentage of popular votes won by the two major parties. In the UK dataset, the quantity of interest is the percentage of popular votes won by the Conservative[2] and Labour parties in each general elections from 1922 onwards. In the US dataset, it is the number of popular votes gathered by Republicans and Democrats in each presidential elections from 1912 onwards.

[1] https://en.wikipedia.org/wiki/United_States_presidential_election#Popular_vote_results

[2] The dataset also includes in Conservative results: National, National Liberal and National Labour candidates for 1931-1935; National and National Liberal candidates for 1945; National Liberal candidates from 1945 to 1970.

4.4.2 The Elysée2017 dataset

I evaluate the EN Model on a Twitter dataset, published by Fraisier et al. (2018) and relating to the 2017 French presidential campaign. It includes 2,414,584 tweets and 7,763,931 retweets from 22,853 Twitter users discussing the election. For each tweet I have: [PostID, TimeStamp, UserID, RePostID]. Users have been manually annotated by experts with political affiliations describing support for one or two of the main competing parties:

FI France Insoumise, far-left party (candidate Jean-Luc Mélenchon),

PS Parti Socialiste, left-wing party (candidate Benoit Hamon),

EM En Marche, centre party (candidate Emmanuel Macron),

LR Les Républicains, right-wing party (candidate François Fillon),

FN Front National, far-right party (candidate Marine Le Pen).

For some users the affiliation is unknown and I remove those from my study. Amongst remaining users, a small percentage are affiliated to two different parties (*e.g.* "PS/EM"). In addition, Papanastasiou and Giovanidis (2023) collected the *follow* graph, which was not part of the original dataset. I removed users who did not tweet nor retweeted anything and restricted myself to the largest strongly connected component of the followers graph. Finally some profiles were not available anymore and I end up with an anonymised dataset $\mathcal{D}$ that features $N = 8,277$ users and $E = 975,168$ edges. In Figure 4.1 I show the follow graph, the retweet graph, and the proportion of users supporting each party. Basic statistics for the dataset are summarized in Table 4.1.

4.4.3 Toy datasets

In Chapter 5, I will use four toy datasets to study the correlation between certain topological metrics and discord probabilities in the EV Model. Basic statistics of these datasets are presented in Table 4.2. Note that I only kept the largest weakly connected component of each network.

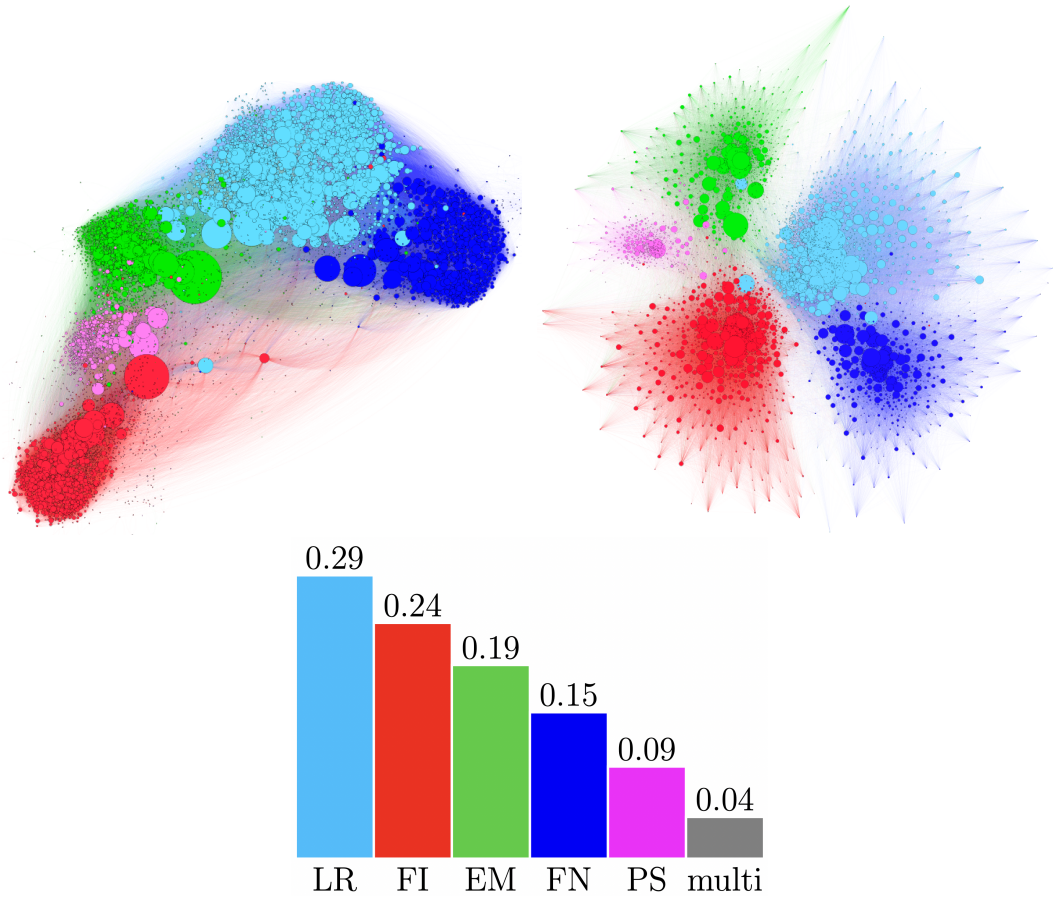

Figure 4.1: Top left: Follow graph of `#Elysée2017fr`. There is an edge from i to j if i is a leader of j. Colors indicate the 5 main parties. **Top right:** Retweet graph of `#Elysée2017fr`. There is an edge from i to j with weight equal to the number of times j retweeted i. **Bottom:** Proportion of users supporting each party. 'multi' stands for users with more two affiliations.

The first is the celebrated Karate Club dataset from Zachary (1977), that represents the social interactions among members of a university karate club. The edges in the dataset indicate friendships or interactions between the members, while the absence of an edge between two nodes implies the lack of a direct connection. During the study, a conflict arose within the club, leading to its later division into two separate clubs. Nodes are labeled by which of these two clubs they adhered to.

The `football` dataset comes from Girvan and Newman (2002). It describes the network of American football games between Division IA colleges during the regular season of fall 2000. Nodes are clubs, and there is an edge between two if they played each other. Nodes are labeled by the conference to which they belong (Atlantic Coast,

Table 4.1: Descriptive statistics on the `#Elysée2017fr` dataset ("#" means "number of").

	`#Elysée2017fr`
Time window	168 days
# users	8,277
# tweets	920,520
# retweets	2,934,830
Mean #tweet/user	111.2
Mean #retweets/user	355.6
Max #tweet	22,833
Max #retweet	23,171
% users with #tweet > 0	72.93
% users with #retweet > 0	98.94
# edges	975,168
Mean # followers	117.8
Max # followers	3,729
Max # leaders	1,551

Table 4.2: Descriptive statistics on the toy datasets.

	`zachary`	`football`	`email`	`polblogs`
# nodes	34	115	986	1,222
# edges	156	1,226	25,552	19,021
Density	0.14	0.09	0.03	0.01
# communities	2	12	42	2
Directed	-	-	✓	✓
Self-loops	-	-	-	✓
Modularity	0.42	0.53	0.43	0.43

Big East, Big Ten, Big Twelve, Conference USA, Independents, Mid-American, Mountain Wes, Pacific Ten, Southeastern, Sun Belt, Western Athletic).

The `email` dataset is taken from Leskovec et al. (2007). It was generated using email data from a large European research institution. The presence of an edge from one node to another indicates that the first person sent at least one email to the other. Nodes are labeled by the department to which the corresponding person belongs.

The `poblogs` dataset is another famous dataset, taken from Adamic and Glance (2005). It represents the directed network of hyperlinks between political blogs during the period leading up to the 2004 United States presidential election. Each blog is labeled by its political leaning, 'liberal' or 'conservative'. It is one of the first and most cited examples of online social network exhibiting homophily.

Chapter 5

The Enhanced Voter Model for opinion evolution in social networks

The Voter Model has been widely studied in the context of opinion dynamics. I develop a highly general framework, the Enhanced Voter Model, that encompasses many important extensions of the traditional Voter Model, and is applicable to any given directed, weighted network, without any requirement regarding the topology or the degree distribution of the network. Doing so, I contribute to the literature by extending important concept and results to a very general setting, while their study had until now mostly relied on approximations and strong assumptions regarding the degree distribution. My framework can accomodate any finite number of opinions (Starnini et al., 2012; Yildiz et al., 2010), zealots (Chinellato et al., 2015; Mobilia et al., 2007), and individual update rates (Masuda et al., 2010). In this context, I derive probabilities of discord between agents, and propose a new way to compute the active links density that accounts for weighted, long-range interactions. Importantly, I make no approximation and the results are exact.

In my application to steering echo chambers I will be interested in complete networks of identical agents, and I characterise my results in this precise settings. Except Section 5.5, the results presented here have been submitted to *Physical Review E* and are currently under review (Vendeville et al., 2023b). In Section 5.5 I assess the model's performance in forecasting the results of elections in the US and the UK, solely based on previous ones. This study was published in Vendeville et al.

(2021).

5.1 General setting

Consider a group of agents interacting through a social network and holding discrete opinions $(\sigma_i)_i$ valued in some set $\mathcal{S} = \{1, \dots, S\}$. Opinions are bound to evolve due to influence from peers, and I let $w_{ij} \geq 0$ denote the strength of influence that agent j exerts on i.

5.1.1 Zealots

To model exogenous sources of influences, I assume that for each opinion s there exists an entity external to the agent graph that promotes it. This entity is called the s-zealot, and exerts an influence of strength $z_i^{(s)}$ on agent i. This influence is an aggregate of all forces other than interactions within the graph, that may push i towards opinion s. In this chapter, I mostly think of it as representing an inner bias of agent i. For the later purpose of steering the echo chamber effect, it will also conveniently model the influence of the recommender system. I call *zealousness* of agent i, and denote by $z_i^{(s)}$, the total amount of influence exerted on them by the s-zealot.

5.1.2 Agent graph

Let $\mathcal{L}_i := \{j \in \mathcal{N} : w_{ij} > 0\}$ be the set of all (non-zealots) agents with influence on i, called *leaders* of i. I allow self-loops, *i.e.* $i \in \mathcal{L}_i$. For the sake of simplicity and without loss of generality, I consider values of influence to be normalised: for any agent $i \in \mathcal{N}$,

$$\sum_{j \in \mathcal{L}_i} w_{ij} + \sum_{s \in \mathcal{S}} z_i^{(s)} = 1. \tag{5.1}$$

The directed, weighted graph of all agents will be denoted by $\mathcal{G}$, where w_{ij} is the weight of edge $j \to i$. I assume $\mathcal{G}$ to be weakly connected—if it is not, one can apply the results to each component of the graph separately. I let W denote the weighted adjacency matrix of the graph. I do not consider zealots to be part of the agent graph, but without ambiguity I allow myself to say that agent i can be "reached" by the s-zealot if either $z_i^{(s)} > 0$ or there exists an agent j with $z_j^{(s)} > 0$ and a path from j to

i.

5.1.3 Dynamics

Each agent is endowed with an exponential clock of parameter 1. When their clock rings, agent i updates their opinion in one the two following way.

Influence of peers agent i adopts the opinion of a leader chosen at random, with probability w_{ij} for leader j.

Exogenous influence agent i adopts the opinion of a zealot, with probability $z_i^{(s)}$ for the s-zealot.

Thus, with rate w_{ij} agent i copies the opinion of j, and with rate $z_i^{(s)}$ they adopt opinion s via the s-zealot.

5.2 Echo chamber effect and opinion diversity

As said in Chapter 4, I am interested by two metrics: the ECE and the AOD. The first quantifies the amount of congruent opinions that users are exposed to. Let ρ_{ij} denote the *discord probability* for users i and j, that is the probability that they hold two different opinions. I define the echo chamber effect experienced by user i as

$$\Gamma_i = \frac{\sum_{j\in\mathcal{L}_i} w_{ij}(1-\rho_{ij})}{\sum_{j\in\mathcal{L}_i} w_{ij}}. \tag{5.2}$$

I demonstrate below how to compute values of ρ. In the application I consider complete, unweighted networks so that both the weighted and unweighted definitions are equivalent. I denote by $\langle\Gamma\rangle$ the average ECE over all users.

My second quantity of interest is the AOD, or the variance of opinions that users are exposed to. I first define the S-sized exposure vector of user i:

$$y_i = \frac{\sum_{j\in\mathcal{L}_i} w_{ij}x_j}{\sum_{j\in\mathcal{L}_i} w_{ij}}. \tag{5.3}$$

The entry of y_i with coordinate s is the average proportion of opinion s amongst the

leaders of i. The AOD for user i is then

$$\Phi_i = \frac{S}{S-1} \sum_{s \in \mathcal{S}} y_i^{(s)} (1 - y_i^{(s)}). \tag{5.4}$$

This is akin to an entropy of y_i. The constant in front ensures that it ranges from 0 (all leaders are fixed on one single, common opinion) and 1 (perfect mix of opinions). I let $\langle \Phi \rangle$ denote the average AOD over all users.

5.3 Mathematical analysis

I now detail how to derive individual opinion distributions and discord probabilities between agents. I let $x_i^{(s)}$ denote the probability for agent i to hold opinion s. The vector x_i is the opinion distribution of i. Discord probabilities are defined by

$$\rho_{ij} = \mathbf{P}(\sigma_i \neq \sigma_j), \tag{5.5}$$

and quantify how often i and j can be found holding different opinions. The quantity $1 - \rho_{ij}$ is called probability of harmony. Both x and ρ are time-dependent quantities, but I omit the time parameter to avoid cumbersome notations.

5.3.1 Evolution of opinions

A straightforward extension of Masuda (2015, eq. 3) gives

$$\frac{dx_i^{(s)}}{dt} = (1 - x_i^{(s)}) \left[\sum_{j \in \mathcal{L}_i} w_{ij} x_j^{(s)} + z_i^{(s)} \right] - x_i^{(s)} \left[\sum_{j \in \mathcal{L}_i} w_{ij} (1 - x_j^{(s)}) + \sum_{r \neq s} z_i^{(r)} \right]. \tag{5.6}$$

This reduces to

$$\frac{dx_i^{(s)}}{dt} = \sum_{j \in \mathcal{L}_i} w_{ij} x_j^{(s)} + z_i^{(s)} - x_i^{(s)}. \tag{5.7}$$

Hence, at equilibrium we have

$$x_i^{(s)} = \sum_{j \in \mathcal{L}_i} w_{ij} x_j^{(s)} + z_i^{(s)}. \tag{5.8}$$

As expressed by Yildiz et al. (2013, Prop. 3.2), $x_i^{(s)}$ is the probability that a backward random walk initiated at i reaches the s-zealot before another zealot. Hence, $x_i^{(s)} > 0$ if and only if i can be reached by the s-zealot.

The dynamics correspond to those of a continuous-time Friedkin-Johnsen model (Section 2.3.3). The number of distinct equilibrium states depends on the topology of the agent graph and the influence of zealots. Assuming that every agent can be reached by at least one zealot, Eq. 5.1 and Lemma 1 (Section B.1 in Appendix) imply that the spectral radius of W is strictly less than 1, and there is a unique equilibrium state. If $z_i^{(s)} = 0$ for all i, s, we uncover a continuous-time DeGroot model (Section 2.3.2). In that case, consensus is reached if there exists an agent able to reach every other.

5.3.2 Discord probabilities

I now turn to the study of discord probabilities, defined by Equation 5.5. Trivially $\rho_{ii} = 0$ and $\rho_{ij} = \rho_{ji}$. The discord probability between i and the s-zealot is simply $1 - x_i^{(s)}$. To enhance readability I denote without distinction by ij or ji the unordered agent pair $\{i, j\}$. It is tempting to simply write

$$\begin{aligned}
\rho_{ij} &= P(\sigma_i \neq \sigma_j) && (5.9)\\
&= \sum_{s \in \mathcal{S}} P(\sigma_i = s, \sigma_j \neq s) && (5.10)\\
&= \sum_{s \in \mathcal{S}} P(\sigma_i = s) P(\sigma_j \neq s). && (5.11)\\
&= \sum_{s \in \mathcal{S}} x_i^{(s)} (1 - x_j^{(s)}). && (5.12)
\end{aligned}$$

However this assumes that the opinions σ_i and σ_j are independent, which is not guaranteed. This assumption is for example violated if i and j are neighbours, as illustrated in Figure 5.1. Let me first focus on the general case, valid for any i, j. Later on I characterise cases where Eq. 5.12 holds.

5.3.2.1 General case

There are two types of events that lead to i adopting an opinion different than j's:

1. i copies agent $k \neq j$, who holds another opinion than j's. This happens at rate $w_{ik}\rho_{jk}$.

2. i copies an s-zealot while j holds an opinion different than s. This happens at rate $z_i^{(s)}(1-x_j^{(s)})$.

Hence i adopts another opinion than j's at rate

$$\sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{s\in\mathcal{S}} z_i^{(s)}(1-x_j^{(s)}). \tag{5.13}$$

The same reasoning gives the rate at which j adopts another opinion than i's, and the pair ij switches from harmony to discord at rate:

$$\Delta_{ij}^{-} = (1-\rho_{ij})\left[\sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{s\in\mathcal{S}} z_i^{(s)}(1-x_j^{(s)}) + \sum_{k\in\mathcal{L}_j} w_{jk}\rho_{ik} + \sum_{s\in\mathcal{S}} z_j^{(s)}(1-x_i^{(s)})\right], \tag{5.14}$$

and from discord to harmony at rate

$$\Delta_{ij}^{+} = \rho_{ij}\left[\sum_{k\in\mathcal{L}_i} w_{ik}(1-\rho_{jk}) + \sum_{s\in\mathcal{S}} z_i^{(s)}x_j^{(s)} + \sum_{k\in\mathcal{L}_j} w_{jk}(1-\rho_{ik}) + \sum_{s\in\mathcal{S}} z_j^{(s)}x_i^{(s)}\right]. \tag{5.15}$$

Subtracting Δ_{ij}^{-} from Δ_{ij}^{+}, we obtain the master equation

$$\frac{d\rho_{ij}}{dt} = \sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{k\in\mathcal{L}_j} w_{jk}\rho_{ik} + \sum_{s\in\mathcal{S}} z_i^{(s)}(1-x_j^{(s)}) + \sum_{s\in\mathcal{S}} z_j^{(s)}(1-x_i^{(s)}) - 2\rho_{ij}. \tag{5.16}$$

The evolution of discord probabilities is thus governed by a system of linear differential equations. Setting the left-hand side to zero gives us the equilibrium discord probability for the pair ij:

$$\rho_{ij} = \frac{1}{2}\left[\sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{k\in\mathcal{L}_j} w_{jk}\rho_{ik} + \sum_{s\in\mathcal{S}} z_i^{(s)}(1-x_j^{(s)}) + \sum_{s\in\mathcal{S}} z_j^{(s)}(1-x_i^{(s)})\right]. \tag{5.17}$$

Let me write this in matrix form as

$$\rho = V\rho + u. \tag{5.18}$$

If there are no zealots, Eq. 5.16 becomes

$$\frac{d\rho_{ij}}{dt} = \sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{k\in\mathcal{L}_j} w_{jk}\rho_{ik} - 2\rho_{ij}, \tag{5.19}$$

and at equilibrium,

$$\rho_{ij} = \frac{1}{2}\left(\sum_{k\in\mathcal{L}_i} w_{ik}\rho_{jk} + \sum_{k\in\mathcal{L}_j} w_{jk}\rho_{ik}\right). \tag{5.20}$$

In a state of consensus all discord probabilities are 0. Otherwise, the various equilibrium states are given by the leading eigenvectors of V.

In the case with zealots, I show in Section B.1 (Appendix) that the spectral radius of V is strictly less than, assuming every agent can be reached by a zealot. Hence, the system has a unique solution, which can be efficiently computed by iterating

$$\rho(k) = V\rho(k-1) + u \tag{5.21}$$

for any initialisation $\rho^{(0)}$ with values in $]0,1[$, and the convergence rate depends on the spectral radius of V. The proof of that statement can be found in Giovanidis et al. (2021, Thm. 4). In practice, I choose to stop when no element of $\rho(k)$ changes more

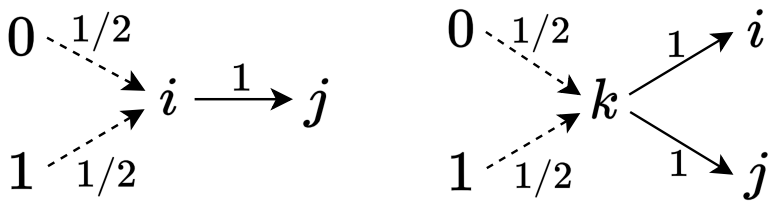

Figure 5.1: Dependency between opinions. Nodes 0 and 1 are the 0- and 1-zealot respectively. Numbers along the arrows denote edge weights. On the left, there is a path from i to j. On the right, there is none but i and j have a common ancestor k. In both cases, the correct formula (5.17) gives $\rho_{ij} = 1/4$, while (5.12) yields $\rho_{ij} = 1/2$.

than 0.1% in a single step, *i.e.* when

$$\max_{i,j} \frac{|\rho_{ij}(k) - \rho_{ij}(k-1)|}{\rho_{ij}(k)} < 10^{-4}. \tag{5.22}$$

Agent pairs with $\rho_{ij} = 0$ are excluded from the calculation.

Finally, note that all the equations I derived are easily adapted to account for various update rates $r_1, \ldots, r_N$ across agents. It suffices to scale each edge weight w_{ij} by r_i, to replace $2\rho_{ij}$ by $(r_i + r_j)\rho_{ij}$ in Eq. 5.16 and $1/2$ by $1/(r_i + r_j)$ in Eq. 5.17. Unless stated otherwise, I stick to the traditional setting $r_i = 1$ for all i.

5.3.2.2 Independent pairs

There are some cases where the opinions σ_i and σ_j are independent, and one can use Eq. 5.12 to calculate ρ_{ij} without having to solve a possibly large linear system. Independence holds if one of the following is verified:

1. σ_i or σ_j is constant, or

2. there is no path from i to j nor from j to i, and i and j have no common ancestor.

The first comes from the fact that a constant is independent from any other random variable. In particular, it is verified if i or j is a zealot. The second assures us that i and j do not influence each other, and that they are exposed to strictly different sources of influence: their opinions evolve in total independence. As illustrated in

Figure 5.1, if one of these assumptions is violated then (5.12) gives an incorrect result.

When i and j have the same opinion distribution x, (5.12) is akin to the entropy of x. Hence the more uniform x is, the higher the discord. Equation 5.12 is also exactly 1 minus the cosine similarity between x_i and x_j. While for dependent agent pairs the discord probabilities are given by (5.16) and (5.17), Eq. 5.12 may still be used for the purpose of measuring dissimilarity of opinion distributions.

5.3.2.3 Generalised active links density

The active links density is the average discord between all neighboring agents. While convenient for regular, unweighted graphs, this definition suffers from two shortcomings when it comes to general networks.

First, not all edges are created equal: if $w_{ij} = 0.80$ and $w_{ik} = 0.01$, then j holds a strong power of influence over i, while k barely has any at all. Discord ρ_{ij} between i and j will thus often be much more relevant to the analysis than ρ_{ik}, a difference not accounted for when taking a simple unweighted average. Second, two agents may be closer than they appear: if $w_{ij} = 0$ but $w_{ik} = w_{kj} = 0.9$, agent j exerts non-negligible influence on i via k, despite them not being directly connected by an edge.

This is why I introduce a novel metric for the study of discord, better suited for complex networks: the *generalised active links density*, defined over all agent pairs by

$$\langle \rho \rangle = \frac{\sum_{i<j}(w^{\infty}_{ij} + w^{\infty}_{ji})\rho_{ij}}{\sum_{i<j}(w^{\infty}_{ij} + w^{\infty}_{ji})}. \tag{5.23}$$

Inspired by Estrada and Benzi (2014), w^{∞}_{ij} is the (i,j)-th component of the matrix exponential

$$e^{W} = \sum_{k=1}^{\infty} \frac{1}{k!} W^{k}. \tag{5.24}$$

Scaling by the inverse of the path length factorial attributes a rapidly decaying importance to longer path, as a way to account for all the combinatorial possibilities that j's opinion is overwritten on its route towards i. The sum $w^{\infty}_{ij} + w^{\infty}_{ji}$ is then a measure of long-range, weighted influence between i and j. The quantity $\langle \rho \rangle$ as defined by Eq. 5.23 is akin to $\langle \Gamma \rangle$ but incorporates long-range interactions.

Note that zero entries in the i^{th} row w_i^∞ correspond to users unable to reach i, and non-zero entries correspond to ancestors of i. The cosine similarity

$$cos(w_i^\infty, w_j^\infty) = \frac{w_i^\infty \cdot w_j^\infty}{\|w_i^\infty\| \|w_j^\infty\|} \tag{5.25}$$

informs us on the similarity of i and j's ancestry, meaning the extent to which they are exposed to the same channels of influence.

5.3.3 Echo chambers and opinion diversity in the complete network

Laying the ground for my applications in Section 5.5 and Section 7.1, I now precise the values of the AOD and ECE in the particular case of a complete network of N identical agents with two opinions $\mathcal{S} = \{0,1\}$. All links have the same weight w. Each agent receives the same influence from zealots $(z^{(1)}, \dots, z^{(S)})$ and I let $z = \sum_{s \in \mathcal{S}} z^{(s)}$. Via Eq. 5.1 this entails

$$w = \frac{1-z}{N-1}. \tag{5.26}$$

5.3.3.1 Opinion diversity

Every agent has the same distribution of opinions x, and Eq. 5.8 becomes

$$x^{(s)} = (N-1)wx^{(s)} + z^{(s)}, \tag{5.27}$$

so that

$$x^{(s)} = z^{(s)}/z. \tag{5.28}$$

This was already proven by Mobilia et al. (2007). In the case $S = 2$, I call $x^{(1)}$ the average equilibrium opinion of the network. Its value is closer to 0 when agents favour opinion 0, and closer to 1 otherwise. As all agents and links are identical, its

average value is immediately given by $z^{(1)}/z$ as per Eq. 5.28. The (AOD) is then

$$\langle \Phi \rangle = \frac{1}{N} \sum_{i \in \mathcal{N}} \Phi_i \tag{5.29}$$

$$= \frac{S}{S-1} \sum_{s \in \mathcal{S}} x^{(s)}(1 - x^{(s)}), \tag{5.30}$$

$$= \frac{2S}{S-1} \sum_{r<s} \frac{z^{(r)} z^{(s)}}{z^2}. \tag{5.31}$$

This AOD is then maximised when all $z^{(s)} = 1/S$ for all $s \in \mathcal{S}$. For $S = 2$, it becomes

$$\langle \Phi \rangle = \frac{4 z^{(0)} z^{(1)}}{z^2}. \tag{5.32}$$

5.3.3.2 Echo chamber effect

Discord probabilities are all equal to the same value ρ which is also the GALD $\langle \rho \rangle$, and Eq. 5.17 reduces to

$$\rho = \frac{1}{2} \left[2(N-2)w\rho + 2 \sum_{s \in \mathcal{S}} z^{(s)}(1 - x^{(s)}) \right] \tag{5.33}$$

$$= (N-2)w\rho + \sum_{s \in \mathcal{S}} z^{(s)} \left[1 - \frac{z^{(s)}}{z} \right] \tag{5.34}$$

$$= (1 - z - w)\rho + \sum_{s \in \mathcal{S}} z^{(s)} \left[\sum_{r \in \mathcal{S} \setminus \{s\}} \frac{z^{(r)}}{z} \right] \tag{5.35}$$

because of Eq. 5.26. Thus,

$$\rho = 2 \sum_{r<s} \frac{z^{(r)} z^{(s)}}{z(z+w)}. \tag{5.36}$$

This is very similar to (5.32), and again it is maximised when $z^{(r)} = 1/S$ for all $s \in \mathcal{S}$. Assuming it is the case, both ρ and the entropy of $(x^{(s)})_{s \in \mathcal{S}}$ converge to 1 as the size S of the opinion space goes to infinity: discord and diversity of opinion go hand in hand in this example. For $S = 2$ we have

$$\rho = \frac{2 z^{(0)} z^{(1)}}{z(z+w)}. \tag{5.37}$$

Then, the ECE $\langle\Gamma\rangle$ is simply given by

$$\langle\Gamma\rangle = 1 - \rho. \tag{5.38}$$

5.3.3.3 Comparison between the metrics

When $z \gg w$, *i.e.* $z \gg 1/N$, ρ becomes equivalent to $\frac{S-1}{S}\langle\Phi\rangle$. In this case, discord and opinion diversity is the same. This is not too surprising. Indeed (5.30) is a rescaled version of the discord between independent agent pairs (5.12), which as I remarked earlier as akin to an entropy of x.

The equilibrium opinions only depend on the relative values of $z^{(0)},\dots,z^{(S)}$ versus one another. Discord depends on those as well, but also on the total amount of zealots present in the system. Multiplying each $z^{(s)}$ by the same amount c will not impact $\langle\Phi\rangle$, while the new discord is given by

$$\langle\rho\rangle = \frac{2c(N-1)z^{(r)}z^{(s)}}{z(1+(N-2)cz)}. \tag{5.39}$$

This effect vanishes in the large z regime, where scaling it up begins to show diminishing returns.

5.3.3.4 Markov chain dynamics with stubborn agents

In Section 5.5 I use a description of the voting dynamics based on Markov chain modelling. This requires the computation of the exponential of the transition rate matrix, an operation that can be costly. This work was done in the first year of the PhD, and I have encountered since other strategies which in retrospect would have been more efficient. The method proposed here can be improved for more efficiency by using the voting dynamics as described in Section 5.3.3.1.

Consider a complete network of size N, and $\mathcal{S} = \{0,1\}$. Let $N_1(t)$ denote the number of nodes with opinion 1 at time t; it will be the quantity of interest. I fix $n_1 := N_1(0)$. I assume that some agents are stubborn and never change opinions. They form an inflexible core of partisans within a group who bear great power of persuasion over the whole population: politicians, journalists, lobbyists... These

agents impact the dynamics in a similar way as zealots do. Thus, when working in this context, I let $z^{(0)}$ denote the number of stubborn agents promoting opinion 0, and $z^{(1)}$ the number of stubborn agents promoting opinion 1.

Because $z^{(0)}$ and $z^{(1)}$ nodes will always be in respective states 0 and 1 no matter what, $N_1(t)$ is comprised between $z^{(1)}$ and $N-z^{(0)}$ for all t. The idea behind my analysis is that it describes a birth-and-death process over the opinion-space $\{z^{(1)},\dots,N-z^{(0)}\}$ with transition rates, for all $z^{(1)} \leq k \leq N-z^{(0)}$,

$$\begin{cases} q_{k,k-1} = (k-z^{(1)})(N-k)/(N-1) \\ q_{k,k+1} = k(N-k-z^{(0)})/(N-1) \\ q_{k,k} = -q_{k,k-1} - q_{k,k+1}. \end{cases} \tag{5.40}$$

Here, $q_{k,l}$ is the transition rate from state $\{N_1(t)=k\}$ to state $\{N_1(t)=l\}$. To move from state k to $k-1$ we need a non stubborn opinion-1 node to adopt the state of an opinion-0 node. There are $k-z^{(1)}$ non stubborn opinion-1 nodes and for each of these, a proportion $(N-k)/(N-1)$ of the others is in state 0, hence $q_{k,k-1} = (k-z^{(1)})(N-k)/(N-1)$. I obtain $q_{k,k+1}$ via an analogous reasoning and define $q_{k,k} = -q_{k,k+1} - q_{k,k-1}$. Since the process only evolves by unit increments or decrements, $q_{k,j} = 0$ if $j \notin \{k-1,k,k+1\}$. As expected we have $q_{z^{(1)},z^{(1)}-1} = 0$ and $q_{N-z^{(0)},N-z^{(0)}+1} = 0$. Finally I let $Q = [q_{ij}]_{i,j}$ denote the transition rate matrix.

From there I am able to compute the distribution of $N_1(t)$ and its expected value at any point in time. Indeed, the probability for N_1 to equal k at time t is

$$p_{n_1,k}(t) := [e^{tQ}]_{n_1,k}. \tag{5.41}$$

Hence,

$$\mathbb{E}N_1(t) = \sum_{k=z^{(1)}}^{N-z^{(0)}} k\, p_{n_1,k}(t). \tag{5.42}$$

is the expected number of opinion-1 nodes at time t.

5.4 Experiments on synthetic networks

I now provide a first insight on the behaviour of discord probabilities on synthetic networks. I analyse empirically how it depends upon both the topology of the network and the influence exerted by zealots. I also study social networks split between two antagonistic factions supporting different opinions.

5.4.1 Dependency between opinions

One may be interested only in certain values of ρ_{ij}, and wishing to avoid the burden of computing them all. In this case it may be tempting to use

$$\tilde{\rho}_{ij} = \sum_{s \in \mathcal{S}} x_i^{(s)} (1 - x_j^{(s)}), \tag{5.43}$$

as given by Eq. 5.12, even if σ_i and σ_j are not independent. While sometimes effective, this approximation does not always fare well—*cf.* Figure 5.1. As the dependency of i and j's opinions relies on the strength of paths joining them and the similarity of their ancestry, I expect ρ_{ij} to decrease with those, and the error made by $\tilde{\rho}_{ij}$ to increase. I verify this at equilibrium on the four toy datasets introduced in Section 4.4.3. Ground-truth communities being given, I set $z_i^{(s)}$ to a random uniform value if s is i's community, and to zero otherwise. To quantify path strength and ancestry similarity I use respectively $w_{ij}^{\infty} + w_{ji}^{\infty}$ and $cos(w_i^{\infty}, w_j^{\infty})$.

Results are shown in Figure 5.2, and confirm my hypotheses: ρ_{ij} decreases with the strength of paths joining (i, j) and the similarity of their ancestry, while the error made by $\tilde{\rho}_{ij}$ increases. Moreover, the error decreases with the total zealousness of i and j. This is not surprising, as higher values of $\|z_i + z_j\|$ mean lighter weights on inter-agent edges and thus less influence from peers. I also show the distribution of errors in Figure 5.3. The errors are quite low on average, but can peak very high for certain agent pairs (maximum error ranges from 15% for `football` to 187% for `polblogs`). The least accurate $\tilde{\rho}_{ij}$ is on `zachary`, probably because of the higher path strengths due to its smaller size.

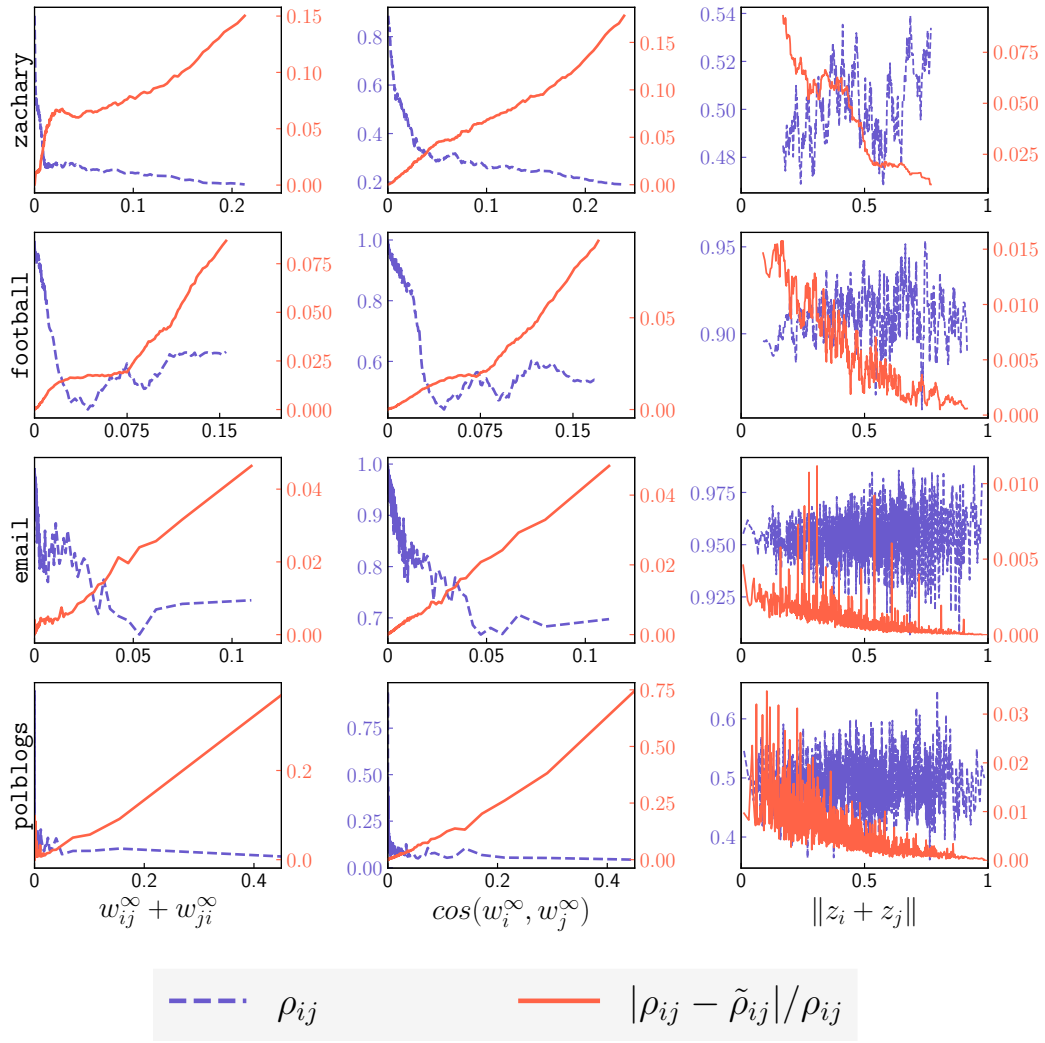

Figure 5.2: Comparison of ρ_{ij} and $\tilde{\rho}_{ij}$ on real-life datasets at equilibrium. X-axis, log scale: path strength $w_{ij}^{\infty} + w_{ji}^{\infty}$, ancestry similarity $cos(w_i^{\infty}, w_j^{\infty})$, total zealousness $\|z_i + z_j\|$. Y-axis, linear scale: moving average for ρ_{ij} (dotted blue line) and relative approximation error $|\rho_{ij} - \tilde{\rho}_{ij}|/\rho_{ij}$ (orange line) for all dependent agent pairs.

5.4.2 Echo chambers and opinion diversity with communities

I now analyse the ECE and the AOD in polarised networks split into antagonistic factions. I generate SBM graphs (*cf.* Section 2.2.3) with two communities $\mathcal{C}_0$ and $\mathcal{C}_1$, supporting opinion 0 and 1, respectively. I then calculate the equilibrium values of $\langle \Gamma \rangle$ and $\langle \Phi \rangle$, for various values of the model parameters. Namely, I fix the in-group link probability to $p_{\text{in}} = 0.1$ and vary the out-group link probability p_{out}. I consider different values for the zealousness of each community. I compute $\langle \Gamma \rangle$ and $\langle \Phi \rangle$ over the whole network and within each community, *i.e.* only considering in-group

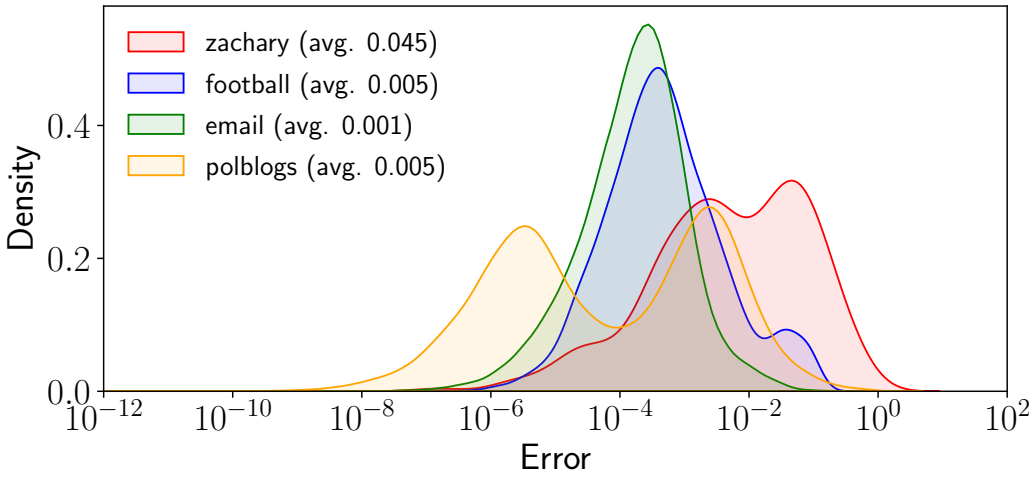

Figure 5.3: Distribution of the relative approximation error $|\rho_{ij} - \tilde{\rho}_{ij}|/\rho_{ij}$ on real-life datasets at equilibrium, obtained via kernel density estimation. Averages are indicated in the legend.

edges. Results are illustrated in Figure 5.4. In the submission to *Physical Review E* I focused on the generalised active links density, for which some plots are shown in Appendix C.

When reinforcing connections between different-minded communities, the novel paths of influence create more diverse information flows, and thus a higher exposition to contradicting beliefs. There are two main consequences one could expect from this. Agents may revise their viewpoint to incorporate adverse ideas, leading the system towards a more consensual state with higher ECE and lower AOD. Alternatively, they may fiercely cling onto their preexisting opinions, resulting in lower ECE and higher AOD—the so-called *backfire effect* (Bail et al., 2018; Schaewitz and Krämer, 2020). As we see now, both scenarios can happen.

Overall, when one community is much more zealous than the other $z = (0.1, 0.5), (0.1, 0.9)$, higher connectivity between them means higher echo chamber effect, and lower diversity of opinions (Figure 5.4, left plots). In other cases, the ECE tends to decrease as more incongruent connections are introduced. The behaviour of the AOD is interesting. For larger values of zealousness, it goes up until $p_{\text{in}} = p_{\text{out}}$, then goes back down. Thus in these cases, while it might be a good idea to create more links between communities, it is crucial to not overdo it. The only case where it

is guaranteed for opinion diversity to increase as I add more out-group links is when $z = (0.1, 0.1)$, that is when both communities are equally—and only lightly—biased.

The evolution of ECE within $\mathcal{C}_0$ exhibits the same trends (Figure 5.4, upper middle plot). When $\mathcal{C}_0$ is much less zealous than $\mathcal{C}_1$, for $z = (0.1, 0.5)$ and $z = (0.1, 0.9)$, adding out-group links can surprisingly increase the ECE within $\mathcal{C}_0$. While adding too few of them will reduce ECE in the community, once they reach a critical mass we observe an increase in $\langle \Gamma \rangle$. This stems from the fact that $\mathcal{C}_1$ has fiercer partisans, meaning opinion 1 gets more and more prevalent in the network as connectivity between the two communities goes up. Thus even within $\mathcal{C}_0$, holding opinion 1 guarantees more agreement with peers. This is also why, in community $\mathcal{C}_1$, the ECE always decreases and the AOD mostly increases—except for high values of z and of p_{out} (Figure 5.4, right plots).

However, this prevalence of opinion 1 negatively impacts the AOD in $\mathcal{C}_0$, as it mostly decreases when there are too much out-group links (Figure 5.4, lower middle plot). Thus, again, it is important to not add too many out-group links if we are looking to increase AOD and reduce ECE. We can have low ECE and low AOD at the same time, which highlights the importance of distinguishing between the two in general settings. Indeed for complete graphs, as I prove in Section 5.3.3 below, AOD is the opposite of ECE and such phenomenons cannot take place.

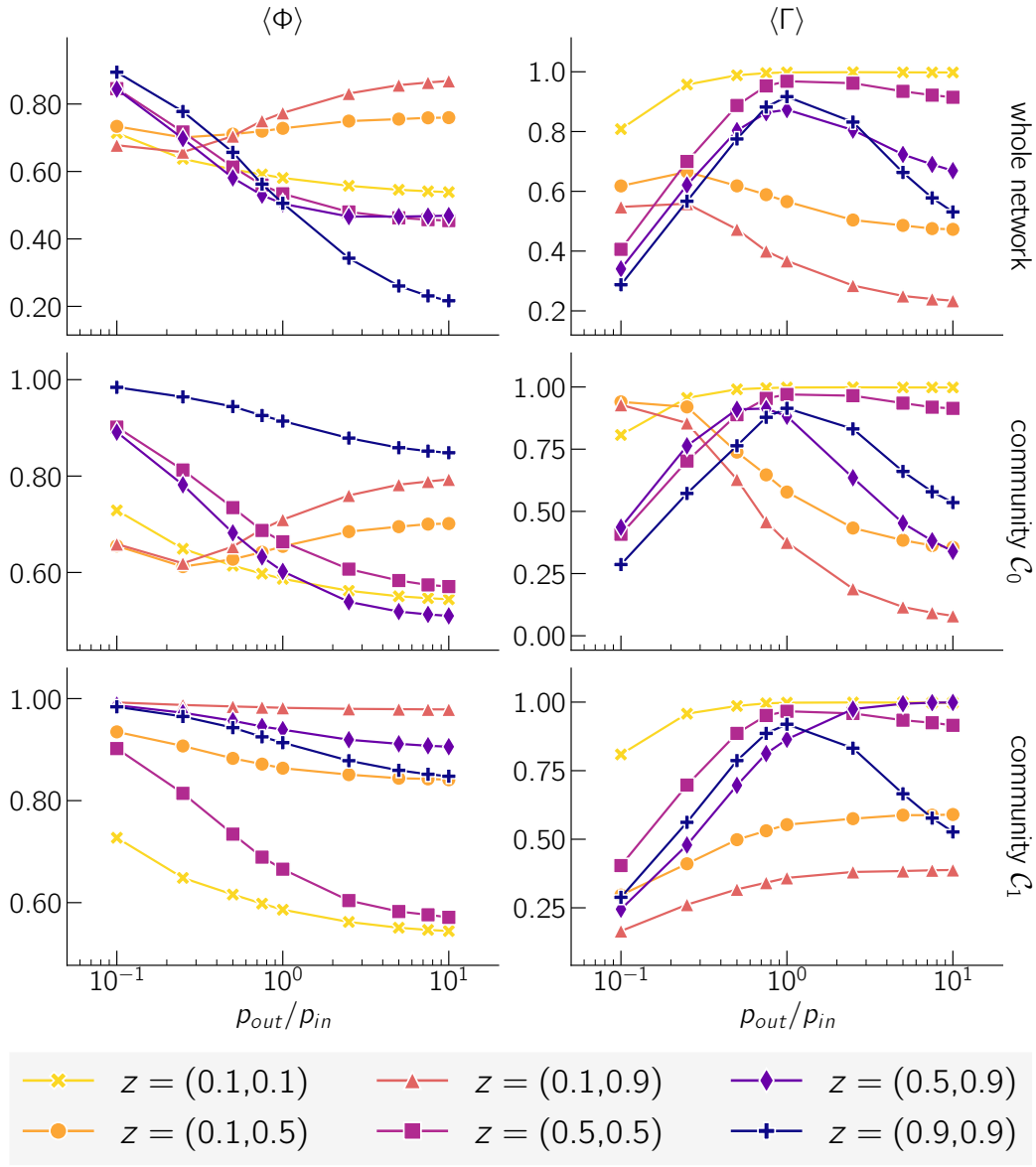

Figure 5.4: Network of $N = 100$ agents with two communities $\mathcal{C}_0$ and $\mathcal{C}_1$. The s-zealot exert a total influence $z^{(s)}$ on each agent in $\mathcal{C}_s$ and none on others. In-group link probability is fixed at $p_{\text{in}} = 0.1$, while the out-group link probability p_{out} and zealousness $z = (z^{(0)}, z^{(1)})$ vary. Results are averaged over 20 agent graphs generated under the Stochastic Block Model, for the whole network (top) and within each community (middle, bottom). **Left:** ECE. **Right:** AOD. The plots have different scales for clarity, but the purpose is to focus on the qualitative dynamics rather than exact values.

5.5 Predicting elections results with the EV Model

My goal is to forecast the results of general elections in the UK and presidential elections in the US, via the theory developed in Section 5.3.3.4 for the EV Model.

I am interested in the percentage of popular votes won by the two major parties – Conservative and Labour in the UK, Republicans and Democrats in the US. The datasets are presented in Section 4.4.1. I assume these quantities correspond to pointwise observations of independent realisations of the EV Model on a complete network with $N = 100$ nodes. I assume some of the voters are stubborn and always vote for the same party, there are $z^{(s)}$ such partisans for party s. The result of each election can then be forecast via Eq. 5.42, provided I have an estimate of the quantity of stubborn nodes $z := (z^{(0)}, z^{(1)})$. Thus, my analysis is done in two steps: first I make for each elections an estimate of z based on previous results, then Eq. 5.42 gives me the expected value for the coming election that I use as a predictor.

5.5.1 Setting

I present my method in the UK case, but note that it directly translates to the US case by replacing Conservative and Labour with Republicans and Democrats. Because the model does not account for decimal values values I round the percentages to the nearest integer. Different parties are present, but because my model applies to a two-sided situation only, I cannot consider all of them at once. Thus, I aggregate all non-Conservative parties under the label 0 while Conservatives are attributed label 1. I let x_i denote the number of seats won by Conservatives on the i^{th} elections and t_i the elapsed time, in years, since the starting point 1922. There have been $M = 27$ elections total, with the last one taking place in 2019. Thus $t_1 = 0$ and $t_M = 2019 - 1922 = 97$. I let x_M denote the percentage of votes won by the conservatives in 2019. To concur with my theoretical framework I consider one seat won by the Conservatives (resp. non-Conservatives) as the observation of an node being in state 1 (resp. 0) amongst $N = 100$ of them. The x_i's then correspond to pointwise observations at times t_i's of a realisation of the process $N_1(t)$. All the reasoning described here and in the following will also be applied independently in the cases Labour versus non-Labour, Republican versus non-Republican (US) and Democrat versus non-Democrat (US).

5.5.2 Estimation of z and forecast

To be able to use Eq. 5.42 to make predictions, I first need to estimate the proportion of potential stubborn nodes in the population, that is the percentage of votes which are guaranteed for or against Conservatives. Let $z^{(0)}$ denote the number of stubborn opinion-0 (non-Conservative) nodes and $z^{(1)}$ that of opinion-1 (Conservative) ones. I look for the values $(z_\star^{(0)}, z_\star^{(1)})$ that maximise the log-likelihood of the observed data. Let's say I want to predict results for the i^{th} election. Because I need at least two datapoints to make an estimation, I require $3 \leq i \leq M+1$. Let $p_{k,l}^{(z^{(0)},z^{(1)})}(t)$ denote the theoretical probability for $N_1(t)$ to go from k to l in t units of time when there are respectively $z^{(0)}$ and $z^{(1)}$ opinion-0 and opinion-1 stubborn nodes. I seek to solve

$$\underset{z^{(0)},z^{(1)}}{\operatorname{argmax}} \sum_{j=1}^{i-2} \log\left(p_{x_j,x_{j+1}}^{(z^{(0)},z^{(1)})}(t_{j+1}-t_j)\right) \tag{5.44}$$

Indeed, $p_{x_j,x_{j+1}}^{(z^{(0)},z^{(1)})}(t_{j+1}-t_j)$ is by definition the probability for Conservatives to win x_{j+1} percent of the votes in the $(j+1)^{\text{th}}$ election knowing they won x_j percent in the j^{th} one. Thus I seek to simultaneously maximise the likelihood of all past elections results. Let $Q^{(z^{(0)},z^{(1)})}$ be the matrix with entries calculated via (5.40). By Eq. 5.42, we have that (5.44) is equivalent to

$$\underset{z^{(0)},z^{(1)}}{\operatorname{argmax}} \sum_{j=1}^{i-2} \log\left[\exp\left((t_{j+1}-t_j)Q^{(z^{(0)},z^{(1)})}\right)\right]_{x_j,x_{j+1}} \tag{5.45}$$

The computation of the matrix exponential is typically done in cubic time and quickly becomes intractable as the size of the matrix increases. Here however, because $N = 100$, the number of possible couples $(z^{(0)}, z^{(1)})$ is small enough here that (5.45) can be solved by directly computing the sum for each of these couples individually. The optimal value $z_\star^{(1)}$ for $z^{(1)}$ then gives me an estimation of the percentage of votes "locked" by the Conservative party, proportion of the population that will always root for them. The optimal value $z_\star^{(0)}$ for $z^{(0)}$ is an estimate of the quantity of such votes for all other parties aggregated.

To make a forecast for the i^{th} election, I just have to apply Eq. 5.42 with

$Q = Q^{(z_\star^{(0)}, z_\star^{(1)})}$, $n_1 = x_{i-1}$ and $t = t_i - t_{i-1}$. Eq. 5.42 then gives me the expected percentage $\tilde{x}_i$ of votes gathered by Conservatives on that occasion. This can then be compared to the actual value x_i to assess the efficacy of my approach. I also calculate the standard deviation, to see how far the datapoints fall from the theoretical predictions.

5.5.3 Notes on the methodology

I set the number of agents to $N = 100$, meaning that a result of 54.7% in the data is translated to $N_1(t) = 55$. Thus, I lose in precision, and would probably obtain better results with a higher value of N. However, this would significantly slow down the optimisation process described by Eq. 5.44-5.45. For each possible pair of $(z^{(0)}, z^{(1)})$ I must compute the matrix exponential of $Q^{(z^{(0)}, z^{(1)})}$, which is done in cubic time. This is still manageable with $N = 100$, but would become way too long with higher values. Future research could look into ways to accelerate this process.

I set the time scale such that one time unit in the Voter Model corresponds to one year in real life. Thus, each agent is supposed to update their opinion once each year, on average. I did try a few other values in the range $[10^{-1}, 10^{1}]$, but obtained worse results. There is no reason to believe that the chosen value is optimal however, and a more thorough search might reveal more adapted values.

My prediction relies on the use of stubborn agents. I assume some of the N agents are strongly in favour of one party against the other. The first step in my prediction method, that is the optimisation described by Eq. 5.44-5.45, is dedicated to inferring the number of stubborn agents. The inference process is dynamical, that is after each election I adjust my estimation of $(z^{(0)}, z^{(1)})$. That is because, I believe this quantity can evolve over time, especially as the results of elections often reshape the political landscape of a country. Moreover, the results of this inference not only serves the prediction, it is also by itself a valuable information. Knowing the "non-swing" voters, that is the proportion of the population backing each party no matter what, gives precious insights on the political landscape of the country. Future research could also discard the possibility of stubborn agents, and use Eq. 5.7 to predict the evolution.

To summarise, Table 5.1 describes the correspondence between the model features and reality.

Table 5.1: Correspondence between model features and reality, and chosen values for the parameters.

Model	Reality	Value
Number of agents N	Size of the voting population	100
Time unit	Average number of opinion changes per voter per year	1
Stubbornness values $z^{(0)}, z^{(1)}$	Number of unconditional supporters for party 0, for party 1	-
Evolution of $z^{(0)}, z^{(1)}$	The result of elections affects the number of unconditional supporters	-

5.5.4 Results for the UK

I show in Table D.1 (Appendix D, left) the estimated values for $(z_\star^{(0)}, z_\star^{(1)})$, updated with each new election. They seem to globally stabilise between 15 and 25 for both parties. Look at the last value in the Labour case for example, which is $(24, 15)$. According to my model, this means there is an estimated proportion of 15% of voters that will *always* vote Labour. On the other side, 24% of voters are found to be stubborn "anti-Labour" – by that I don't mean that they are fundamentally against the Labour party but rather that they will *never* vote for it. Note that these estimates fluctuate according to the variability of the data. For example in 1922 and 1923 there were twice in a row 38% votes for Conservative[1] and as a result it was estimated that 38% of all voters are stubborn pro-Conservative and the 62% are stubborn against the party. This is indeed what maximises the likelihood, with this configuration yielding a probability of one for the observed values. On the other hand, with pro-Conservative votes jumping from 38% to 61% in 1935, estimated values of $z^{(0)}$ and $z^{(1)}$ dropped significantly to account for the wide range covered by the data.

In Figure 5.5 (top plots) I compare my predictions, that is the expectations $\tilde{x}_i$, with the real outcomes x_i. I plot both values for each election starting with the

[1] Remember that those value are rounded to the nearest integer to fit the needs of my model—the actual results were 38.5% and 38%.

third one that took place in 1924, because the optimisation problem (5.45) requires $i \geq 3$. For both parties, most values seem to fluctuate around the 40% mark. The global tendency of the real outcomes looks respected by the predictions, albeit with less variability. Also note that most real values appear to fall within one standard deviation from the predictions.

To get a better insight I look at the absolute errors $|\tilde{x}_i - x_i|$ of my predictions. I plot running averages over the last 5 elections in Figure 5.6 (top). After a few erratic first years they seem to stabilise between 2% and 8%. More precisely, if I discard the first few years up until 1960 where the model lacks sufficient amount of data to properly calibrate, I get MAEs of respectively 4.63% and 5.23% for Conservative and Labour. Minimal values of 0.06% for Conservatives in 1979 and 0.40% for Labour in 2001 are observed, showing that my method was able to make very accurate predictions in these cases. Surprisingly however, the errors do not seem to monotonically decrease over time, but rather fluctuate. As a matter of facts, peak absolute errors were observed in 1983 (Labour, 13.0%) and 1997 (Conservative, 13.6%).

5.5.5 Results for the US

I apply the exact method described above to the case of presidential elections in the United States. As before I independently consider two cases, Republicans versus non-Republicans and Democrats versus non-Democrats. Presidential elections in the US take place every four years and I start with the year 1912, then 1920, 1924, and so on. Here again, keep in mind that due to how the American system work, the party with the most popular votes does not necessarily win the elections. The first estimation I am able to make is based on the first two elections and thus my first prediction is for 1924.

I observe similar results as in the UK case. Stubborn values (Appendix D, right) estimated $(z_\star^{(0)}, z_\star^{(1)})$ are close, albeit a little bit lower – stabilising at (18,17) for Republicans and (16,14) for Democrats. Again, a majority of the real values fall within one standard deviation of the prediction (Figure 5.5, bottom plots). The prediction curves also looks more stable than the slightly spiky ones with real

values. Note that because of the two-party system in place in the United States, both Republicans and Democrats see their share of popular votes fluctuate around the 50% mark. In the previous case, it was rather around 40% because of the space occupied by smaller parties such as Liberal Democrats or Scottish National Party amongst others. The two-sided aspect of my model – always one party versus another – may thus be more adapted to the study of the US system.

As for the errors, running averages over the last 5 elections are shown in Figure 5.6 (bottom). Here again after a few erratic first years values appear to be comprised between 2% and 8%. However, where errors in the UK case seemed to increase in the last few years, here they to are dropping down. In fact, my most accurate forecast regarding Democrat votes is for 2016, with only 0.04% error. For Republicans it is in 1940 with 0.10%. Peak errors were again around 13% for both parties, in 1972 (Republicans, 14.0%) and 1964 (Democrats, 12.3%). The MAE over all elections, starting in 1940 when forecasts start to stabilise, is 4.27% for Republicans and 4.83% for Democrats. This is slightly better than in the UK case (4.63% and 5.23%). The MAE error over both cases is then 4.74%.

5.5.6 Comparison with other methods

I compare these errors with those obtained using a naïve prediction method, and a baseline. The naïve method predicts for an election the same result as the previous one. The baseline method predicts a random results amongst all past ones.

For the UK, The naïve method yields an average errors 4.75% and 3.44% for Labour and Conservative, respectively. It is thus more effective than my method in this case. The baseline fared worse, with average errors of 7.53% and 7.24%. For the US, both alternatives yielded worse results than my method, with errors of 5.35% (naïve, democrats), 6.60% (naïve, republicans), 8.58% (baseline, democrats) and 9.48% (baseline, republicans).

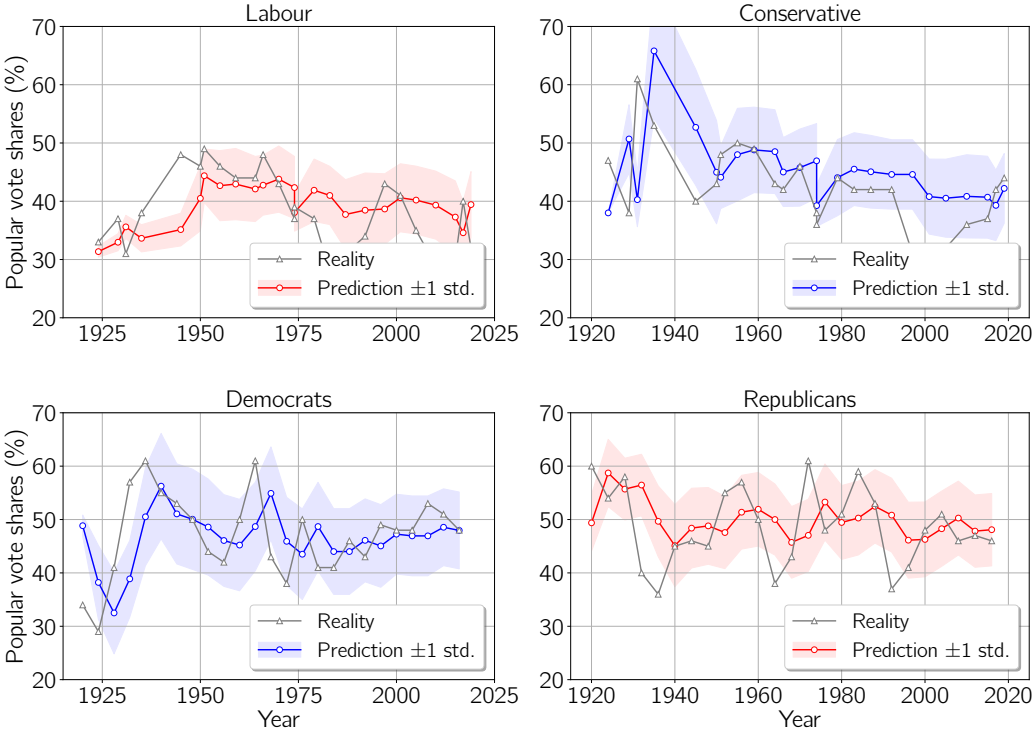

Figure 5.5: Percentage of votes for each party over years, prediction and reality. The prediction is computed as the average of the distribution of $N_1(t)$, obtained via Equations 5.45 and 5.42. Shaded areas cover one standard deviation from the mean. **Top:** United Kingdeom. **Bottom:** United States.

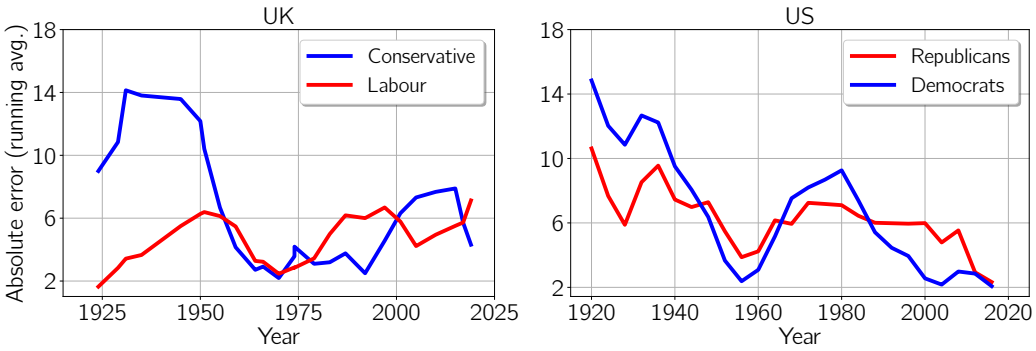

Figure 5.6: Absolute error made by the model, running average over the last 5 elections. **Left:** UK. **Right:** US.

5.6 Discussion

In this chapter, I analysed the Enhanced Voter Model, that generalises the traditional voter model. I first provided some novel theoretical results in a general context, applicable to many complex settings. Then, I evaluated the accuracy of the model prediction in a basic case of elections forecasting.

5.6.1 Theoretical findings

I expressed the metrics of interest, the ECE and the AOD, in this context. Importantly, I demonstrated how to compute discord probability between any two agents. In some cases the opinion distributions are independent and the calculation is straightforward. Otherwise, these probabilities are solution of a large linear system of differential equations. Laying the ground for my method to steer the echo chamber effect, I computed the exact ECE and AOD in complete networks.

Knowing discord probabilities allows for a precise computation of the active links density, a widely studied order parameter of the voter model for which no general formula was known. I extended its definition and proposed the generalised active links density, to account for long-range, weighted interactions.

Through experiments on toy datasets, I showed that, when agents are closely connected or share ancestors, *(i)* discord probabilities diminish, and *(ii)* the approximation errors made by assuming independent distributions augment. Highly zealous pairs of agents were also more forgiving in the second case, yielding low approximation errors.

I also analysed the evolution of the ECE and the AOD in polarised networks split into antagonistic factions. Interestingly, I observed that adding links between groups can sometimes reduce ECE and increase AOD, and sometimes the opposite. Using such strategies in real life should thus be met with precautions, in order to not entail adverse effects. Moreover, except for simple cases such as the complete graph, the behaviour of ECE and AOD is not always symmetrical. This highlights the importance of using both metrics in complex settings.

5.6.2 Empirical evaluation

To predict election results, I considered official results of past elections as observations of independent realisations of the voter model on a complete network. From there I was able to perform time-evolving estimates of the model parameters and use them to forecast an outcome. My model yielded an MAE of 4.74%, reaching absolute errors as low as 0.04% and as high as 14%.

In their review, Gayo-Avello (2013) suggest that any model used to predict

the elections outcome should not have an MAE higher than 1% or 2%. This is because the result of an election is more often than not the matter of just a few percents. According to this standard, my MAE is not low enough to reliably predict the outcome of an election. Moreover, my method performed worse than both a naïve and a baseline method in the case of the UK. The results were better for the US however, which might be due to the stricter two-party nature of the system.

Although my method did not yield significant enough results here, I believe it is an interesting step in a novel direction. First of all, it only relies on official data. Second, my model does not only forecast the elections results, it also gives me estimates of the proportion stubborn voters, that is the proportion of individuals who will *always*—or *never*—vote for the considered parties. This provides meaningful insight on the political landscape of the considered areas.

Several extensions of the model could be considered to improve its accuracy. First of all, the number of agents N and the time unit could be adjusted for better precision. Second, dynamical estimates could be made without accounting for the presence of stubborn agents. Third, adding in-between election polls to the data would go a long way in improving the estimates. With a few years gap from one election to another, it is too wide a range of possibilities for the model to account for. Fourth, one could take a deeper look into the past of a country's results and try to detect tendencies about landslide victories, incumbency reelection and so forth. I believe that having a deeper understanding of the specific country one is working with could substantially improve the model calibration process. Finally, combining my method with Twitter data-based estimations may lead to higher accuracy.

Chapter 6

The Extended Newsfeed Model for opinion diffusion in online social platforms

I extend a previous model from Giovanidis et al. (2021, 2019), that describes the diffusion of content throughout an OSP. The original works were interested in user-to-user influence. Here, I introduce opinions and calculate the AOD and ECE, as described in Section 4.1. I also propose a novel feature, *preferential reposting*, that improves the accuracy of the model when compared with real data. I perform an empirical evaluation of the `#Elysée2017fr` dataset. This was published in Vendeville et al. (2023a). Except from Section 6.1 where I introduce the original model, all the material in this chapter is a novel contribution from this PhD.

6.1 General setting

I now briefly describe the general setting, as proposed in the original model. Consider N users who interact on a social platform. I let $\mathcal{L}_i$ denote the set of all leaders of user i, that is users that i follows. Each user repeatedly creates new content (*self-posts*) or spread content created by others (*reposts*). The newsfeed of a user contains posts propagated by their leaders. All newsfeeds are of finite size M. I make the following fundamental modelling assumptions:

Markovian activity User n is endowed with two exponential clocks of respective

parameters λ_n and μ_n. Whenever the first one rings they create a new self-post, while the second one prompts them to visit their newsfeed and select an item to repost amongst all content available there.

Random selection The selection of an item to repost when visiting the newsfeed is made uniformly at random.

Random eviction Any new entry into a full newsfeed will evict an older one chosen uniformly at random.

These assumptions are made for the sake of analytical tractability, and it is shown in Giovanidis et al. (2021) that they can be relaxed.

6.2 Introducing opinions

In the present extension, I add the labelling of posts by the opinion they express. To do so, I assume that user n produces self-posts expressing opinion s at rate $\lambda_n^{(s)}$, so that $\sum_{s\in\mathcal{S}} \lambda_n^{(s)} = \lambda_n$. In other words a proportion $\lambda_n^{(s)}/\lambda_n$ of all self-posts from n expresses opinion s. I call s-post a post expressing opinion s.

Let $p_n^{(s)}$ denote the average proportion of s-posts on the newsfeed of n at equilibrium. It is obtained via

$$p_n^{(s)} \sum_{k\in\mathcal{L}_n} (\lambda_k + \mu_k) = \sum_{k\in\mathcal{L}_n} (\lambda_k^{(s)} + \mu_k p_k^{(s)}). \tag{6.1}$$

Assuming the user graph is strongly connected and at least one user has $\lambda > 0$, the system has a unique solution (Giovanidis et al., 2021). Equation 6.1 is a balance equation that equates the input and output rates of s-posts on the newsfeed of n induced by the activity of its leaders. On the left-hand side is the rate at which s-posts are evicted at random from the newsfeed to be replaced by fresher content. On the right-hand side is the rate at which s-posts propagated by the leaders of n enter the newsfeed. It decomposes in the self-posting rates of leaders of n about party s, and the reposting rates of content s from leaders of n, through their newsfeeds. Note that $p_n^{(s)}$ is the probability to uniformly select an s-post at equilibrium.

To efficiently compute (6.1), I rely on the same iterative algorithm as for (5.21) in the previous chapter. The convergence rate depends on the spectral radius of the matrix defining the linear system. A proof of convergence is found in (Giovanidis et al., 2021, Thm. 4).

6.2.1 Echo chamber effect and opinion diversity

Let me define for every user i,

$$\mathcal{S}_i = \{s \in \mathcal{S} : \lambda_i^{(s)} > 0\}. \tag{6.2}$$

It is the set of all opinions that user i agrees with and posts about. I let S_i denote its size. To quantify the ECE for i I use:

$$\Gamma_i = \frac{1}{S_i} \sum_{s \in \mathcal{S}_i} p_i^{(s)}, \tag{6.3}$$

which is the average proportion of opinions on the newsfeed of i that they can agree with. I let $\langle \Gamma \rangle$ denote the average of this quantity over all users.

I quantify the AOD via the Φ-score:

$$\Phi_i = \frac{S}{S-1} \sum_{s=1}^{S} p_i^{(s)} (1 - p_i^{(s)}). \tag{6.4}$$

A value of 0 indicates that the newsfeed of i only contains a single opinion, describing a perfect echo chamber. On the other hand when $\Phi_i = 1$ all opinions are equally represented on the newsfeed with the same average proportion of $1/S$. Again, $\langle \Phi \rangle$ will denote the average of (6.4) over all users.

Note that I could have chosen to measure opinion diversity via the entropy $-\sum^{(s)} p_i^{(s)} \log p_i^{(s)}$ of the newsfeed of i—and similarly for the EVM. Its use is ubiquitous and its extrema lie in the same place as Φ_i. However the latter has the nice property of being quadratic in p_i which allows for considerably more efficient optimisation.

6.2.2 Reposting preferences

For higher accuracy, Vendeville et al. (2023a) introduces reposting preferences: when user n visits their newsfeed, they choose what to repost not uniformly at random but with certain probabilities, that describe their preferences towards certain type of content. This seems like a natural feature to incorporate, and indeed the retweet graph exhibits sparser connections between parties than the follow graph (Figure 4.1).

The probability for n to repost an s-post is set to $\nu_n^{(s)} = \lambda_n^{(s)}/\lambda_n$. Unfortunately, this makes the mathematical analysis more complicated, and I am not able to derive analytical formulas for the steady-state. Indeed, Eq. 6.1 becomes

$$p_n^{(s)} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) = \sum_{k \in \mathcal{L}_n} \left(\lambda_k^{(s)} + \mu_k \frac{\nu_k^{(s)} p_k^{(s)}}{\sum_{r \in \mathcal{S}} \nu_k^{(r)} p_k^{(r)}} \right), \quad \forall n, s. \tag{6.5}$$

The system is not linear anymore, and it will be more difficult to solve optimisation problems that rely on it. It is also not obvious how to treat the case where the denominator in the right-hand side is zero, *i.e.* the user has no interest for items that appear on their newsfeed. This is why, I restrict myself to the case with non-preferential reposting. I will however present some results obtained with preferential reposting via simulations. Note that, as we see in Figure 6.1, with reposting preferences the size M of the newsfeeds has an impact on the steady-state of the system. This impact is salient for $M = 1, 2$ but diminishes with higer values. In my experiments with reposting preferences I usually set $M = 5$ or $M = 10$.

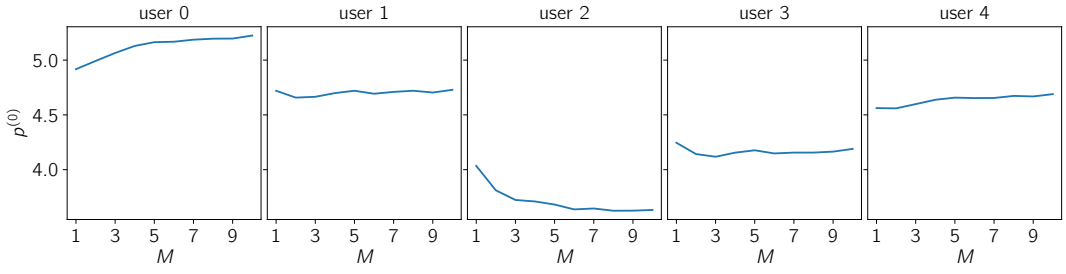

Figure 6.1: Impact of the newsfeed size M on the steady-state of the system when using preferential reposting. Complete network with $N = 5$ and two opinions 0 and 1. Activity rates are chosen uniformly at random in $[0, 1]$. For each M I run a simulation with 10^5 events and compute estimates of $p^{(0)}$ by averaging over the last 9×10^4 events. I skip a reposting event by user n if no item on their newsfeed has $\nu_n^{(s)} > 0$. I plot $p^{(0)}$ per user.

6.3 Simulating the model

I describe how to obtain values of $p_n^{(s)}$ for all n, s via a single simulation. To simulate the model, I start with newsfeeds and walls filled with content labeled uniformly at random. Let $news(n)$ be a S-size vector, that contains the initial proportion of posts with each label on the newsfeed of n. I initialise p_n as a zero vector of size S.

6.3.1 Base algorithm

At each step, I draw $2N$ random exponential variates $X_1, \dots X_N, Y_1, \dots, Y_N$ with parameters $\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_N$. The waiting time before the next self-post of user n is given by X_n (*resp.* repost, Y_n). Overall, the next event will happen in

$$dt := \min \{X_1, \dots X_N, Y_1, \dots, Y_N\} \tag{6.6}$$

time units. If $dt = X_n$, I create a new post. The label of this post is s with probability $\lambda_n^{(s)}/\lambda_n$. If $dt = Y_n$, I select a post at random from the newsfeed of n to be reposted. The freshly created or reposted piece of content is then inserted in the newsfeeds of all followers of n. Each insertion replaces an older post, chosen uniformly at random, and I update *news* accordingly. Then I update p as follows:

$$p_n \leftarrow p_n + dt \times news(n). \tag{6.7}$$

At the end, I divide p_n by the total simulation time to obtain its final value. The simulation must be long enough for the system to reach equilibrium, and earlier steps must be discarded from the computation. Typically I simulate $N \times 10^k$ steps and discard the first $N \times 10^{k-1}$.

I do not study the convergence time here, however given the form of Equation 6.1, I believe it must be related to the spectral radius of the matrix describing the linear system. Simulations for the original model (Giovanidis et al., 2021) exhibited a difference between simulated averages and equilibrium values that decreases exponentially with time, and was about 1.5% after $N \times 10^7$ simulation steps.

6.3.2 Extensions

The algorithm is straightforwardly adapted to account for other selection and eviction policies, such as studied in Giovanidis et al. (2021). With preferential reposting, I do not use a uniform distribution when selecting an item to repost amongst those on a newsfeed. Instead I choose to repost an item labeled s with probability ν_s. When n visits their newsfeed and find no item labeled s such that $\nu_n^{(s)} = 0$, they do not repost anything. In Section 7.2.3 I will need to account for personalised recommendations. To simulate user n receiving recommendations supporting party s at rate $y_n^{(s)}$, I simply draw an additional exponential variable of parameter $y_n^{(s)}$ at each iteration. When it realises the minimum, I insert a post labeled s in the newsfeed of n.

6.3.3 Memory-less property

Each of $X_1, \dots X_N, Y_1, \dots, Y_N$ gives me the waiting time before a future event. Storing them all in memory, and rearranging the list every time a new event is drawn, is computationally cumbersome. The memory-less property of the exponential distribution provides a work-around, allowing me to simply draw $2N$ values, keep the minimum, and re-draw at the next iteration. Indeed, if X is distributed under the exponential law, it holds that:

$$\mathbf{P}(X > s+t | X > s) = \mathbf{P}(X > t). \tag{6.8}$$

This memory-less property means that the past of X does not affect its future behaviour. The time we expect to wait before the occurrence of an exponentially distributed event is independent of how long we have already been waiting. In our context, the occurrence of the next posting or reposting event is not influenced by the previous ones. I refer the interested reader to Norris (1997, Theorem 2.3.1 and Section 5.2) for a proof of Eq. 6.8, and a deeper analysis of how this property applies to the simulation of stochastic processes.

6.3.4 Example dynamics

To observe the evolution of the system in a simple case, I simulate the model

dynamics on the `Zachary` dataset—*cf.* Section 4.4.3. Each user is assigned activity rates at random. There are two ground-truth communities, I set $\lambda_n^{(s)}$ to a random uniform number in $[0.5, 1]$ is n is in community s, and to a random uniform number in $[0, 0.5]$ otherwise. I set the newsfeed size to $M = 10$ and perform two simulations: one without preferential reposting, and the other with. I plot the evolution of $p^{(0)}$ for each user in Figure 6.2.

Unsurprisingly, in both cases users in community 0 almost always have a higher $p^{(0)}$ than those in community 1. This effect is more pronounced in the preferential reposting scenario, are users are even less likely to propagate outside content in their community. Moreover, users with more outside connections have values of $p^{(0)}$ closer to 0.5, which reflects the fact that they are exposed to more diverse content.

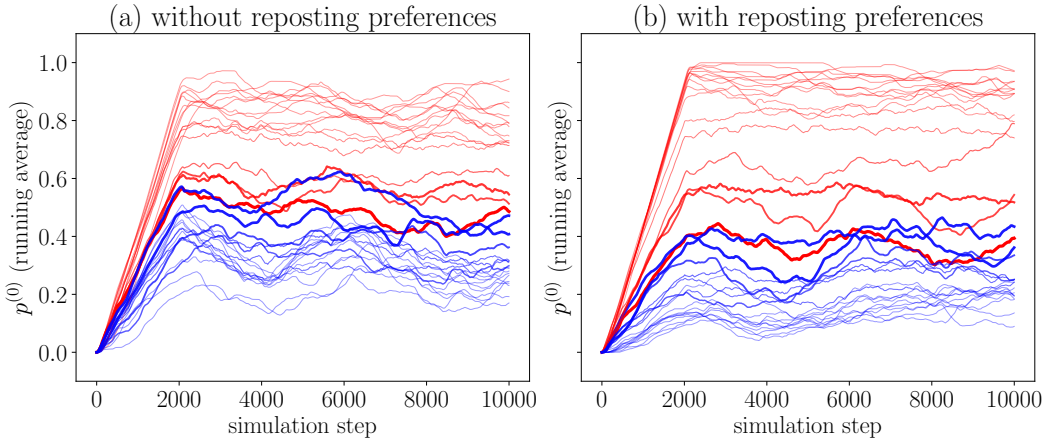

Figure 6.2: Simulation of the `Zachary` dataset. I perform 10^4 simulation steps and compute running averages of $p^{(0)}$ for each user, with a window size of 2,000. Red lines show the running averages for users of community 0, and blue ones for users of community 1. Thicker lines indicate users with a higher number of connections with the other community. **Left:** without reposting preferences. Final average $p^{(0)}$ in community 0: 0.65, in community 1: 0.36. **Right:** with reposting preferences. Final average $p^{(0)}$ in community 0: 0.73, in community 1: 0.16.

6.4 The Newsfeed model on OSP data

I now compare the equilibrium the system as predicted by the model with empirical estimates made on the `#Elysée2017fr` dataset. I refer to Section 4.4.2 for a description of the dataset. Are given the follow graph, and a list of posts and reposts with timestamps and user ids (the *trace*). This allows me to infer the values of λ and μ.

For the Φ-score I am concerned with opinions, thus I additionally require knowledge of the political leanings of users, which is given in `#Elysée2017fr`. First I explain how to infer the model parameters, that can be used to derive theoretical values of p. Then I explain how to make empirical estimates of p. These theoretical and empirical values are then compared, and I derive corresponding values of Φ.

6.4.1 Parameters inference

To compute the p I need, for every user n: posting rate λ_n, reposting rate μ_n, and the list of leaders $\mathcal{L}_n$. The latter is given by the *follow* graph. I estimate λ_n (*resp.* μ_n) as the total number of tweets (*resp.* retweets) posted by n divided by the duration covered by the dataset. To compute the Φ-score I additionally need to know how is λ_n distributed over the parties $\mathcal{S}$. How to estimate this distribution may differ depending on the precision level of political leanings available. In `#Elysée2017fr` I am provided with the set $\mathcal{S}_n$ for each user, of size one or two. Then, I set

$$\lambda_n^{(s)} = 0 \text{ if } s \notin \mathcal{S}_n, \text{ else } 1/|\mathcal{S}_n|. \tag{6.9}$$

In practice people may post about other parties than those they support, and I leave to future research the estimation of more precise labels based on richer features. Once the parameters are known, one can use (6.1) to compute the newsfeed distributions p.

Most of the time, the data I have access to is incomplete. It only covers a certain period of time, and some tweets and retweets might be missing. In consequence, the estimated parameters are error-prone and most of the time would not match those obtained with an extended or reduced version of the dataset. Ideally, I would like to know how high these errors are. To evaluate them and thus have an idea of the accuracy of the parameters I estimate, I can perform several parameter estimations on different subsets of the data and compare the results. I did not do so here because of time constraints, but further research shall take that into account.

6.4.2 Empirical estimation of Φ-score

In order to estimate empirical values of Φ, I use a method (hereafter, the *emulator*) developed in Giovanidis et al. (2021, 2019) that uses the trace to estimate p. I first label each tweet by the affiliation of its creator. If this information is not available I label by '?'. Because the size of the newsfeeds does not matter according to the model (Giovanidis et al., 2021), I assume for simplicity that all users have newsfeeds of size 1. I do not know the initial content of each newsfeed and assume they all contain a post of unknown origin, labelled '?'. As soon as a user tweets or retweets something, the post is inserted into the newsfeeds of their followers, evicting any previous post that was there. Some users are affiliated to two different parties, so that the corresponding label is two-sized: (s, s'). If during a period of time the newsfeed of n contained a post with such a label, I assume that half the time the newsfeed contains a post labeled s and the other half a post labeled s'. Finally to obtain $p_n^{(s)}$ I compute for each user n and label s the proportion of time their newsfeed contained a post labelled s. I disregard periods during which the newsfeed contained a post labelled '?'.

6.4.3 Opinion distribution on newsfeeds and echo chamber effect in `#Elysée2017fr`

I evaluate the accuracy of theoretical values for the Φ-score (6.4), made via Eq. 6.1, against empirical estimates. I use the `#Elysée2017fr` dataset described in Section 4.4.2. Theoretical and empirical values of p are derived as described in Section 6.4.1 and Section 6.4.2. Again, model parameters are inferred via Section 6.4.1, and I obtain empirical estimates of p via the emulator described in Section 6.4.2.

6.4.3.1 Opinion distribution on newsfeeds

I display in Figure 6.3 the values of $p_n^{(s)}$ and Γ_n from both the empirical evaluation and the theoretical model. The difference between both is 0.093 on average, and overall tendencies are respected as the Pearson correlation coefficients are close to 1. Moreover the ranking of parties based on their overall share of content, defined for party s as the average of $p^{(s)}$ over all users, is the same as for empirical estimations

(Figure 6.5). Note that it is also identical to the ranking of parties according to the number of users affiliated to them (Figure 4.1).

However the model is too moderate as it tends to overestimate the small values of $p_n^{(s)}$ and underestimate the high ones. This is due to the random reposting policy: in the model users repost content without distinction for the opinion expressed within. This is not the case in practice, as people are more inclined to interact with congenial content—about 89% of retweets in the dataset are between supporters of the same party.

In Figure 6.4 I compare empirical estimates with averages obtained in simulations using reposting preferences, as in Section 6.2.2. Whenever user n visits their newsfeed at time t, the choice of which item to repost is made at random under the following probability: if $p^{(n)}(t)$ is the M-sized vector describing the state of n's newsfeed at that time, the user will repost an item from party s with probability $\nu_s^{(n)} p_s^{(n)}(t)$. If no item comes from a party with $\nu > 0$ then nothing is reposted. I set $K = 1$ and $M = 5$—I obtain similar results with $M = 2, 10$. This time, I observe a very tight fit between the resulting values of p and empirical estimates. Thus, further research shall strive to solve analytically the case with reposting preferences.

6.4.3.2 Echo chamber effect

I can obtain corresponding values of the ECE and of the AOD via (6.3) and (6.4), respectively. Additionally, I also calculate their values under preferential reposting behaviour, via simulations. I show in Figure 6.6 the resulting cumulative distributions of Φ and Γ, obtained by kernel density estimation[1].

The base model tends to underestimate the ECE and overestimate the AOD. This is because of uniform reposting, meaning users are prone to repost a wider variety of content, even if that content expresses opinions that clash with their personal beliefs.

As we saw in Section 6.4.3.1, the simulations with preferential reposting yield results closer to empirical estimates. The newsfeeds exhibit a strong echo chamber effect and low diversity of opinions, as only 20% of users have $\Gamma < 0.50$, and same

[1] I use the function `kdeplot` from the Python package Seaborn (https://seaborn.pydata.org/generated/seaborn.kdeplot.html)

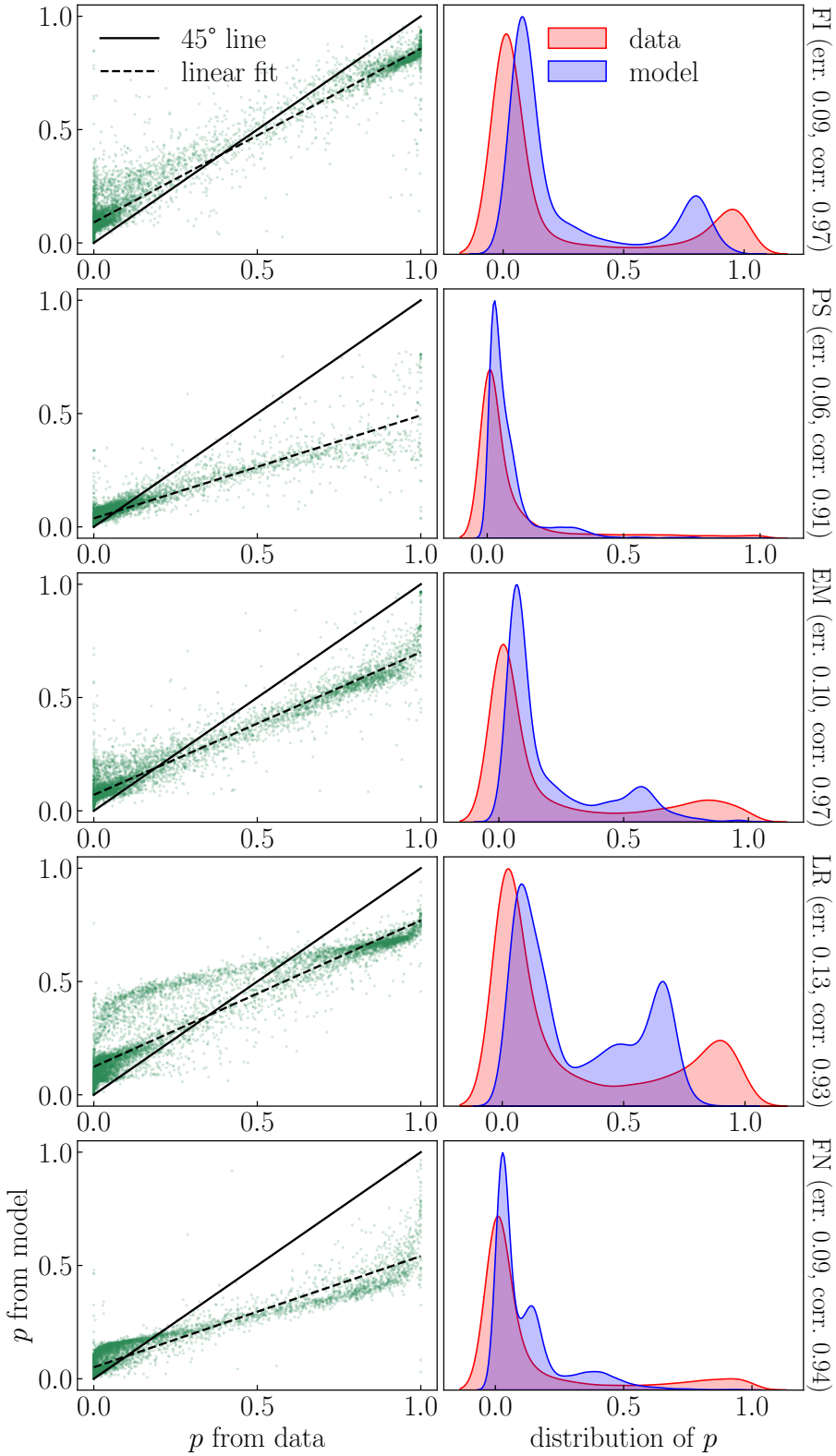

Figure 6.3: Comparison between p estimated from data and p given by the model. Average errors ('err.') and Pearson correlations ('corr.') between estimations and model are indicated above. **Left:** scatter plots. **Right:** distributions.

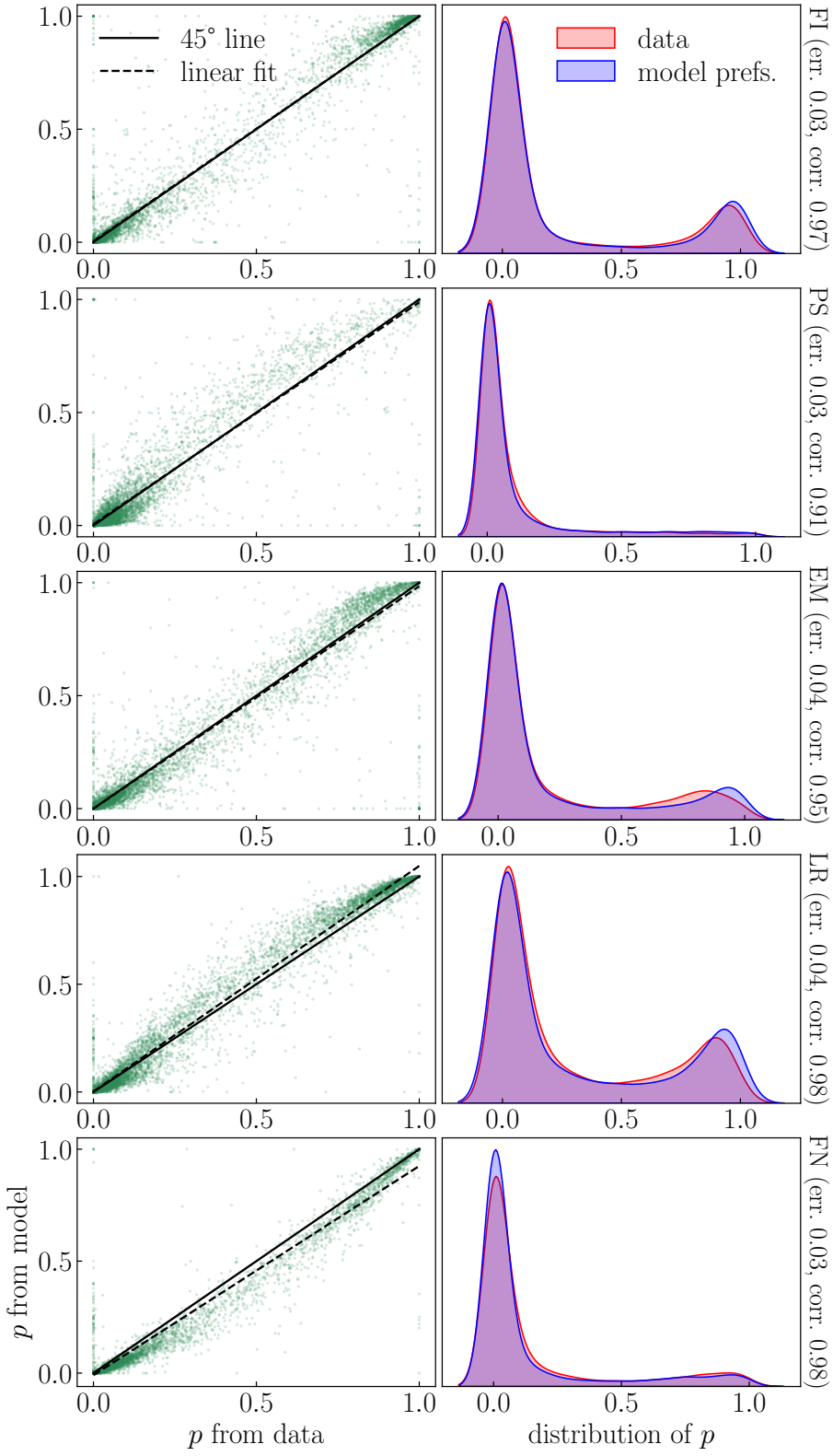

Figure 6.4: Comparison between p estimated from data and p given by the model with reposting preferences (simulations with newsfeed size $M = 5$). Average errors ('err.') and Pearson correlations ('corr.') between estimations and model are indicated above. **Left:** scatter plots. **Right:** distributions.

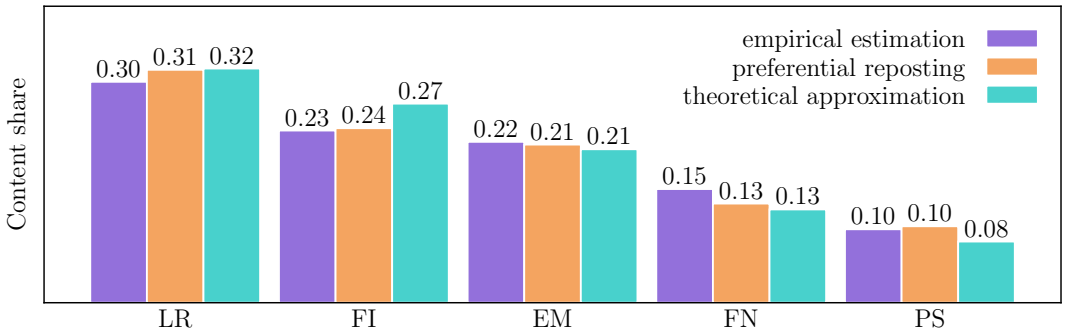

Figure 6.5: Average proportion of content from each party on the newsfeeds, as estimated empirically (purple), as per the the model without preferential reposting (cyan), and as per the model with preferential reposting (orange).

for $\Phi > 0.50$. About 60% of users have $\Gamma > 0.75$ and are thus exposed to more than 75% of congruent opinions.

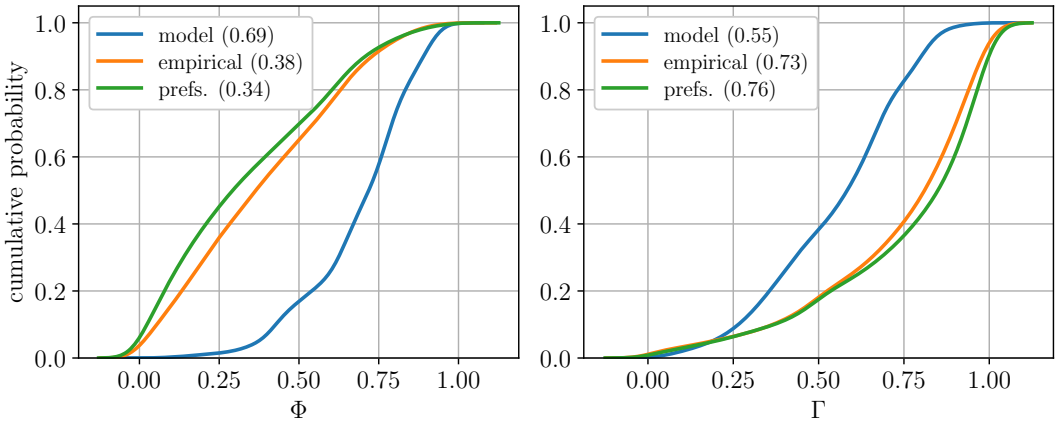

Figure 6.6: Cumulative distributions of Φ **(left)** and Γ **(right)**, as given by the model, the emulator, and simulations with reposting preferences (newsfeed size $M = 10$). Averages are precised in parentheses.

The intensity of the echo chamber effect is not too surprising. Indeed, as we see in Figure 6.7, there is great homophily in the network. For each party, the average proportion of leaders supporting the same views is between 62% (PS) and 88% (FI). This entails the biased distribution of opinions on the newsfeeds, as we see on the right plot that on average, between 53% and 80% of opinions on a newsfeed are congruent.

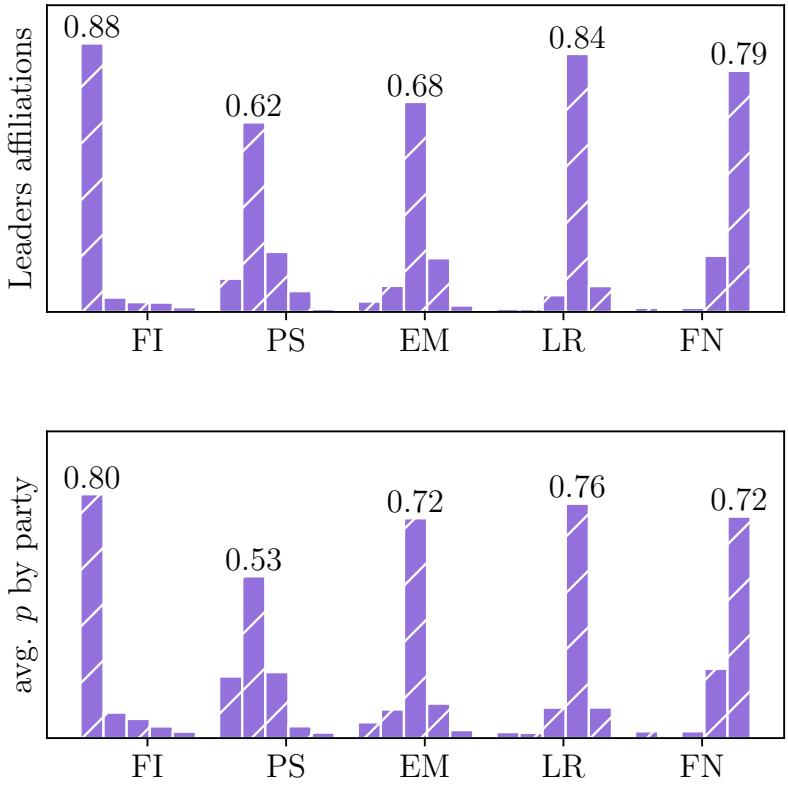

Figure 6.7: **Top:** distribution of leaders' party for each party. **Bottom:** average distribution of empirically estimated p by party. Each bar represents a party, in the same order as the x-axis. All values are empirical and independent from the model.

6.5 Discussion

In this chapter, I presented the Newsfeed Model and the Extended Newsfeed Model, that incorporates opinions. The models describe the flow of content throughout an OSP, as users create new posts or repost items from their newsfeeds.

6.5.1 Theoretical findings

I expressed my metrics of interest, the ECE and the AOD, in this context. They rely on the equilibrium state of the users' newsfeeds. The newsfeeds are interdependent through a linear system of size N, and I gave a closed-form formula to compute their equilibrium distributions. I also explained how to simulate the model and its possible variations.

The model can be further extended by assuming a preferential reposting mechanism, where users choose items to repost according to their concordance with

personal views. However, this induces non-linearities in the model equations, making it difficult to solve analytically. I can still study the model behaviour in this context via simulations. In that case, the size of the newsfeeds has an impact on the equilibrium of the system, which is not the case otherwise.

Finally, I detailed how to infer model parameters, and make empirical estimates of newsfeed distributions, ECE, and AOD, from real-life datasets. I applied the methods on `#Elysée2017fr` and highlighted the existence of an important echo chamber effect, as given by the ECE metric. I also observed that simulations with preferential reposting yielded results closest to empirical estimates. Solving the model with this feature is thus of primary importance for future research concerned with real-life applicability. Future research shall also look into the convergence time of the dynamics.

6.5.2 Empirical evaluation

I evaluated the accuracy of the model on the `#Elysée2017fr` dataset. I obtained Φ-scores, both via theoretical computations and empirical estimations. I also compared theoretical and empirical distributions of opinions on the newsfeeds. The model was able to capture general trends, but fails to return precise results in many cases. I found that it was due to the random reposting behaviour. Indeed, when I assume that users choose items to repost based on personal preferences, the correspondence between empirical and theoretical values is very tight. This highlights the importance to try to incorporate this behaviour in the analytical model, as for now I am only able to simulate it. Note that Giovanidis et al. (2021) contains additional empirical evaluations of the model on different datasets, with encouraging results.

Chapter 7

Steering the echo chamber effect

This chapter lays the final contribution of this project, which is the steering of the echo chamber effect. I start with the macroscopical approach and finish with the microscopical one. In both cases, I demonstrate how to increase the diversity of opinions that users are exposed to, via content recommendation, accounting for possible backfire effects. The results for the macroscopic approach with the EV Model were published—albeit under a slightly different form—in Vendeville et al. (2022c). Most of the results for the microscopic approach with the EV Model were published in Vendeville et al. (2023a).

7.1 Macroscopical approach with the Enhanced Voter Model

I assume a complete network of N individuals, who unilaterally support opinion 0. I model this by assuming all receive the same influence from zealots, as follows:

$$\forall n \in \mathcal{N}, \quad \begin{cases} z_n^{(0)} = z^{(0)} > 0, \\ z_n^{(1)} = z^{(1)} = 0. \end{cases} \tag{7.1}$$

The completeness assumptions acts like a simplifying hypothesis when precise connections are unknown. It is also realistic in certain cases to assume that everyone sees what everyone else posts, as it is the cases for Subreddits or Facebook groups for example.

7.1.1 Without backfire effect

Because $z^{(0)} > 0$ and $z^{(1)} = 0$, the community is homogeneous: each member will end up adopting opinion 0 no matter what. To steer this phenomenon, I suggest increasing $z^{(1)}$, which can be achieved in practice by recommending content supporting opinion 1. My goal is to maximise the AOD, given by Eq. 5.32:

$$\langle \Phi \rangle = \frac{4 z^{(0)} z^{(1)}}{z^2}, \tag{7.2}$$

where $z = z^{(0)} + z^{(1)}$. Here, $\langle \Phi \rangle = 0$ initially. Maximising this quantity is equivalent to solving the following problem:

$$\underset{z^{(1)} \leq B}{\operatorname{argmin}} \left(\frac{z^{(1)}}{z} - \frac{1}{2} \right)^2. \tag{P0}$$

The upper bound B acts as a budget: I do not want to flood users with recommendations, so I limit the amount they will receive. Under no budget constraints, $B = 1 - z^{(0)}$ due to Eq. 5.1. When this bound is reached, users are not exposed to content created by others anymore, but solely to the recommendations, while still being subject to their inner biases towards opinion 0.

By writing the problem as (P0), we see that one can tune the amount diversity to be reached. To do so, it suffices to replace the $1/2$ in the objective by any target value $\tilde{x}^{(1)} \in [0, 1]$. For the sake of generality, I will thus look to solve

$$\underset{z^{(1)} \leq B}{\operatorname{argmin}} \left(\frac{z^{(1)}}{z} - \tilde{x}^{(1)} \right)^2 \tag{P}$$

for a fixed $\tilde{x}^{(1)}$ that I assume chosen. Typically it should be around $1/2$ if the goal is to transform the echo chamber into a diverse sphere of opinions.

Differentiating the objective with respect to $z^{(1)}$ and finding the zeros of the derivative, gives us the solution of (P) as

$$z_\star^{(1)} = \min \left\{ B, \frac{\tilde{x}^{(1)}}{1 - \tilde{x}^{(1)}} z^{(0)} \right\}. \tag{7.3}$$

If the second element is smaller than the first, the target $\tilde{x}^{(1)}$ is reached. Otherwise, one or both of α and $\tilde{x}^{(1)}$ is too high, and we can only hope to approach the target but not reach it exactly.

7.1.2 With backfire Effect

Numerous studies suggest that presenting certain users with opposing views might actually entrench them even deeper in their beliefs. This is known as the backfire effect. To account for it I study the scenario where in reaction to the apparition of recommendations supporting opinion 1, users will reinforce their inner bias towards opinion 0. Formally, I set that for $z^{(1)} > 0$, the value of $z^{(0)}$ is incremented by $\alpha z^{(1)}$, with $\alpha < 1$. The average opinion at equilibrium is now given by

$$\frac{z^{(1)}}{z^{(0)} + z^{(1)} + \alpha z^{(1)}} \tag{7.4}$$

and the target $\tilde{x}^{(1)}$ is exactly reached with

$$z_\star^{(1)} = \tilde{x}^{(1)} z^{(0)} / d \tag{7.5}$$

where $d := 1 - (1+\alpha)\tilde{x}^{(1)}$. If $d > 0$ then $z_\star^{(1)} > 0$ and we can inject this quantity of users into the network. If $d \leq 0$ however this becomes impossible as (7.5) is then either undefined or negative. In this case, I find that the function

$$z^{(1)} \mapsto \left(\frac{z^{(1)}}{z^{(0)} + z^{(1)} + \alpha z^{(1)}} - \tilde{x}^{(1)} \right)^2 \tag{7.6}$$

is strictly positive and decreasing towards $(\tilde{x}^{(1)} - (1+\alpha)^{-1})^2$ over $\mathbb{R}_{>0}$. Thus the larger $z^{(1)}$ the closer we get to the target diversity, up to a certain point. This means that the backfire effect is too strong and the target is unreachable. Thus, we have the optimal values for $z^{(1)}$:

$$\begin{cases} z_\star^{(1)} = \min\left(B, \tilde{x}^{(1)} z^{(0)} d^{-1}\right) & \text{if } d > 0, \\ z_\star^{(1)} = B & \text{if } d \leq 0. \end{cases} \tag{7.7}$$

Because after optimisation we have $z = z^{(0)} + (1+\alpha)z_\star^{(1)}$, the default budget is given by

$$B(\alpha, z^{(0)}) = \frac{1 - z^{(0)}}{1+\alpha}. \tag{7.8}$$

7.1.3 Results on synthetic data

I present the results in Figure 7.1. I consider three different targets $\tilde{x}^{(1)} = (0.1, 0.25, 0.5)$ and intensities of the backfire effect $\alpha = (0, 0.1, 0.25, 0.5, 0.75, 0.9)$. The first value $\alpha = 0$ corresponds to the absence of a backfire effect. The last target $\tilde{x}^{(1)} = 0.5$ corresponds to the maximisation of the AOD *per se*, whereas with the other I am not trying to reach a perfect balance between opinions 0 and 1 but rather to nudge opinions towards 1 just slightly. I first do not assume any budget constraint and simply set B to its default value $B(\alpha, z^{(0)})$ as per (7.8). The ECE depends upon the total number of users and I set $N = 10^3$.

Inflexion points in the curves correspond to the threshold above which the target is not reachable anymore, and we can only hope to approach it by injecting the maximum possible amount of recommendations B. As α and $\tilde{x}^{(1)}$ increase, this threshold gets lower, and so does B. For low enough values of α and $\tilde{x}^{(1)}$, we are able to reach the target without even using the maximal amount of recommendations possible—*cf.* last row of plots.

In the left and middle columns, the AOD never reaches 1 because the target is lower than 0.5. In the last column, it is encouraging to observe that we can reach an AOD higher than 0.5 for most values of α and $z^{(0)}$. In many of these cases however, the optimal proportion of recommendations $z_\star^{(1)}$ is quite high (*cf.* fourth row of plots). This means that users are less subject to influence of peers and more impacted by recommendations and inner bias. With just $\tilde{x}^{(1)} = 0.5$ for example, as soon as the inner bias gets higher than 0.5, users have no influence on each other anymore.

In Figure 7.2 I set B to 0.1 (or $(1 - z^{(0)})/(1+\alpha)$ if it is lower), so that users don't see more than 10% of recommended content. As expected the errors grow way quicker, but we can still sustain an error lower than 0.1 for most values of z_0 and α. The evolution of the metrics is this time divided in three phases: a first one when

optimal diversity is reached, a second one when $z^{(1)} = B$ and AOD and ECE start decreasing, and a third one where $\alpha, z^{(0)}$ are jointly too high, we cannot even take $z^{(1)} = B$ anymore, and AOD and ECE decrease even faster.

Finally, notice that the ECE evolves symmetrically to the AOD in both Figure 7.1 and Figure 7.2, except for very low values of $z^{(0)}$. As I remarked in Section 5.3.3.3, when $z \gg 1/N$ both opinion diversity $\langle \Phi \rangle$ and discord $\langle \rho \rangle$ are equivalent. Because ECE is in this case given by $\langle \Gamma \rangle = 1 - \rho$ as per (5.38), maximising AOD has exactly the same effect as minimising ECE. But this equivalence vanishes when $z \gg N^{-1}$ does not hold anymore. This is confirmed by Figure 7.3: here I have set $N = 10^3$, and we see that the ECE starts mirroring the AOD at about $z^{(0)} = 10^{-2}$, or $10 \times N^{-1}$.

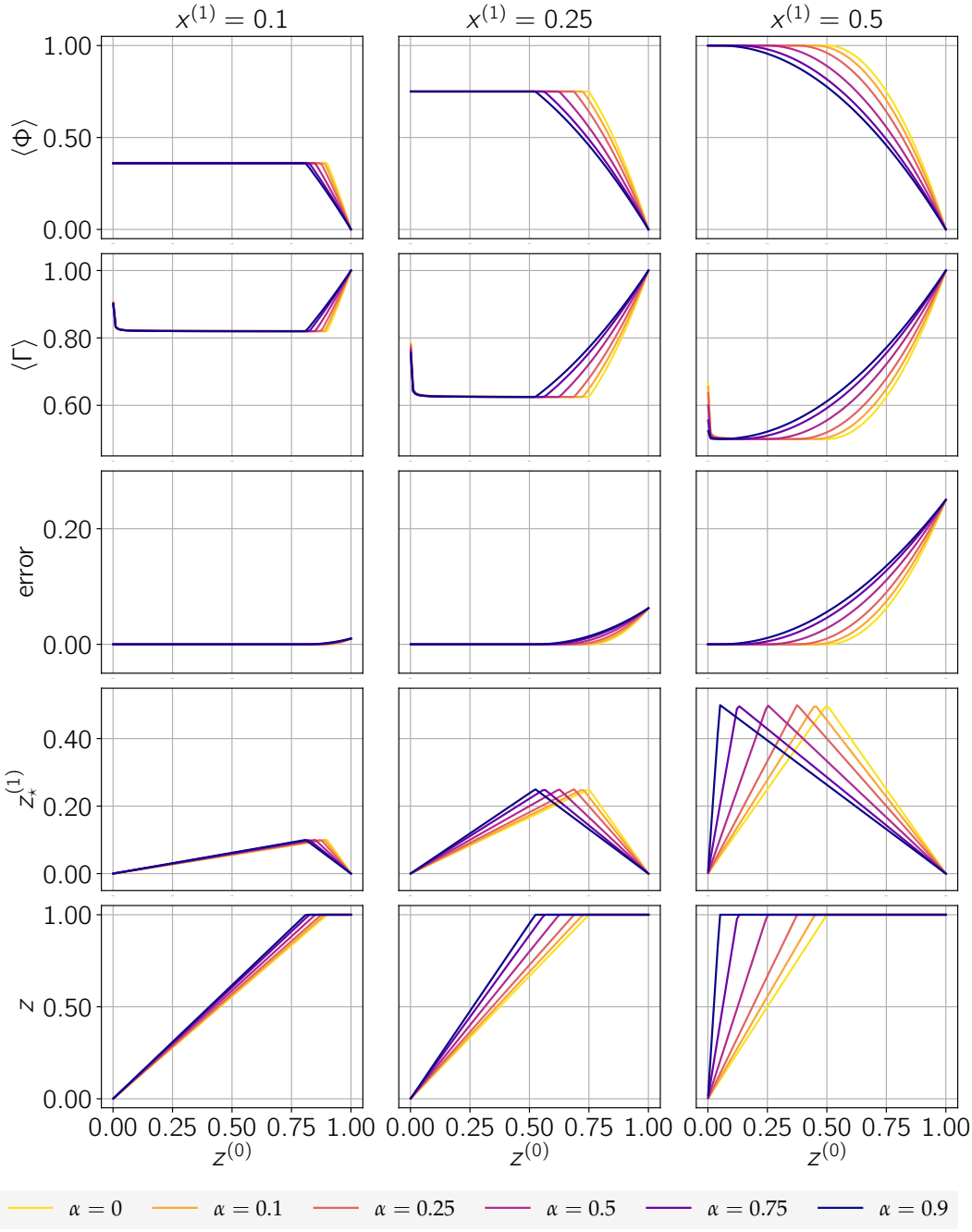

Figure 7.1: Maximisation of AOD on a complete network of size $N = 10^3$, for various intensities α of the backfire effect. In each column I am trying to reach a different target $\tilde{x}^{(1)}$. The budget is maximal (7.8). **First row:** AOD after optimisation. **Second row:** ECE after optimisation. **Third row:** squared error between $z_\star^{(1)}/z$ and $\tilde{x}^{(1)}$ after optimisation. **Fourth row:** optimal $z_\star^{(1)}$ needed to maximise. **Last row:** total zealousness z after optimisation. I impose no constraint on $z^{(1)}$ except the natural one $z^{(1)} \leq 1 - z^{(0)}$.

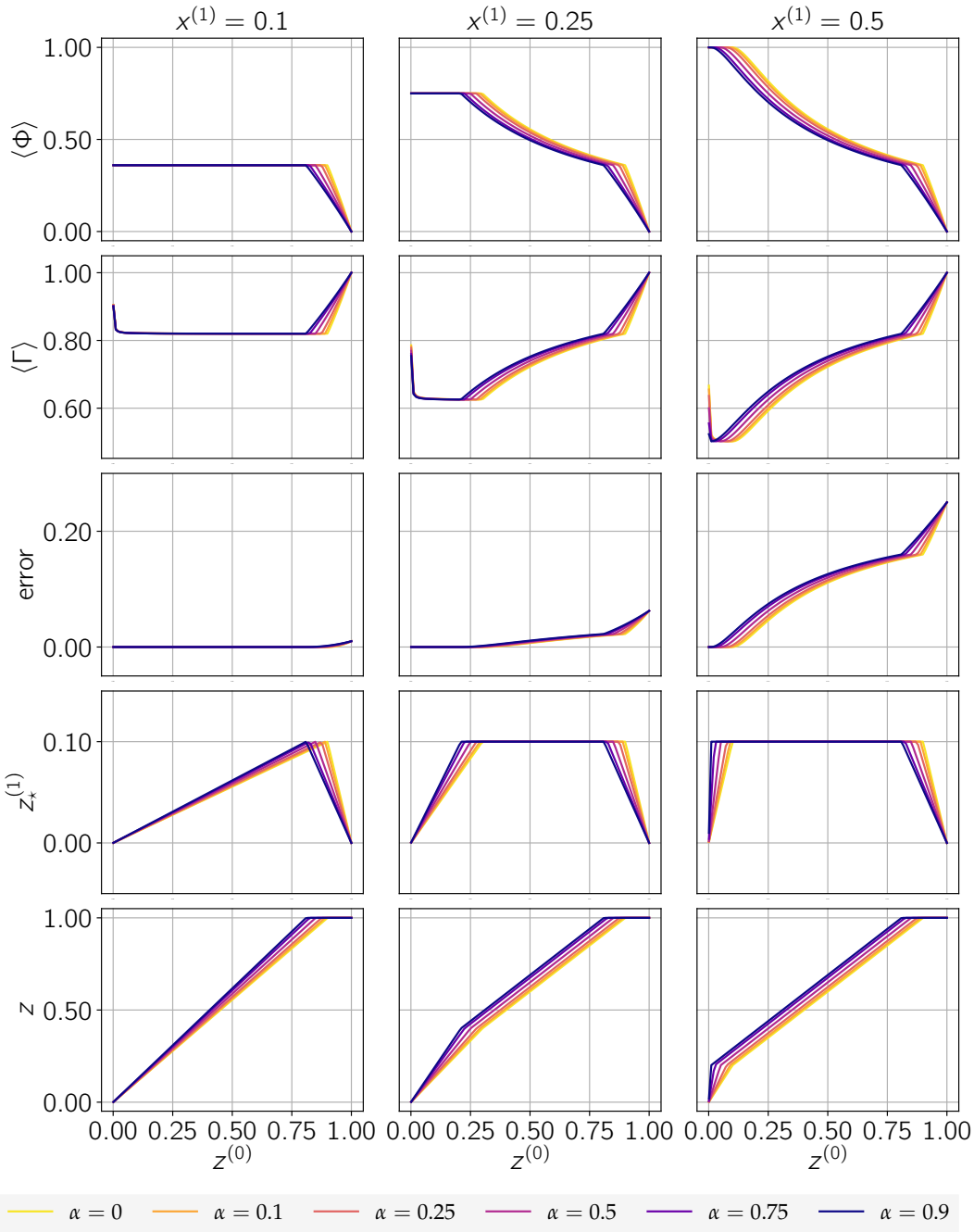

Figure 7.2: Same as Figure 7.1, except this time I set $B = \min\,(0.1, B(\alpha, z^{(0)}))$.

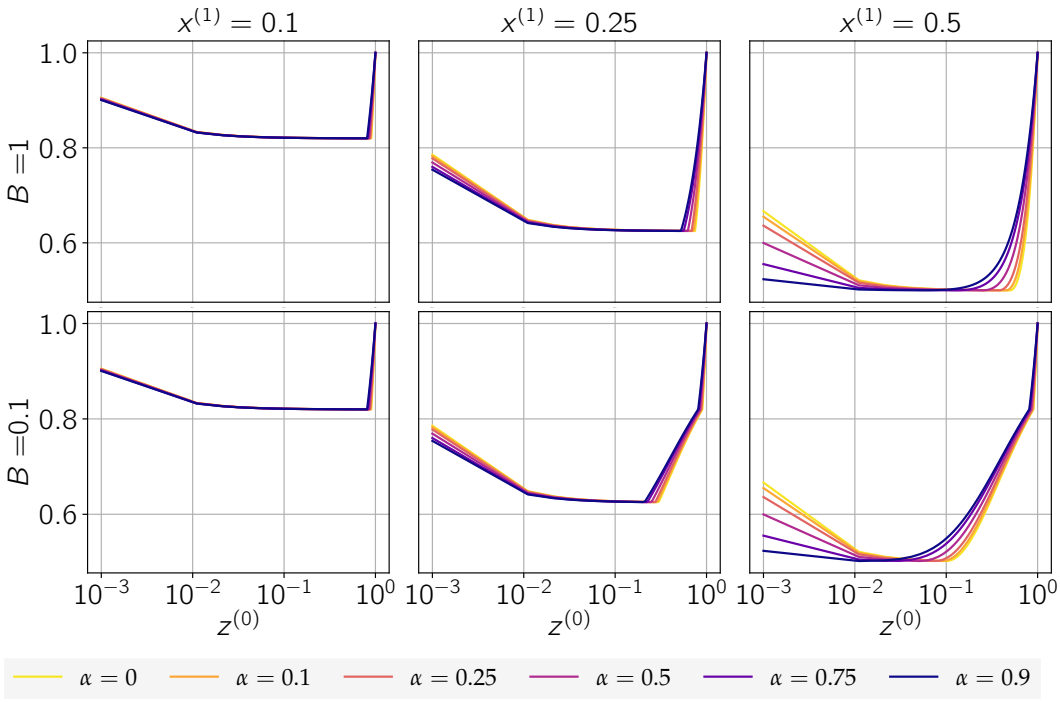

Figure 7.3: Optimal ECE with a log-scaled x-axis, for budgets $B = 1$ (top) and $B = 0.1$ (bottom).

7.2 Microscopical approach with the Extended Newsfeed Model

I now turn to the microscopic approach, that relies on the EN Model. Assume given a (possibly directed) network of users, as well as posting and reposting rates λ_n, μ_n. As per (6.4), the AOD is given by

$$\langle \Phi \rangle = \sum_n \Phi_n(p)/N, \tag{7.9}$$

which is the average diversity of opinions on the newsfeeds. I am looking to solve an optimisation problem of the form

$$\underset{y,p}{\text{argmax}} \langle \Phi \rangle \tag{P0}$$

where y are personalised recommendation policies that describe what opinions should be inserted (in the form of posts) by the platform into the newsfeeds and at what rate. The dependency of $\langle \Phi \rangle$ on y is detailed below. Importantly the impact of inserting an item into a newsfeed is not limited to an immediate change therein, but may also include a broader effect on the whole network as the concerned user can share it to their followers, who may share it further, and so on. The spreading behaviour of the individuals on the platform thus affects the results of the recommendation policies.

7.2.1 Newsfeeds with recommendations

The system can be acted upon by the platform administrator via personalised recommendations, that consist in selecting posts to insert into the newsfeeds of others. Let $y_n^{(s)}$ be the rate at which an s-post (*i.e.* a post expressing opinion s) is inserted into the newsfeed of user n this way. I am looking for the values of y that maximise the AOD. For the sake of equity I would like all newsfeeds at equilibrium to contain on average the same proportion $B < 1$ of recommended posts. This value should not be too high, lest the users will be flooded with recommendations. Users with very active leaders and thus fast-changing newsfeeds should get new recommendations

more often than those with quieter leaders. Formally I require for any user n:

$$\sum_s y_n^{(s)} = \frac{B}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k). \tag{7.10}$$

This ensures that a proportion ω of all content arriving on the newsfeed of n is a recommendation. The recommender system can be seen as if each user n has an artificial leader controlled by the platform, who creates s-posts at rates $y_n^{(s)}$ and these immediately appear on the newsfeed of user n. Hence, the steady-state of the newsfeeds is now given for all n and s by

$$p_n^{(s)} \left(\sum_s y_n^{(s)} + \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) \right) = y_n^{(s)} + \sum_{k \in \mathcal{L}_n} \left(\lambda_k^{(s)} + \mu_k p_k^{(s)} \right). \tag{7.11}$$

This is straightforward from (6.1). Inserting (7.10) in (7.11) we obtain

$$\frac{p_n^{(s)}}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) = y_n^{(s)} + \sum_{k \in \mathcal{L}_n} \left(\lambda_k^{(s)} + \mu_k p_k^{(s)} \right). \tag{7.12}$$

Such values $p_n^{(s)}$ exist and are unique, as proved in Appendix B.2.

7.2.2 Optimisation problem

The optimal recommendation rates that maximise the AOD under budget B can be computed via the following quadratic program with linear constraints.

$$\begin{aligned} \underset{y,p}{\text{argmax}} \quad & \langle \Phi \rangle \\ \text{s.t.} \quad & \frac{p_n^{(s)}}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) = y_n^{(s)} + \sum_{k \in \mathcal{L}_n} (\lambda_k^{(s)} + \mu_k p_k^{(s)}), \quad \forall n, s, \\ & \sum_s y_n^{(s)} = \frac{B}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k), \quad \forall n, \\ & y_n^{(s)}, p_n^{(s)} \geq 0 \quad \forall n, s. \end{aligned} \tag{Q0}$$

Note the presence of p in the optimisation variables, due to its values being dependent on the recommendations y. To avoid any backfire effect, I would like to limit the

recommendation of posts supporting incongruent opinions. To do so, I can tweak the problem (Q0) to penalise the difference between the newsfeeds before and after optimisation. Formally:

$$
\begin{aligned}
\underset{y,p}{\text{argmax}} \quad & \langle \Phi \rangle - \omega \sum_{n,s} \left(\frac{\tilde{p}_n^{(s)} - p_n^{(s)}}{\tilde{p}_n^{(s)} + \varepsilon} \right)^2 \\
\text{s.t.} \quad & \frac{p_n^{(s)}}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) = y_n^{(s)} + \sum_{k \in \mathcal{L}_n} (\lambda_k^{(s)} + \mu_k p_k^{(s)}), \quad \forall n, s, \\
& \sum_s y_n^{(s)} = \frac{B}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k), \quad \forall n, \\
& y_n^{(s)}, p_n^{(s)} \geq 0 \quad \forall n, s.
\end{aligned}
\tag{Q}
$$

In the new objective function, $\tilde{p}_n$ is the newsfeed of n before optimisation, $\omega > 0$ is the strength of penalisation, and $\varepsilon > 0$ ensures that the denominator is positive. The penalty term thus quantifies the total relative change in each newsfeed. With $\omega = 0$ we recover the initial problem (Q0). Note that if one wishes, it is possible to set particular values of ω for each user without impacting the complexity. For example, one may set a stronger penalty for more radical users.

7.2.3 Application on `#Elysée2017fr`

I solve (Q) for the social network described by the `#Elysée2017fr` dataset (follow graph). More details on this dataset are found in Section 4.4.2, and inference methods to estimate models parameters are in Section 6.4.1. With $N = 8,277$ nodes and five categories, problem (Q) has 82,770 variables and 49,662 linear constraints. I solve (Q) for the following budgets and penalty strengths:

$$B = 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20.$$
$$\omega = 0, 1, 10, 100.$$

The case $\omega = 0$ corresponds to the optimisation problem without backfire effect. I also use $B = 0$ to denote (theoretical) values of $\langle \Phi \rangle$ and $\langle \Gamma \rangle$ before optimisation. I set $\varepsilon = 10^{-4}$.

7.2.3.1 Optimal AOD and impact on ECE

In Figure 7.4, we see that, as the budget increases, so does the AOD while the ECE decreases: the newsfeeds get more and more diverse and the prevalence of congruent opinions diminishes. Unsurprisingly, the lower the backfire effect, the better the results. If the penalty strength is high enough, the AOD starts to decrease while the ECE increases when the budget becomes too high. Eventually, for the highest budget and backfire effect, the AOd and ECE are even worse after than before optimisation. This is because, as the backfire effect gets stronger, we have to propose more and more congruent opinions to the users, thus reducing the diversity of their newsfeeds.

The two plots on the right underline how important the diffusion aspect captured by the model is. If I assume that recommended content is never propagated by users, the states of the newsfeeds are given by

$$p_n = (1-B)\tilde{p}_n + By_n/\|y_n\|_1, \tag{7.13}$$

where $\tilde{p}$ are (theoretical) distributions of opinions on the newsfeeds before optimisation. In that case, the evolution of $\langle\Phi\rangle$ and $\langle\Gamma\rangle$ as ω increases follows the same trend as before, but at a smaller magnitude. Thus, users spreading the recommended content is crucial to the effectiveness of the method. Of note, this is also what would happen if users do not appreciate or simply ignore the recommended posts. This is why it is important to account for the backfire effect when solving the optimisation problem.

One may argue that the recommendation algorithm relies on the model equation that do not include preferential reposting, a mechanism with which the model is more truthful to empirical observations. This is why, I use simulations to compute values of $\langle\Phi\rangle$ and $\langle\Gamma\rangle$ obtained with preferential reposting, and with added recommendations as given per the solution of (Q), for $\omega = 0, 10$. I find that recommendations are still effective in increasing AOD and reducing ECE, albeit considerably less so. The trend of the curves is very similar to the no diffusion case. Recall that the vector of preferences for user n is the same as its vector of self-posting λ_n, which contains only zeros except for the one (96% cases) or two (4% cases) entries that correspond to n's

affiliation—*cf.* Figure 4.1. Thus, items supporting incongruent opinions are never reposted, which is what the recommender is mostly inserting into the newsfeeds (*cf.* Figure 7.6, left). Hence, there is little spreading of the recommended content, and the results resembles the case without diffusion. The two curves are shifted when compared to the no diffusion case, due to the fact that reposting preferences induce different values of p, with less AOD and more ECE are users favourise congruent opinions.

7.2.3.2 Analysis of the recommendation rates

Figure 7.5 shows the average rate at which each opinion, *i.e.* party, is recommended. That is for all s, I plot

$$\frac{1}{N}\sum_{n=1}^{N}\frac{y_n^{(s)}}{\|y_n\|_1}. \tag{7.14}$$

Recall that I do not enforce that all users receive recommendations at the same rate, but only that their newsfeeds contain on average the same proportion ω of recommended posts. This is why here, each vector y_n is normalised, so that all users equally impact the mean.

For lower values of $\omega = 0,1$, the rates are rather spread for low budgets and converge towards 0.2 as the budget increases. Indeed when $B \to 1$ the newsfeeds only contain recommended posts so that maximal diversity is achieved with a balanced representation of all parties. Note also that the more a party was initially represented (*cf.* Figure 4.1,6.5), the least they get recommended. As ω gets larger, the tendency reverses. Recommendation are equally distributed between parties for low budgets, but as ω increases, the ranking of parties by their rate of recommendation follows their overall presence on the network (*cf.* Figure 4.1,6.5). This is not surprising, as the backfire effect is so strong that we are more and more forced to recommend congruent opinions. Thus, the overall distribution of recommendations follows the distribution of parties throughout the network.

This is confirmed in Figure 7.6 (left), where I compare vector of recommendation rates y_n and the initial newsfeed distributions $\tilde{p}_n$. The squared difference between them, average over all users, declines towards zero as the budget increases,

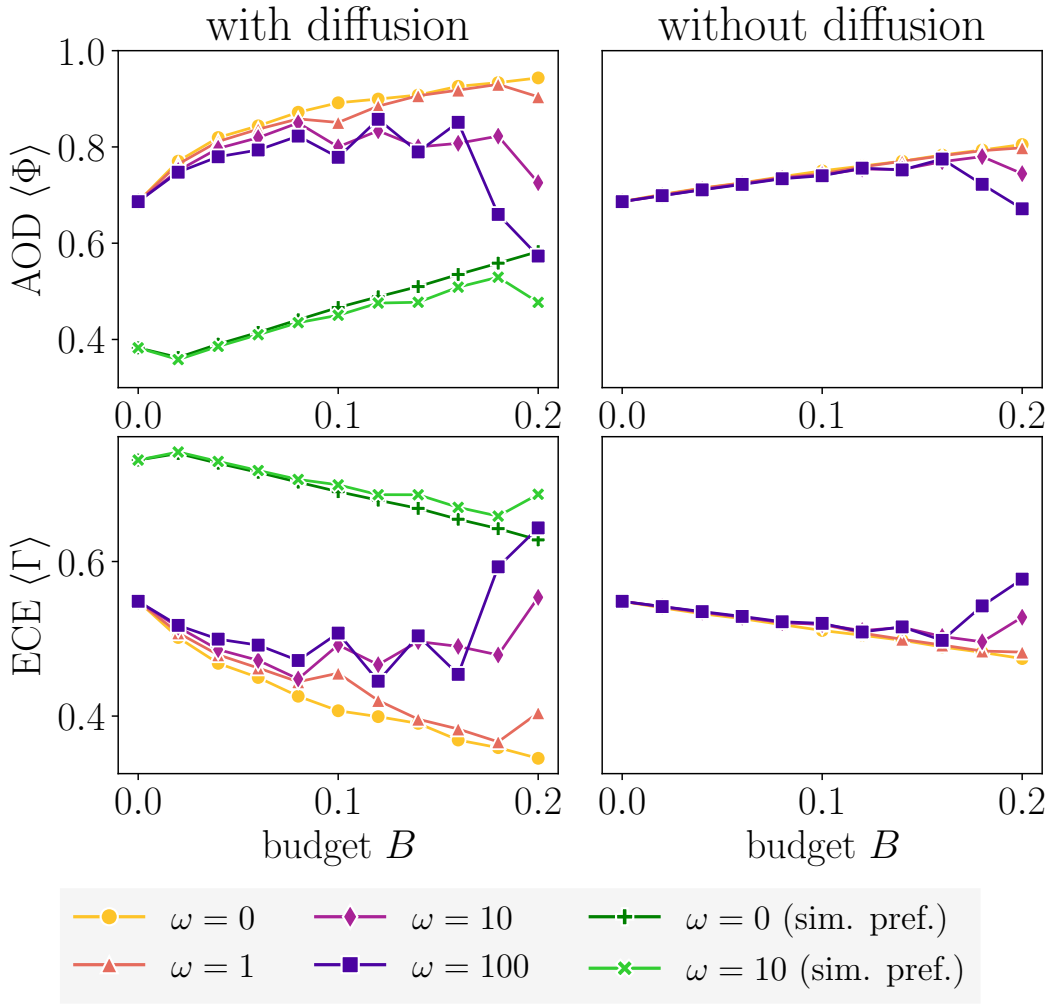

Figure 7.4: Optimisation of $\langle\Phi\rangle$ with increasing budget ω and various penalty strengths ω. **Top:** optimal AOD $\langle\Phi\rangle$. **Bottom:** impact on ECE $\langle\Gamma\rangle$. Budget $B = 0$ corresponds to initial values before optimisation. On the right I show the results without diffusion of the recommended posts. The green lines represent values obtained in simulations, with reposting preferences and a recommender system that follows the optimal recommendation rates (newsfeed size $M = 5$).

at a pace that increases with the backfire effect.

Finally, in Figure 7.6 (right) I show for each value of ω the relative change in $\langle\Phi\rangle$ and $\langle\Gamma\rangle$ per budget unit, for $\omega = 10$. The curves are decreasing, showing that higher budgets have diminishing returns. Adding more and more recommendations into the newsfeeds is less and less effective. We observe similar curves for other values of ω, showing that this behaviour is independent from the intensity of the backfire effect.

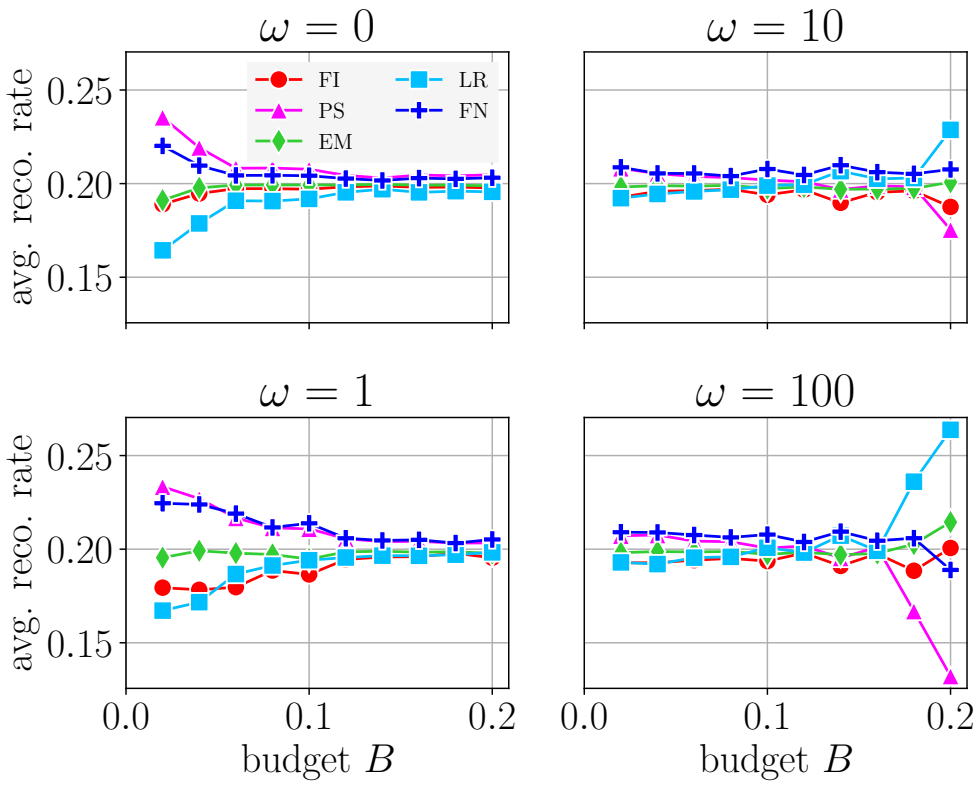

Figure 7.5: Average rate at which each party is recommended. Rates are normalised per user so that $\|y_n\|_1 = 1$ for all $n \in \mathcal{N}$.

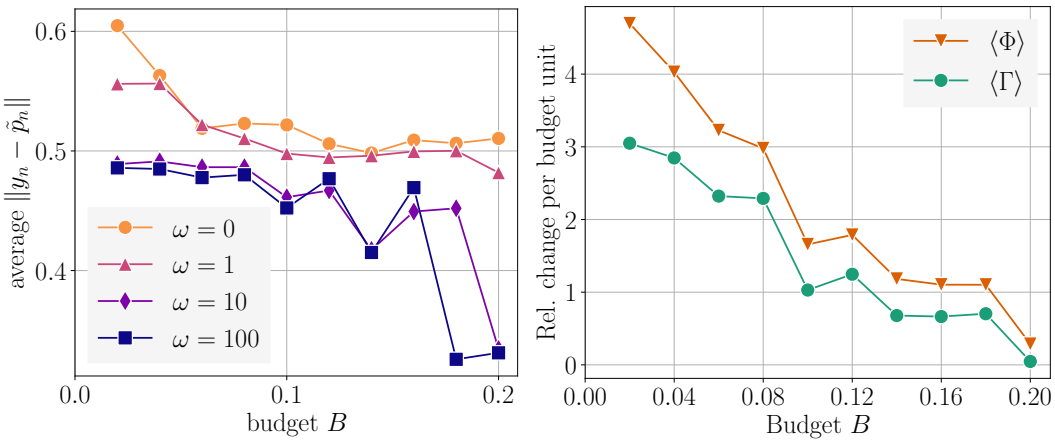

Figure 7.6: Left: average squared difference between recommendation vector and initial newsfeed distribution. **Right:** relative change in the metrics of interest per budget unit, for $\omega = 10$.

7.2.3.3 Implementation and numerical precision

The problem (Q) has a quadratic objective and linear constraints. The time complex-

ity to solve these problems is often considered to be polynomial in the number of variables and constraints (Kozlov et al., 1980). The experiments are run on a virtual machine with 40 vCPUs and 256GB RAM. For the solution of the optimization problem, I configure a Gurobi[1] solver with the barrier algorithm. The runtime is less than fifteen minutes for all parameters values.

While the problem is easily solved, I had trouble with the numerical precision. Indeed, I have thousands of equality constraints. They will never be perfectly accurate, and depending on how tolerant I am on the numerical error, the optimisation might take a very long time—if the software does not return an error before it is finished. I did two things to accelerate the processus:

1. I set the parameter `NumericFocus` to 1, which favours speed against numerical precision.

2. I transformed the equality budget constraint

$$\sum_s y_n^{(s)} = \frac{B}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k) \tag{7.15}$$

into an inequality, and added a penalty term of the form

$$\sum_s y_n^{(s)} - \frac{B}{1-B} \sum_{k \in \mathcal{L}_n} (\lambda_k + \mu_k), \tag{7.16}$$

with a very high penalty coefficient.

Both of the above allow for a higher tolerance to numerical errors, while preserving precision as much as possible. However, the "wobbly" aspect of the high backfire curves in Figure 7.4 and Figure 7.5 makes me believe some errors subsist. To verify that, I compute for each constraint of (Q) the relative difference between both sides of the equation defining the constraint, after optimisation. I show the average and maximum numerical errors for all values of B, ω in Figure 7.7.

We observe that indeed, stronger backfire effects entail higher errors. Average errors are rather low on average, below 0.1%. Maximum errors on the other hand,

[1] https://www.gurobi.com/

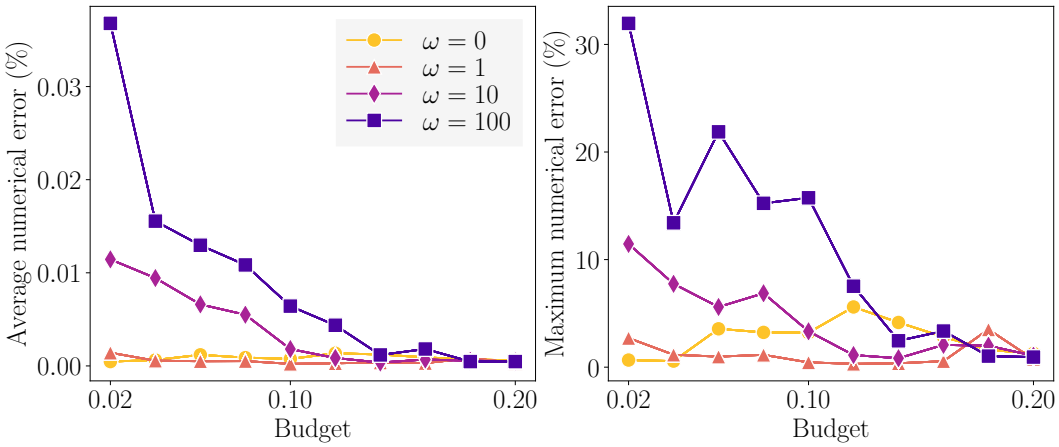

Figure 7.7: Average (left) and maximal (right) numerical error for each budget. Errors are computed as the relative difference between both sides of each constraint in the optimisation problem (Q).

reach as much as $30\%(B = 0.02, \omega = 100)$ and are often in the range of $5-10\%$ for $\omega = 10, 100$. This might explain the aforementioned "wobbly" aspect of some of the plots in this Section. Due to time constraints, I did not perform more experiments to reduce the numerical error. Further research shall try to determine more adapted settings of the optimization software to avoid these issues.

7.3 Discussion

The two proposed methods are effectively able to increase the AOD, and decrease the ECE, for most budgets and backfire effect intensities.

7.3.1 Macroscopical approach

For the macroscopical approach, the method does not depend on the size of the system and is thus completely scalable. We saw that maximising the AOD and minimising the ECE were equivalent, except when we do not have $z \gg 1/N$. In that case, one has to be mindful that this equivalence vanishes. For future research, it may be interesting to give time constraints in the form of a maximum duration tolerated for the group to reach the target diversity. Because convergence time of the model decreases with z, this comes down to imposing a lower bound on $z^{(1)}$.

The method I just presented is fairly simple. I assume a complete network where all agents share the exact same characteristics. They are all biased towards

opinion 0 with the same strength and react to the recommendations in the same way. These assumptions are the strength and the weakness of this method at the same time: I lose in precision but I gain in universality, simplicity, and tractability. As mention earlier, this also makes the method highly adaptable to contexts where we have very few information of the system at hand, merely its general preference for an opinion rather than the other.

The microscopical method is precisely developed to take fine-grained, individual features into account. However, if one wants more precise methods that use the Voter Model, there are several possible extension. Non complete network, individual levels of bias, and individual recommendation rates first spring to mind. As more partisan people are often more prone to backfire, I could make the intensity of the backfire effect proportional to the initial bias of the agent. Some agents could also simply be immune to backfire, or others immune to recommendations. I could take also age into account, so that older agents are less likely to change opinion. Another idea to model the backfire effect is to rely on discord: if a user experiences too much discord with their neighbours, they radicalise into a zealot, with the mode of their opinion distribution as their new, definitive opinion. These are all interesting leads, that future research shall look into.

7.3.2 Microscopical approach

For the microscopical approach, the optimisation method relies on the model equations that do not include preferential reposting, which we saw was crucial to fit empirical data. However, the optimal recommendation rates returned are able to have the desired effect under preferential reposting—albeit less so—as proved through simulations. Additionally, it is possible to set particular values of ω for each user without impacting the complexity. For example, one may set a stronger penalty for more radical users.

We also saw that high budgets and high backfire effects induce an undesirable outcome: the recommender injects too much congruent opinions into the newsfeed, thus reinforcing the ECE instead of attenuating it. In this context, perhaps it would be wiser to simply recommend less, for example by replacing the equality budget

constraint by an inequality. Doing so would induce an inequity is the proportion of recommendations shown to different users, wish might not be something to wish for.

The recommender did favourise some parties more than others, because of their unequal numbers of supporters in the dataset. This might also entail ethical issues, and an interesting lead to avoid it is to enforce equality of all parties in terms of share of recommendations. This could be done via an additional penalty in the objective of the optimisation problem (Q).

Finally, implementing the optimisation problem is not without difficulties. There exists a trade-off between speed of execution and numerical precision. In certain cases, we saw the software (Gurobi) make significant numerical errors, that were reflected in some of the constraints not being exactly respected. It is difficult to strike a good balance between accuracy and efficiency here, and future research shall dive into this issue.

7.3.3 In practice

There are a few things one need if they wish to use my methods in practice. The intensities α, ω of the backfire effects can be chosen in various ways depending on the context. For example, we can treat more radical communities with more caution by choosing higher values. The budget B should be chosen to reflect how often we want each user to be recommended content.

For the Microscopical method, we would need precise estimation of the model parameters λ, μ as well as the user graph. The platform administrator has access to ground-truth values, but otherwise, they can be estimated as described in Section 6.4.1. The error made in the process can themselves also be estimated—again see Section 6.4.1.

In the Macroscopical method, the initial bias $z^{(0)}$ can be treated in several ways. It may be estimated, although it is not obvious how to estimate the bias of a community. It could inferred based on the number of times they mention specific politicians, or post partisan newspaper articles. Another way of proceeding is to act similarly as for the backfire effect: we do not know the bias, but the more partisan we believe the community is, the higher bias we attribute to it, which forces us to be

more careful.

So, which one to choose? This will depend on the problem at hand, the information available as well as time constraints and computation capacities. The strength of the Macroscopical method lies in its simplicity and tractability. When dealing with highly homogeneous communities with all-to-all connections, such as some sub-Reddits or Facebook pages, the assumptions of the method can be realistic. In that case, Macroscopical will be very fast and efficient. It is also useful when not much user-level information is available, or is too cumbersome to collect. On the other hand, in more heterogeneous networks with known user-level features, Microscopical will perform better and is thus preferable. It can also accomodate any number of opinions, while Macroscopical only works with two. The complexity of the optimisation problem, as well as the potential for numerical errors, might however make the process much longer and require a precise configuration of the solver software. Under constraints on time and computation capacity, Macroscopical can then be preferred. Finally, it would be interesting to compare the outcome of both approaches on a same dataset, real or synthetic. We could get a better grasp on how they fare one versus the other depending on the structure of the network, and the precise context of the datasets.

Chapter 8

Conclusion

The regulation of online social platforms is one of the major challenges faced by our societies today. They have profoundly affected the informational landscape and the political debate, posing serious threats to democracy and integrity of information. There is in particular growing concern about the echo chamber effect, where a lack of opinion diversity in the content presented to users is leading to a fragmentation of the society in polarised clusters. It has become crucial to develop principled tools to assess *(i)* the extent of this phenomenon, *(ii)* the impact of the personalisation algorithms employed by the platforms, and *(iii)* possible avenues towards healthier personalisation algorithms.

At the same time, the advent of social platforms has blessed us with rife, fine-grained, readily accessible data on human behaviour. This is a huge leap forward when compared to previous existing methods that relied on surveys, making them cumbersome, imprecise and difficult to scale. Thus has come a unique opportunity to better understand social phenomenons. The modelling of opinions dynamics and social structures, which has interested scholars for many years, is largely benefitting from this upheaval to propose more realistic and better calibrated mathematical models.

There is a lack of overarching works that improve on existing theoretical models for opinion dynamics to propose empirically verified frameworks to describe the echo chamber effect, and investigate how it can be acted upon via the platforms algorithms. This was the aim of this doctoral research. First, I designed novel mathematical

models to describe and quantify the echo chamber effect in social networks. I verified their accuracy on empirical datasets. Then, I developed recommendation methods in order to increase the diversity of opinions that users in a social network are exposed to. I proposed a macroscopical and a microscopical method, each adapted to a different level of granularity. The latter was specifically tailored for online social platforms, while the former was more general and applicable in any type of social network.

These methods are thought of as proofs of concepts rather than guidelines on how should OSPs be regulated. The backbone of this project was to provide tools that can useful in the broad study of opinion evolution, diffusion, and echo chambers phenomenons online. These tools are destined to other researchers, who wish to study these phenomenons. I demonstrated the effectiveness of these tools for the specific problem of maximising opinion diversity, but I do not advocate that it is what must be done, and that this is the best way to do it. This are merely destined to serve as a basis for decision-making by policymakers or platform administrators. Moreover, I believe that as more problems and challenges arise concerning social media, my methods can be adapted to gain more insight on how to address them.

8.1 Achievements

My contributions mostly benefit two domains: the study of opinion dynamics through mathematical models, and the regulation of echo chambers in OSPs. The first domain I contributed to via the novel models I proposed, the Enhanced Voter Model (EV Model) and the Extended Newsfeed Model (EN Model), that I also evaluated on empirical data. Bridging with the second domain, I introduced metrics to quantify the echo chamber effect and the diversity of opinions that users in a social network are exposed to. I expressed these metrics in the context of my models, and developed recommendation methods to steer their values.

8.1.1 Contributions to the field of opinion dynamics

The EV Model generalises the traditional Voter Model to directed, weighted agent graphs with exogenous influences (zealots), multiple opinions and individual update

rates. Due to the origins of the Voter Model in the field of statistical physics, efforts have mostly focused on regular graphs and mean-field approximations. Thus, few works have focused on the application of the model to real-life networks. I derived a closed-form formula for the probability of discord between any two agents, allowing for the computation of the order parameter known as active links density, that I extended to general cases. I was able to demonstrate that, adding links between polarised communities may sometimes increase or decrease the echo chamber effect and the diversity of opinions. I believe this is an important stone in the study of the voter model in complex networks, and opens the road for refined investigation on real-life datasets.

I also extended the Newsfeed Model, that describes the flow of content throughout the newsfeeds of users in an OSP, by incorporating the notion of opinions. This let me quantify how political views spread throughout the network and what distributions of opinions are users exposed to. This model takes into account the different activity rates across users, and complex topologies, making it highly applicable to real-life settings. I introduced preferential reposting, whence users choose content to repost based on personal preferences.

The two models were evaluated on real-life datasets. The EN Model exhibited very good correspondence between theoretical predictions and empirical estimates, when applied on data from Twitter. It was able to identify the distributions of opinions on the newsfeeds of users, with a very high accuracy if a mechanism of preferential reposting is used. As for the EV Model, I studied its ability to predict the outcome of democratic elections, based on passed ones. On one hand, my method allowed me to derive proportions of stubborn voters who never change opinion, making it a valuable tool for the study of political landscapes.

8.1.2 Contributions towards the regulation of echo chambers

I introduced novel metrics to measure the echo chamber effect, and the diversity of opinions, that users are exposed to. Both can be calculated in the context of the EV Model and the EN Model, as functions of the equilibrium states of the system considered. The two measures are complementary: closely related, and sometimes

equivalent, but not always. In particular, I highlighted in Section 5.4.2 how one being high does not mean the other is low, and vice-versa. Both are thus important to consider when studying the impact of recommendation algorithms in OSPs.

My final contribution was to develop recommendation methods to steer the echo chamber effect. For each model, I searched for optimal recommendation rates to maximise the diversity of opinions that users are exposed to. I improved on the existing literature by proposing both a macroscopical and a microscopical approach, and taking into account possible backfire effects, that make users susceptible to reinforce their pre-existing beliefs when presented with incongruent opinions.

I took a macroscopical approach with the EV Model, considering a complete network of like-minded agents. This benefits general settings where the overall leaning of the population is known, but fine-grained data is not available. It is also particularly applicable in some real cases, such as Facebook groups or Subreddits, where the agent network is effectively complete as everyone has access to the content created by everyone else. When backfire effects and inner biases are not too high, it is possible to effectively increase the overall opinion diversity. This has the advantage of requiring very little computation.

I took a microscopical approach with the EN Model, allowing the craft of individual recommendation policies in a setting with complex topologies, and various political views in the same network. This is done by the means of a quadratic optimisation problem with linear constraints. I demonstrated the effectiveness of the method in the context of the `#Elysée2017fr` dataset. The optimisation problem does not include preferential reposting for the reasons stated above. However, I showed via simulations, that the optimal recommendation rates also had the intended impact on the system when simulating with preferential reposting. The importance of taking the diffusion aspect of the model into account was also highlighted.

For people willing to use this work in practice, I also explained in details how to proceed: from inferring the model parameters—and estimating the associated error—to choosing optimisation parameters and implementing the optimisation programs. Most importantly, I laid the advantages and drawbacks of each method, helping

to choose between both depending on the context. In general, the Macroscopical method should be used either in *(i)* homogeneous contexts, *(ii)* in the absence of information on user-level features, or *(iii)* under certain constraints on time or computing capacity. In all other cases, the Microscopical method should return more refined results, allowing for a more precise steering of the echo chamber effect. Note also that anyone looking to apply the EV Model on large-scale networks should in priority seek to reduce the complexity of the computation of discord probabilities. This one of the most important limitations of this work, as I discuss now.

8.2 Limitations

In the EV Model, computing discord probability however relies on solving a large linear system of size $\mathcal{O}(N^2)$, which can quickly become intractable for large systems. Accelerating the computation of finding accurate approximations is thus necessary if one wishes to compute discord probabilities for large systems. As for the empirical evaluation, I obtained an average error of 4.7% when forecasting the outcome of an election. This is a non-negligible gap in politics, where elections are often decided by a few percentage points. The model parameters, such as the number of agents N and the time unit, could be adapted to enhance accuracy. Dynamical estimates could also be made without accounting for the presence of stubborn agents. Note that, in absence of more precise information, I considered a complete network where everyone is connected to everyone else. This is probably an unrealistic assumption. Further research shall thus look to investigate the validity of the EV Model on complex topologies.

For the steering of the echo chamber effect with the EC Model, I also considered complete graphs. Developing similar methods for complex topologies might entail intractability, and require heuristics or greedy approximations. Introducing varying degrees of bias at the individual level, as well as considering individual recommendation rates, are also interesting possible extensions. As more partisan individuals tend to be more susceptible to the backfire effects, we could try to adjust the intensity of the penalty depending on the initial bias of the agent. Another strategy for modeling

the backfire effect could involve discord probabilities: if a user experiences too much discord with their peers, it may lead to radicalisation, lmeading them to stick with their opinion and not update anymore.

As for the EN Model, we saw that preferential reposting was a crucial feature if one wishes to obtain empirically accurate values. I was however only able to incorporate it in simulations, as it makes the model equations non linear and harder to analyse. Thus, the main limitation of the model as it stands is the analytical intractability of this feature. This does not impede steering of the ECE too much however. We saw that optimisation results obtained without accounting for preferential reposting, yielded good results when injected into a simulator that incorporates a recommender system and a preferential reposting mechanism.

The problem I studied for steering the ECE does not enforce equity between the different political parties, and some were much more recommended than others. The may raise ethical concerns, and could be avoided via an additional penalty in the objective of the optimisation problem. I also obtained undesirable outcomes with high budgets and high backfire effects, that could be avoided by replacing the equality budget constraint by an inequality. This would allow the recommender to have varying rates between users, effectively making some less exposed than others to recommendations. An important and fairly straightforward extension is also to use individual penalties for the backfire effect, in order to account for varying levels of partisanship, and susceptibility to incongruent opinions, in the population. Finally, we saw that specific attention should be devoted to the trade-off between speed of execution and numerical precision in the implementation of the optimisation problem.

8.3 Future research

I already laid some leads for future research. The EV Model lacks a fast algorithm to compute discord probabilities, an empirical evaluation of its predictions on fine-grained datasets, and a recommendation method for complex topologies with different-minded agents. The EN Model lacks analytical tractability when

considering preferential reposting, which has turned out to be a crucial feature for real-life applicability. The optimisation problem to find optimal recommendation rates could also be further refined, to enforce equity between political parties, for example.

There is an important aspect of social relations that I did not take into account: enmity. Indeed, I did not consider antagonistic connections between agents. They are often represented by negative edges in social graphs, and can be crucial to accurately describe the phenomenons of polarisation or echo chambers (Keuchenius et al., 2021). For example, Williams et al. (2015) found that there was a lot of communication between climate activists and climate sceptics on Twitter, but it was mostly hostile and aggressive—a feature that my models, as presented here, do not incorporate. I believe this to be a primary lead for further investigations.

Another area for improvement is the design of recommendation policies. In this project, I only studied recommendation rates per opinion, but not how exactly they are implemented. The exact type of content recommended, and the recommendation of users to one another, is important to consider. The ranking of items in the newsfeeds, and filtering policies that hide content, are examples of other important personalisation features. To be able to further refine recommendation policies, it is crucial to have access to the actual data employed by the platforms, and how there personalisation algorithms are built. The EU has started to take action in this direction with the Digital Service Act (EU, 2022), and Twitter has already made their algorithm public.

The models I developed and the results I presented in this thesis are not limited to the sole purpose of steering the echo chamber effect. The theoretical aspect of the models, especially the EV Model, makes them applicable in many other contexts. One can think of using my models to predict the evolution of opinions in a population, to replace traditional polls or surveys. Doing so, they could also help better understand the impact of political elites or media outlets on the public opinion, by incorporating such entities in the models and trying to isolate their impact on the dynamics. The EN Model, with its good description of content propagation

in OSPs, can be used for marketing purposes. For example, to find when and to whom a product should be advertised in order to yield the most visibility. Beyond social platforms and politics, there is an interesting field of applicability for the EV Model: collective intelligence. In particular for *robot swarms*, coordinated groups of robots that can solve problems or take decisions by interacting with each other via pre-determined rules (Zavala-Rio et al., 2013). The Voter Model dynamics may be an interesting type of interaction rule for this purpose.

Overall, I believe that research shall keep striving to improve descriptions of social phenomenons, and especially how they intertwine with the structure and functionality of online social platforms. First to provide us with a deeper understanding of ourselves, and our societies in general. But also because, in the current times, the need for appropriate regulation of OSPs is stronger than ever. It is not obvious how exactly this should be done, and the precise effects that it will have. This is why, it is very important that researchers develop refined, empirically accurate, characterisations of the phenomenons taking place. Once we have a profound understanding of the impact of OSPs, we can propose adequate policies to keep their nefarious effects at bay.

Appendix A

Notations

Table A.1: General notations.

Symbol	Definition	Range	Equation
N	number of users	$\mathbb{N}$	-
$\mathcal{N}$	user set	-	$\{1,\dots,N\}$
S	number of possible opinions	$\mathbb{N}$	-
$\mathcal{S}$	opinion set	-	$\{1,\dots,S\}$
i,j,k,n	a user	$\mathcal{N}$	-
r,s	an opinion	$\mathcal{S}$	-
$\mathcal{G}$	agent/user graph/network	-	-
$\mathcal{E}$	edge set of $\mathcal{G}$	-	-
W	adjacency matrix of the user graph $\mathcal{G}$	$[0,1]^{N\times N}$	-
w_{ij}	weight of the directed edge $j \to i$	$[0,1]$	
$\mathcal{L}_n$	set of leaders of user n	$\mathcal{P}(\mathcal{N})$	$\{j \in \mathcal{N} : w_{ij} > 0\}$
$\langle\cdot\rangle$	overall average at equilibrium	-	-
$A^\top$	transpose of the matrix A	-	$A^\top_{ij} = A_{ji}$
$u^\top$	transpose of the vector u	-	-
e^A	matrix exponential of A	-	$\sum_{k\in\mathbb{N}} A^k/k!$
$\mathbf{1}\{\bullet\}$	indicator function	$\{0,1\}$	0 if $\bullet$ is false, 1 if true
$\mathbf{P}$	probability	$[0,1]$	-
$\lvert\cdot\rvert$	absolute value (scalar), cardinal (set)	-	-
Γ_n	ECE for user n	$[0,1]$	-
$\langle\Gamma\rangle$	average ECE at equilibrium	$[0,1]$	$\sum_{n\in\mathcal{N}} \Gamma_n/N$
Φ_n	AOD for user n	$[0,1]$	-
$\langle\Phi\rangle$	average AOD at equilibrium	$[0,1]$	$\sum_{n\in\mathcal{N}} \Phi_n/N$

Table A.2: General notations for the EVM, Chapters 5 and 7.

Symbol	Definition	Range	Equation
σ_i	opinion of user i	$\mathcal{S}$	-
$z_i^{(s)}$	zealousness of user i towards opinion s	$[0,1]$	-
$x_i^{(s)}$	probability of user i having opinion s	$[0,1]$	$\sum_j w_{ij} x_j^{(s)} + z_i^{(s)}$
$y_i^{(s)}$	average exposure of user i to opinion s	$[0,1]$	$\frac{\sum_j w_{ij} x_j^{(s)}}{\sum_j w_{ij}}$
r_i	update rate of user i	$\mathbb{R}_{>0}$	-
ρ_{ij}	probability of discord between agents i and j	$[0,1]$	(5.17),(5.12)
$\langle\rho\rangle$	generalised active links density (GALD)	$[0,1]$	(5.23)
w_{ij}^{∞}	coordinate (i,j) of the matrix exponential of W	$\mathbb{R}_{>0}$	-
$cos(u,v)$	cosine similarity between u and v	$[-1,+1]$	$u^\top v/\|u\|\|v\|$

Table A.3: Notations for the EVM, sections 5.3.3.1,5.3.3.2,5.3.3.3, and Chapter 7.

Symbol	Definition	Range	Equation
$z^{(s)}$	zealousness of users towards opinion s	$\mathbb{R}_{>0}$	-
z	total zealousness of users	$\mathbb{R}_{>0}^2$	$\sum_{s\in\mathcal{S}} z^{(s)}$
$x^{(s)}$	probability of users having opinion s	$\mathbb{R}_{\geq 0}$	$z^{(s)}/z$
B	optimisation budget	$[0,1]$	-
α	backfire effect intensity	$[0,1]$	-
$B(\alpha, z^{(0)})$	maximal budget	$[0,1]$	$(1-z^{(0)})/(1+\alpha)$
$z_\star^{(1)}$	optimal recommendation rate of opinion 1	$\mathbb{R}_{>0}$	(7.3),(7.7)
$\tilde{x}^{(1)}$	target support for opinion 1	$[0,1]$	-

Table A.4: Notations for the EVM, Section 5.3.3.4 and Section 5.5.

Symbol	Definition	Range	Equation
$N_1(t)$	number of nodes with opinion 1 at time t	$\mathbb{N}$	-
n_1	initial number of nodes with opinion 1	$\mathbb{N}$	$N_1(0)$
$q_{k,l}$	transition rate from state k to state l	$\mathbb{R}$	-
Q	transition rates matrix	-	-
$p_{n_1,k}(t)$	probability for $N_1(t)$ to equal k	$\mathbb{R}$	$[e^{tQ}]_{n_1,k}$
M	total number of elections	$\mathbb{N}$	-
x_i	percentage of votes for the party in the i^{th} election	$\mathbb{N}$	-
$(z_\star^{(0)}, z_\star^{(1)})$	most likely percentages of stubborn agents	$\mathbb{N}^2$	(5.45)

Table A.5: Notations for the Newsfeed model, Chapters 6,6.4,7.

Symbol	Definition	Range	Formula
M, K	newsfeeds and walls size	$\mathbb{N}$	-
λ_n	posting rate of user n	$\mathbb{R}_{>0}$	-
$\lambda_n^{(s)}$	posting rate of user n for opinion s	$[0,\lambda_n]$	-
μ_n	reposting rate of user n	$\mathbb{R}_{>0}$	-
$\nu_n^{(s)}$	preferences of n towards opinion s when reposting	$[0,1]$	$\lambda_n^{(s)}/\lambda_n$
$p_n^{(s)}$	proportion of posts supporting opinion s on the newsfeed of n	$[0,1]$	-
B	optimisation budget	$[0,1]$	-
ω	backfire effect	$\mathbb{R}_{>0}$	-
$y_n^{(s)}$	optimal recommendation rate of s-posts to user n	$\mathbb{R}_{>0}$	-
$\tilde{p}_n^{(s)}$	initial values of p before optimisation	$[0,1]$	-

Appendix B

Mathematical proofs

B.1 Unicity of ρ

I demonstrate that the spectral radius of V is strictly less than 1, which implies that Eq. 16 has a unique solution. I make use of the following technical lemma.

Lemma 1 (Azimzadeh (2018), Lemma 2.1)**.** *Let A be the adjacency matrix of a graph $\mathcal{F}$ so that a_{ij} is the weight of the edge $j \to i$. The spectral radius of A is strictly less than 1 if and only if for every row i, one of the following holds:*

- *row i sums to strictly less than 1, or*
- *there is a path $k \to \ldots \to i$ in $\mathcal{F}$ and row k sums to strictly less than 1.*

The matrix V can be seen as the adjacency matrix of a new graph $\mathcal{F}$. Its nodes correspond to agent pairs, and there is an edge from node $i'j'$ to node ij if and only if one of i' or j' is a leader of i or j in the original agent graph $\mathcal{G}$. Let v_{ij} denote the sum of the row of V that corresponds to node ij. We have

$$v_{ij} = \frac{1}{2}\left(\sum_{k\in\mathcal{L}_i} w_{ik} + \sum_{k\in\mathcal{L}_j} w_{jk}\right). \tag{B.1}$$

Lemma 1 tells us that it suffices to prove, for every ij: either $v_{ij} < 1$, or there exists another node $i'j'$ with $v_{i'j'} < 1$ and a path from $i'j'$ to ij in $\mathcal{F}$. If $v_{ij} = 1$, assuming every agent can be influenced by a zealot, there exists an agent k such that $z_k^s > 0$ for some s and a path from k to i. Hence there is path from ik to ij in $\mathcal{F}$, and $v_{ik} < 1$ as shown by Eq. 2 in the main text.

B.2 Unicity of p with recommendations

I prove that Eq. 7.12 has a unique solution. Let us write it in matrix form: $p_s = Ap_s + b$. As long as the spectral radius $\rho(A)$ of A is stricly less than 1, the system has a unique solution $p_s = (I - A)^{-1}b$. The entries of A are given by

$$a_{ij} = (1 - B)\frac{\mu^{(j)}}{\sum_{k \in \mathcal{L}^{(i)}} \lambda^{(k)} + \mu^{(k)}} \mathbf{1}_{j \in \mathcal{L}^{(i)}} \tag{B.2}$$

and because $B > 0$ it holds that $\sum_j a_{ij} < 1$ for any row i. But from Horn and Johnson (1990, Thm. 8.1.22) we have $\rho(A) \leq \max_i \sum_j a_{ij}$ and thus p_s exists and is unique.

Appendix C

Discord and communities

An immediate application of discord in the EVM (Chapter 5) is the analysis of polarised networks. If two groups support different ideas, how fiercely do they disagree? To what extent does it depend on the connections between them, and on the influence of zealots in each camp? To study these questions I calculate the generalised active links density of networks partitioned in two communities $\mathcal{C}_0$ and $\mathcal{C}_1$, for various values of the model parameters, at equilibrium. The rest of this section is dedicated to the discussion of the results, illustrated in Figure C.1.

When reinforcing connections between different-minded communities, the novel paths of influence creat more diverse information flows, and thus a higher exposition to contradicting beliefs. There are two main consequences one could expect from this. Agents may revise their viewpoint to incorporate adverse ideas, leading the system towards a more consensual state of lower discord. Alternatively, they may fiercely cling onto their prexisting opinions, thus creating more tension and reinforcing discord—the so-called *backfire effect* (Bail et al., 2018; Schaewitz and Krämer, 2020).

When community $\mathcal{C}_0$ is much less zealous than $\mathcal{C}_1$, for $z = (0.1, 0.5)$ and $z = (0.1, 0.9)$, outgroup links can surprisingly reduce the discord within $\mathcal{C}_0$. While adding too few of them will introduce more discord in the community, once they reach a critical mass we observe a decrease in $\langle \rho \rangle$. This stems from the fact that $\mathcal{C}_1$ has fiercer partisans, meaning opinion 1 gets more and more prevalent in the network as connectivity between the two communities goes up—see the support for opinion

0 dropping in the bottom plots. Thus even within $\mathcal{C}_0$, holding opinion 1 guarantees more agreement with peers.

It is striking that contrary to $\langle\rho\rangle$, the average difference in opinion $\langle\Delta x\rangle$ (middle plots) always decreases within communities, given enough outgroup links. The opinion difference between i and j is defined by $\Delta x_{ij} = \|x_i - x_j\|_2$. Hence despite an surge of discord amongst like-minded agents, the distributions of their equilibrium opinions converge. This might be due to the fact that as edges are added, the network gets closer to a complete one. Individual node particularities thus fade as agents become more similar to one another. But as opinions distributions approach the uniform distribution, the probabilities of drawing two different values increase. This highlights the importance of distinguishing between similarity of opinion distribution and discord probabilities.

Between the communities it seems that discord always diminishes with more outgroup links. However, as shown in the inset (upper right plot), for equally low values of zealousness $z = (0.1, 0.1)$, first $\langle\rho\rangle$ decreases but it quickly goes back up as p_{out} gets larger. Thus, when agents are not very zealous, relations with opposite minded-others become more tumultuous as connections between them increase. For high values of z, discord does decrease with p_{out}, but stays at rather high values.

Finally, the lower left plot showcases the importance of using other measures than simply average opinions. This was already highlighted in Vazquez and Eguíluz (2008). When the zealousness is the same on both sides, average opinions over the whole network (bottom left plot) do not change with more links, but we observe a rich behaviour in the evolution of $\langle\rho\rangle$ and Δx (top left, center left plots).

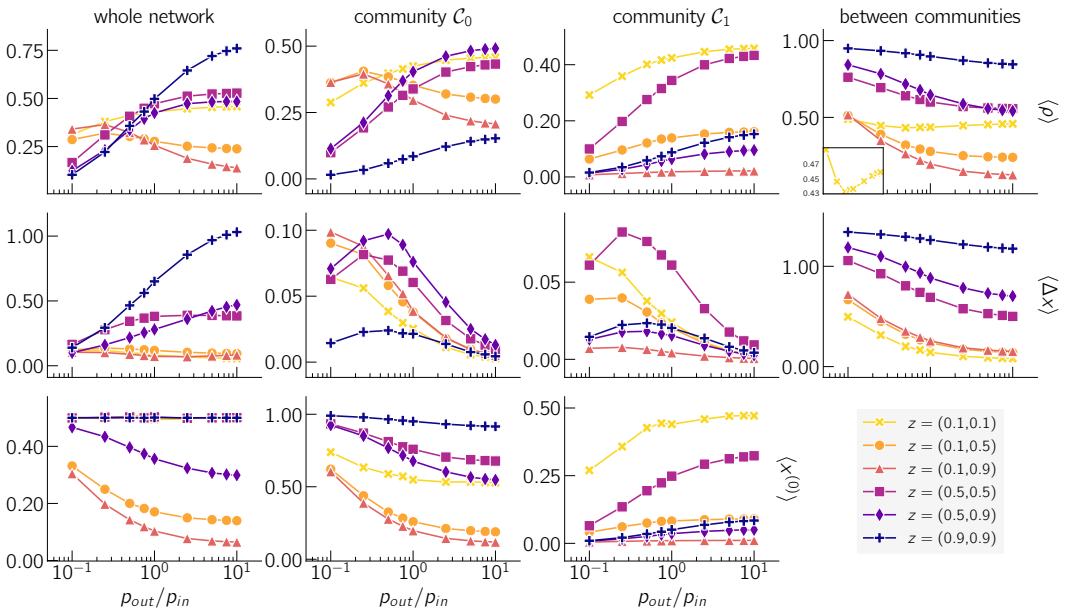

Figure C.1: Network of $N = 100$ agents with two communities $\mathcal{C}_0$ and $\mathcal{C}_1$. The s-zealot exert a total influence $z^{(s)}$ on each agent in $\mathcal{C}_s$ and none on others. In-group link probability is fixed at $p_{in} = 0.1$, while the out-group link probability p_{out} and zealousness $z = (z^{(0)}, z^{(1)})$ vary. Results are averaged over 20 agent graphs generated under the Stochastic Block Model, for the whole network (left), within each community (center) and between them (right). **Top:** Generalised active links density. **Middle:** Opinion difference. **Bottom:** Support for opinion 0. The plots have different scales for clarity, but the purpose is to focus on the qualitative dynamics rather than exact values.

Appendix D

Elections table

Table D.1: Evolution of the estimates for the proportion of stubborn agents $(z_\star^{(0)}, z_\star^{(1)})$ over time. Left: United Kingdom. Right: United States.

Year	Conservative	Labour
1924	(62, 38)	(65, 30)
1929	(20, 21)	(61, 30)
1931	(28, 20)	(55, 30)
1935	(1, 5)	(53, 27)
1945	(9, 10)	(48, 26)
1950	(11, 10)	(26, 17)
1951	(13, 12)	(23, 16)
1955	(13, 12)	(23, 16)
1959	(15, 14)	(22, 16)
1964	(16, 15)	(25, 18)
1966	(18, 16)	(25, 18)
1970	(18, 16)	(24, 18)
1974	(19, 17)	(26, 19)
1974	(19, 16)	(26, 19)
1979	(19, 16)	(26, 19)
1983	(20, 17)	(28, 20)
1987	(20, 17)	(22, 15)
1992	(22, 18)	(21, 14)
1997	(22, 18)	(23, 15)
2001	(19, 15)	(24, 16)
2005	(18, 14)	(24, 16)
2010	(17, 13)	(24, 16)
2015	(18, 13)	(22, 14)
2017	(18, 13)	(22, 14)
2019	(19, 14)	(22, 14)
2024	(19, 14)	(24, 15)

Year	Republicans	Democrats
1920	(23, 23)	(44, 42)
1924	(15, 21)	(18, 12)
1928	(18, 23)	(15, 8)
1932	(18, 23)	(18, 11)
1936	(16, 18)	(10, 8)
1940	(13, 13)	(7, 7)
1944	(14, 14)	(9, 8)
1948	(15, 15)	(9, 8)
1952	(17, 16)	(10, 9)
1956	(16, 16)	(11, 10)
1960	(16, 16)	(11, 10)
1964	(17, 17)	(12, 11)
1968	(17, 16)	(10, 10)
1972	(17, 16)	(12, 11)
1976	(15, 15)	(11, 10)
1980	(16, 16)	(12, 11)
1984	(16, 16)	(13, 11)
1988	(16, 16)	(13, 11)
1992	(16, 16)	(14, 12)
1996	(15, 15)	(14, 12)
2000	(16, 15)	(15, 13)
2004	(16, 15)	(15, 13)
2008	(16, 16)	(15, 13)
2012	(17, 16)	(16, 14)
2016	(17, 16)	(16, 14)
2020	(18, 17)	(16, 14)

My publications during the PhD

Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2022a). Recommendation of content to mitigate the echo chamber effect. In *Conference on Complex Systems*, Palma de Mallorca, Spain. Extended abstract.

Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2023a). Opening up echo chambers via optimal content recommendations. In *Complex Networks and Their Applications XI*, pages 74–85.

Vendeville, A., Guedj, B., and Zhou, S. (2021). Forecasting elections results via the voter model with stubborn nodes. *Applied Network Science*, 6(1):1.

Vendeville, A., Guedj, B., and Zhou, S. (2022b). Active links density in the Voter Model with zealots. In *Conference on Complex Systems 2022*, Palma de Mallorca, Spain. Extended abstract.

Vendeville, A., Guedj, B., and Zhou, S. (2022c). Towards control of opinion diversity by introducing zealots into a polarised social group. In *Complex Networks & Their Applications X*, pages 341–352.

Vendeville, A., Guedj, B., and Zhou, S. (2023b). Discord in the multi-state voter model with zealots. *In review*.

Bibliography

Abebe, R., Chan, T.-H. H., Kleinberg, J., Liang, Z., Parkes, D., Sozio, M., and Tsourakakis, C. E. (2021). Opinion dynamics optimization by varying susceptibility to persuasion via non-convex local search. *ACM Trans. Knowl. Discov. Data*, 16(2):1–34.

Acemoglu, D., Dahleh, M. A., Lobel, I., and Ozdaglar, A. (2011). Bayesian learning in social network. *The Review of Economic Studies*, 78(4):1201–1236.

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 36–43, New York, NY, USA. Association for Computing Machinery.

Aldayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing and Management*, 58(4):102597.

Altafini, C. (2012). Dynamics of opinion forming in structurally balanced social networks. *PLOS ONE*, 7(6):e38135. Publisher: Public Library of Science.

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

Andris, C., Lee, D., Hamilton, M. J., Martino, M., Gunning, C. E., and Selden, J. A. (2015). The rise of partisanship and super-cooperators in the U.S. house of representatives. *PLOS ONE*, 10(4):1–14.

Arugute, N., Calvo, E., and Ventura, T. (2023). Network activated frames: Content sharing and perceived polarization in social media. *Journal of Communication*, 73(1):14–24.

Audickas, L., Cracknell, R., and Loft, P. (2020). UK Election Statistics: 1918-2019 – A Century of Elections.

Avena, L., Baldasso, R., Hazra, R. S., den Hollander, F., and Quattropani, M. (2022).

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *J. Conflict. Resolut.*, 41(2):203–226.

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

Azimzadeh, P. (2018). A fast and stable test to check if a weakly diagonally dominant matrix is a nonsingular m-matrix. *Math. Comp.*, 88(316):783–800.

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

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

Banerjee, A. and Fudenberg, D. (2004). Word-of-mouth learning. *Games and Economic Behavior*, 46(1):1 – 22.

Banerjee, S., Jenamani, M., and Pratihar, D. K. (2020). A survey on influence maximization in a social network. *Knowl Inf Syst*, 62(9):3417–3455.

Banisch, S. and Olbrich, E. (2019). Opinion polarization by learning from social feedback. *J Math Sociol*, 43(2):76–103.

Barabasi, A.-L. (2016). *Network Science*. Cambridge University Press.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit Anal*, 23(1):76–91. Publisher: Cambridge University Press.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: is online political communication more than an echo chamber? *Psychol Sci*, 26(10):1531–1542. Publisher: SAGE Publications Inc.

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., and Starnini, M. (2021). Emergence of polarized ideological opinions in multidimensional topic spaces. *Phys. Rev. X*, 11:011012.

Beckett, L. and Wong, J. C. (2020). The misinformation media machine amplifying Trump's election lies. *The Guardian*.

Bessi, A., Petroni, F., Vicario, M. D., Zollo, F., Anagnostopoulos, A., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2016). Homophily and polarization in the age of misinformation. *Eur Phys J Spec Top*, 225(10):2047–2059.

Betsch, C. and Sachse, K. (2013). Debunking vaccination myths: Strong risk negations can increase perceived vaccination risks. *Health Psychology*, 32(2):146–155.

Bickert, M. (2020). Removing Holocaust denial content. *Facebook Newsroom*.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Bizyaeva, A., Franci, A., and Leonard, N. E. (2023). Nonlinear opinion dynamics with tunable sensitivity. *IEEE Transactions on Automatic Control*, 68(3):1415–1430.

Blake, A. (2021). Facebook removes popular anti-mask group for violating policy against coronavirus misinformation. *The Washington Times*.

Bogost, I. (2021). People aren't meant to talk this much. *The Atlantic*.

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *J Comput Mediat Commun*, 23(1):17–33. Publisher: Oxford Academic.

Bronner, G. (2021). *Apocalypse cognitive*. Presse Universitaire de France.

Caridi, I., Manterola, S., Semeshenko, V., and Balenzuela, P. (2019). Topological study of the convergence in the voter model. *Appl Netw Sci*, 4(1):1–13.

Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81(2):591–646. Publisher: American Physical Society.

Castellano, C., Marsili, M., and Vespignani, A. (2000). Nonequilibrium phase transition in a model for social influence. *Phys. Rev. Lett.*, 85(16):3536–3539. Publisher: American Physical Society.

Cen, S. H. and Shah, D. (2020). Regulating algorithmic filtering on social media. arXiv: 2006.09647.

Cencetti, G., Contreras, D. A., Mancastroppa, M., and Barrat, A. (2023). Distinguishing simple and complex contagion processes on networks. *Phys. Rev. Lett.*, 130:247401.

Centola, D. (2018). *How Behavior Spreads: The Science of Complex Contagions*. Princeton Analytical Sociology Series. Princeton University Press.

Chan, M.-p. S. and Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nat Hum Behav*.

Chavalarias, D. (2022). *Toxic data: Comment les réseaux manipulent nos opinions*. Flammarion.

Chen, W., Pacheco, D., Yang, K.-C., and Menczer, F. (2021). Neutral bots probe political bias on social media. *Nat Commun*, 12(1):5580.

Chen, W., Wang, C., and Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 1029–1038, New York, NY, USA. Association for Computing Machinery.

Chinellato, D. D., Epstein, I. R., Braha, D., Bar-Yam, Y., and De Aguiar, M. A. M. (2015). Dynamical response of networks under external perturbations: Exact results. *J Stat. Phys.*, 159(2):221–230.

Chitra, U. and Musco, C. (2020). Analyzing the impact of filter bubbles on social network polarization. *Proc. of WSDM2020*, page 115–123.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proc. Natl. Acad. Sci.*, 118(9):e2023301118.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2020). Echo chambers on social media: A comparative analysis. arXiv: 2004.09603.

Cinus, F., Minici, M., Monti, C., and Bonchi, F. (2021). The effect of people recommenders on echo chambers and polarization. *arXiv:2112.00626 [physics]*.

Clifford, P. and Sudbury, A. (1973). A model for spatial conflict. *Biometrika*, 60(3):581–588.

Collins, B. and Zadrozny, B. (2020a). Facebook bans QAnon across its platforms. *NBC News*.

Collins, B. and Zadrozny, B. (2020b). Twitter bans 7,000 QAnon accounts, limits 150,000 others as part of broad crackdown. *NBC News*.

Collins, B. and Zadrozny, B. (2020c). YouTube bans QAnon, other conspiracy content that targets individuals. *NBC News*.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political Polarization on Twitter. In *ICWSM*.

Cook, J. and Lewandowsky, S. (2012). *The debunking handbook*. Skeptical Science. OCLC: 768864362.

Cota, W., Ferreira, S. C., Pastor-Satorras, R., and Starnini, M. (2019). Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Sci*, 8(1):1–13.

Culliford, E. and Paul, K. (2020). With fact-checks, Twitter takes on a new kind of task. *Reuter*.

Cuthbertson, A. (2020). Facebook membership of anti-mask groups shoots up nearly 2000% since August. *The Independent*.

Daly, E. M., Geyer, W., and Millen, D. R. (2010). The network effects of recommending social connections. In *Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10*, page 301, Barcelona, Spain. ACM Press.

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

De Francisci Morales, G., Monti, C., and Starnini, M. (2021). No echo in the chambers of political interactions on Reddit. *Sci Rep*, 11(1):2818.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98.

DeGroot, M. H. (1974). Reaching a consensus. *J. Am. Stat. Assoc.*, 69(345):118–121.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2017a). Modeling confirmation bias and polarization. *Sci Rep*, 7(1):40391.

Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., and Quattrociocchi, W. (2017b). Mapping social dynamics on Facebook: The Brexit debate. *Soc Netw*, 50:6–16.

Diaz-Diaz, F., San Miguel, M., and Meloni, S. (2022). Echo chambers and information transmission biases in homophilic and heterophilic networks. *Sci Rep*, 12(1):9350.

Dimock, M., Kiley, J., Keeter, S., and Doherty, C. (2014). Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life. *Pew Research Center*.

Dubois, E. and Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Inf Commun Soc*, 21(5):729–745.

Duggins, P. (2017). A psychologically-motivated model of opinion change with applications to American politics. *JASSS*, 20(1):13.

Dunlap, R. E., McCright, A. M., and Yarosh, J. H. (2016). The political divide on climate change: Partisan polarization widens in the U.S. *Environment: Science and Policy for Sustainable Development*, 58(5):4–23.

Erdös, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290.

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

EU (2022). Digital Services Act: regulating platforms for a safer online space for users. Press release from the European Parliament, ref. 20220114IPR21017.

Even-Dar, E. and Shapira, A. (2007). A note on maximizing the spread of influence in social networks. In Deng, X. and Graham, F. C., editors, *Internet and Network Economics*, pages 281–286, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fernandez-Gracia, J., Suchecki, K., Ramasco, J. J., San Miguel, M., and Eguíluz, V. M. (2014). Is the voter model a model for voters? *Phys. Rev. Lett.*, 112:158701.

Ferraz de Arruda, H., Maciel Cardoso, F., Ferraz de Arruda, G., R. Hernández, A., da Fontoura Costa, L., and Moreno, Y. (2022). Modelling how social network algorithms can influence opinion polarization. *Information Sciences*, 588:265–278.

Fraisier, O., Cabanac, G., Pitarch, Y., Besançon, R., and Boughanem, M. (2018). #Élysée2017fr: The 2017 French Presidential Campaign on Twitter. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

French, J. R. (1956). A formal theory of social power. *Psychol. Rev.*, 63:181–94.

Friedkin, N. E. and Johnsen, E. C. (1990). Social influence and opinions. *J. Math. Sociol.*, 15(3-4):193–206.

Funk, C. and Tyson, A. (2020). Partisan differences over the pandemic response are growing. *Pew Research Center*.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2016). Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*, pages 33–42. ACM Press.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017a). The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 43–52, New York, NY, USA. Association for Computing Machinery.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017b). Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 81–90, New York, NY, USA. Association for Computing Machinery.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*,

WWW '18, pages 913–922. International World Wide Web Conferences Steering Committee.

Garimella, K., Gionis, A., Parotsidis, N., and Tatti, N. (2017c). Balancing information exposure in social networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4666–4674, Red Hook, NY, USA. Curran Associates Inc.

Garimella, V. R. K. and Weber, I. (2017). A long-term analysis of polarization on Twitter. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 528–531. AAAI PRESS.

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the united states, 1974 to 2010. *Am Sociol Rev*, 77(2):167–187.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from twitter data. *Soc. Sci. Comput. Rev.*, 31(6):649–679.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2016). Measuring group differences in high-dimensional choices: Method and application to congressional speech.

Giovanidis, A., Baynat, B., Magnien, C., and Vendeville, A. (2021). Ranking online social users by their influence. *IEEE/ACM Transactions on Networking*, 29(5):2198–2214.

Giovanidis, A., Baynat, B., and Vendeville, A. (2019). Performance analysis of online social platforms. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 2413–2421.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goyal, M., Chatterjee, D., Karamchandani, N., and Manjunath, D. (2019). Maintaining ferment. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5217–5222.

Grabisch, M., Mandel, A., and Rusinowska, A. (2023). On the design of public debate in social networks. *Operations Research*, 71(2):626–648.

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

Grevet, C., Terveen, L. G., and Gilbert, E. (2014). Managing political differences in social media. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 1400–1408, Baltimore, Maryland, USA. ACM Press.

Gruzd, A. and Roy, J. (2014). Investigating political polarization on Twitter: A Canadian perspective. *Policy Internet*, 6(1):28–45.

Grömping, M. (2014). 'echo chambers': Partisan Facebook groups during the 2014 Thai election. *Asia Pac Media Educ*, 24(1):39–59. Publisher: SAGE Publications India.

Guerra, P. C., Jr, W. M., Cardie, C., and Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries.

Haidt, J. (2022). Why the past 10 years of american life have been uniquely stupid. *The Atlantic*. Accessed on June 6, 2022.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *J Public Econ*, 143:73–88.

Harary, F. (1953). On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143 – 146.

Hazla, J., Jin, Y., Mossel, E., and Ramnarayan, G. (2020). A geometric model of opinion polarization. arXiv: 1910.05274.

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

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112. PMID: 21010780.

Hills, T. T. (2019). The dark side of information proliferation. *Perspect Psychol Sci*, 14(3):323–330.

Himelboim, I., McCreery, S., and Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *J Comput Mediat Commun*, 18(2):40–60.

Holley, R. A. and Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann Probab*, 3(4):643–663.

Holme, P. and Newman, M. E. J. (2006). Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E*, 74:056108.

Hong, S. (2013). Who benefits from twitter? Social media and political competition in the U.S. House of Representatives. *Gov Inf Q*, 30(4):464 – 472.

Hong, S. and Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Gov Inf Q*, 33(4):777–782.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press.

Hosseinmardi, H., Ghasemian, A., Clauset, A., Rothschild, D. M., Mobius, M., and Watts, D. J. (2020). Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube. arXiv: 2011.12843.

Hu, J., Meng, K., Chen, X., Lin, C., and Huang, J. (2014). Analysis of influence maximization in large-scale social networks. *SIGMETRICS Perform. Eval. Rev.*, 41(4):78–81.

Huberman, B. A., Romero, D. M., and Wu, F. (2009). social networks that matter twitter under the microscope. *First Monday*.

Jadbabaie, A. (2012). Non-Bayesian social learning. *Games and Economic Behavior*, page 16.

Kaiser, B. (2019). *Targeted*. Harper Collins.

Kan, U., Feng, M., and Porter, M. A. (2023). An adaptive bounded-confidence model of opinion dynamics on networks. *Journal of Complex Networks*, 11(1):cnac055.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1).

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 137–146, New York, NY, USA. Association for Computing Machinery.

Keuchenius, A., Törnberg, P., and Uitermark, J. (2021). Why it is important to take into account negative ties when studying Twitter: polarization results not from separation but confrontation. In *Networks2021*.

Kirdemir, B. and Agarwal, N. (2022). Exploring Bias and Information Bubbles in YouTube's Video Recommendation Networks. In *Complex Networks & Their Applications X*, pages 166–177.

Klamser, P. P., Wiedermann, M., Donges, J. F., and Donner, R. V. (2017). Zealotry effects on opinion dynamics in the adaptive voter model. *Phys. Rev. E*, 96.

Kozlov, M. K., Tarasov, S. P., and Khachiyan, L. (1980). The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20:223–228.

Kurka, D. B., Godoy, A., and Von Zuben, F. J. (2016). Online social network analysis: A survey of research applications in computer science. arXiv: 1504.05655.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opin Q*, 80(S1):392–410. Publisher: Oxford Academic.

Leonard, N. E., Lipsitz, K., Bizyaeva, A., Franci, A., and Lelkes, Y. (2021). The nonlinear feedback dynamics of asymmetric political polarization. *Proc. Natl. Acad. Sci. U.S.A.*, 118(50):e2102149118.

Lerman, K. (2016). Information is not a virus, and other consequences of human cognitive limits. *Future Internet*, 8(2).

Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 915–924, New York, NY, USA. Association for Computing Machinery.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2–es.

Li, Y., Chen, W., Wang, Y., and Zhang, Z.-L. (2013). Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 657–666, New York, NY, USA. Association for Computing Machinery.

Liao, Q. V. and Fu, W.-T. (2013). Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2359–2368, New York, NY, USA. Association for Computing Machinery.

Liu, S., Ying, L., and Shakkottai, S. (2010). Influence maximization in social networks: An ising-model-based approach. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 570–576.

Liu, Z. and Weber, I. (2014). Is Twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *Social Informatics*, volume 8851, pages 336–347. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol*, 37(11):2098–2109.

Lynn, C. W. and Lee, D. D. (2016). Maximizing influence in an Ising network: a mean-field optimal solution. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2495–2503, Red Hook, NY, USA. Curran Associates Inc.

Mackin, E. and Patterson, S. (2019). Maximizing diversity of opinion in social networks. In *2019 American Control Conference (ACC)*, pages 2728–2734.

Macy, M. W., Ma, M., Tabin, D. R., Gao, J., and Szymanski, B. K. (2021). Polarization and tipping points. *Proceedings of the National Academy of Sciences*, 118(50):e2102144118.

Martin-Gutierrez, S., Losada, J. C., and Benito, R. M. (2023). Multipolar social systems: Measuring polarization beyond dichotomous contexts. *Chaos, Solitons & Fractals*, 169:113244.

Masuda, N. (2015). Opinion control in complex networks. *New J. Phys.*, 17(3):033031.

Masuda, N., Gibert, N., and Redner, S. (2010). Heterogeneous voter models. *Phys. Rev. E*, 82:010103.

Matakos, A., Tu, S., and Gionis, A. (2020). Tell me something my friends do not know: Diversity maximization in social networks. *Knowl Inf Syst*, 62(9):3697–3726.

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

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

Milgram, S. (1967). The small world problem. *Psychology Today*.

Minh Pham, T., Kondor, I., Hanel, R., and Thurner, S. (2020). The effect of social balance on social fragmentation. *J. R. Soc. Interface.*, 17(172):20200752.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, page 29–42, New York, NY, USA. Association for Computing Machinery.

Mobilia, M. (2003). Does a single zealot affect an infinite group of voters? *Phys. Rev. Lett.*, 91:028701.

Mobilia, M., Petersen, A., and Redner, S. (2007). On the role of zealotry in the voter model. *J Stat Mech Theory Exp*, 2007:P08029–P08029.

Monti, C., Cinelli, M., Valensise, C., Quattrociocchi, W., and Starnini, M. (2023). Online conspiracy communities are more resilient to deplatforming.

Moreno, G. R., Chakraborty, S., and Brede, M. (2021). Shadowing and shielding: Effective heuristics for continuous influence maximisation in the voting dynamics. *PLOS ONE*, 16(6):e0252515. Publisher: Public Library of Science.

Moussaïd, M., Brighton, H., and Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *PNAS*, 112(18):5631–5636.

Musco, C., Musco, C., and Tsourakakis, C. E. (2018). Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 369–378, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Mäs, M. and Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLOS ONE*, 8(11):e74516. Publisher: Public Library of Science.

Mønsted, B., Sapieżyński, P., Ferrara, E., and Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using twitter bots. *PLOS ONE*, 12(9):1–12.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc Nat Acad Sci*, 103(23):8577–8582.

Ngak, C. (2011). Then and now: a history of social networking sites. *CBS News*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol*, 2(2):175–220.

Nikolov, D., Oliveira, D. F. M., Flammini, A., and Menczer, F. (2015). Measuring online social bubbles. *PeerJ Comput Sci*, 1:e38. Publisher: PeerJ Inc.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press, Cambridge.

Notarmuzi, D., Castellano, C., Flammini, A., and et al. (2022). Universality, criticality and complexity of information propagation in social media. *Nat Commun*, 13:1308.

Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Polit. Behav.*, 32(2):303–330.

Osho, A., Waters, C., and Amariucai, G. (2020). An information diffusion approach to rumor propagation and identification on twitter. *arXiv:2002.11104 [cs, stat]*.

OSoMe (2020). Tracking public opinion about unsupported narratives in the 2020 Presidential election. *Indiana University Observatory on Social Media*.

Papanastasiou, E. and Giovanidis, A. (2023). Constrained expectation-maximisation for inference of social graphs explaining online user-user interactions. *Journal of Social Network Analysis and Mining (Springer)*.

Papasavva, A., Blackburn, J., Stringhini, G., Zannettou, S., and Cristofaro, E. D. (2021). "Is it a Qoincidence?": A first step towards understanding and characterizing the QAnon movement on Voat.co. In *Proceedings of WWW 2021*.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group.

Parsegov, S. E., Proskurnikov, A. V., Tempo, R., and Friedkin, N. E. (2017). Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87:925–979.

Peralta, A. F., Kertész, J., and Iñiguez, G. (2022). Opinion dynamics in social networks: From models to data. *arXiv:2201.01322 [nlin, physics:physics]*.

Peralta, A. F., Ramaciotti, P., Kertész, J., and Iñiguez, G. (2023). Multidimensional political polarization in online social networks.

Perra, N. and Rocha, L. E. C. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Sci Rep*, 9(1):7261.

Phadke, S., Samory, M., and Mitra, T. (2020). What makes people join conspiracy communities?: Role of social factors in conspiracy engagement. arXiv: 2009.04527.

Pougué-Biyong, J., Gupta, A., Haghighi, A., and El-Kishky, A. (2023). Learning stance embeddings from signed social graphs. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 177–185, New York, NY, USA. Association for Computing Machinery.

Proskurnikov, A., Matveev, A., and Cao, M. (2014). Consensus and polarization in altafini's model with bidirectional time-varying network topologies. In *53rd IEEE Conference on Decision and Control*, pages 2112–2117.

Qiao, M., Yu, J., Bian, W., Li, Q., and Tao, D. (2019). Adapting stochastic block models to power-law degree distributions. *IEEE Transactions on Cybernetics*, 49(2):626–637.

Ramaciotti Morales, P. and Cointet, J.-P. (2021). Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. In *RecSys '21: 15th ACM Conference on Recommender Systems*, Amsterdam, Netherlands.

Ramaciotti Morales, P., Cointet, J.-P., Muñoz Zolotoochin, G., Fernández Peralta, A., Iñiguez, G., and Pournaki, A. (2022). Inferring attitudinal spaces in social networks. *Soc. Netw. Anal. Min.*, 13(1):14.

Ramirez, L., San Miguel, M., and Galla, T. (2022). Local and global ordering dynamics in multistate voter models. *Phys. Rev. E*, 106:054307.

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

Rossi, W. S., Polderman, J. W., and Frasca, P. (2021). The closed loop between opinion formation and personalised recommendations. *IEEE Transactions on Control of Network Systems*, pages 1–1.

Sabin, S. (2020). 1 in 4 social media users say QAnon conspiracy theories are at least somewhat accurate. *Morning Consult*.

Santos, F. P., Lelkes, Y., and Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proc Natl Acad Sci USA*, 118(50):e2102141118.

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *J Comput Soc Sc*.

Schaewitz, L. and Krämer, N. C. (2020). Combating disinformation: Effects of timing and correction format on factual knowledge and personal beliefs. In

Disinformation in Open Online Media, Lecture Notes in Computer Science, pages 233–245. Springer International Publishing.

Schank, T. and Wagner, D. (2005). Approximating clustering coefficient and transitivity. *J. Graph Algorithms Appl.*, 9(2):265–275.

Shearer, E. and Matsa, K. E. (2018). News use across social media platforms 2018. *Pew Research Center*.

Shi, F., Shi, Y., Dokshin, F. A., Evans, J. A., and Macy, M. W. (2017). Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nat Hum Behav*, 1(79).

Shi, G., Proutiere, A., Johansson, M., Baras, J. S., and Johansson, K. H. (2016). The evolution of beliefs over signed social networks. *Operations Research*, 64(3):585–604.

Sikder, O., Smith, R. E., Vivo, P., and Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. *Sci Rep*, 10(1):5493.

Sood, V., Tibor, A., and Redner, S. (2008). Voter models on heterogeneous networks. *Phys. Rev. E*, 77:041121.

Spocchia, G. (2020). Engineer derails train near Navy mercy ship over coronavirus conspiracy theory. *The Independent*.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Bus Inf Rev*, 34(3):150–160.

Sprague, D. A. and House, T. (2017). Evidence for complex contagion models of social contagion from observational data. *PLOS ONE*, 12(7):1–12.

Starnini, M., Baronchelli, A., and Pastor-Satorras, R. (2012). Ordering dynamics of the multi-state voter model. *J. Stat. Mech. Theory Exp.*, 2012(10):P10027.

Suchecki, K., Eguíluz, V. M., and Miguel, M. S. (2005). Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution. *Phys. Rev. E*, 72(3):036132.

Tacchi, J., Boldrini, C., Passarella, A., and Conti, M. (2022). Signed ego network model and its application to twitter.

Takikawa, H. and Nagayoshi, K. (2017). Political polarization in social media: analysis of the "Twitter political field" in Japan. *BigComp 2017*.

Takács, K., Flache, A., and Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. *PLOS ONE*, 11(6):e0157948. Publisher: Public Library of Science.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge*. Penguin, New York, NY.

The Guardian, 2021 (2021). US lawmakers ask FBI to investigate Parler app's role in Capitol attack.

Tong, G., Wu, W., Tang, S., and Du, D.-Z. (2017). Adaptive influence maximization in dynamic social networks. *IEEE/ACM Transactions on Networking*, 25(1):112–125.

UK (2023). Online safety bill.

Vaccari, C., Valeriani, A., Barberá, P., Jost, J., Nagler, J., and Tucker, J. (2016). Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Soc Media Soc*, 2(3).

Vazquez, F. and Eguíluz, V. M. (2008). Analytical solution of the voter model on uncorrelated networks. *New J. Phys.*, 10(6):063011. Publisher: IOP Publishing.

Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2022a). Recommendation of content to mitigate the echo chamber effect. In *Conference on Complex Systems*, Palma de Mallorca, Spain. Extended abstract.

Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2023a). Opening up echo chambers via optimal content recommendations. In *Complex Networks and Their Applications XI*, pages 74–85, Cham. Springer International Publishing.

Vendeville, A., Guedj, B., and Zhou, S. (2021). Forecasting elections results via the voter model with stubborn nodes. *Appl Netw Sci*, 6(1):1.

Vendeville, A., Guedj, B., and Zhou, S. (2022b). Active links density in the Voter Model with zealots. In *Conference on Complex Systems 2022*, Palma de Mallorca, Spain. Extended abstract.

Vendeville, A., Guedj, B., and Zhou, S. (2022c). Towards control of opinion diversity by introducing zealots into a polarised social group. In *Complex Networks & Their Applications X*, pages 341–352, Cham. Springer International Publishing.

Vendeville, A., Guedj, B., and Zhou, S. (2023b). Discord in the multi-state voter model with zealots.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, Y. and Wang, Y. (2023). Opinion-aware influence maximization in online social networks.

Weatherbed, J. (2023). Twitter replaces its free API with a paid tier in quest to make more money. *The Verge*.

Weber, D., Nasim, M., Falzon, L., and Mitchell, L. (2020). #ArsonEmergency and Australia's "Black Summer": Polarisation and misinformation on social media. In *Disinformation in Open Online Media*, Lecture Notes in Computer Science, pages 159–173. Springer International Publishing.

Williams, H. T., McMurray, J. R., Kurz, T., and Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Glob Environ Change*, 32:126 – 138.

Wojcieszak, M. (2010). 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media Soc*. Publisher: SAGE PublicationsSage UK: London, England.

Xia, Y., Chen, T. H. Y., and Kivelä, M. (2020). Spread of tweets in climate discussions. arXiv: 2010.09801.

Yang, Q., Qureshi, K., and Zaman, T. (2022). Mitigating the backfire effect using pacing and leading. In Benito, R. M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L. M., and Sales-Pardo, M., editors, *Complex Networks & Their Applications X*, volume 1016, pages 156–165. Springer International Publishing, Cham.

Ye, M., Trinh, M. H., Lim, Y.-H., Anderson, B. D., and Ahn, H.-S. (2020). Continuous-time opinion dynamics on multiple interdependent topics. *Automatica*, 115:108884.

Yi, Y., Castiglia, T., and Patterson, S. (2021). Shifting opinions in a social network through leader selection. *IEEE Transactions on Control of Network Systems*, 8(3):1116–1127.

Yi, Y. and Patterson, S. (2019). Disagreement and polarization in two-party social networks. arxiv: 1911.11338.

Yildiz, M. E., Ozdaglar, A., Acemoglu, D., Saberi, A., and Scaglione, A. (2013). Binary opinion dynamics with stubborn agents. *ACM Trans Econ Comput*, 1(4).

Yildiz, M. E., Pagliari, R., Ozdaglar, A., and Scaglione, A. (2010). Voting models in random networks. *Proc. of 2010 Information Theory and Applications Workshop (ITA)*, pages 1–7.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33(4):452–473.

Zafeiris, A. (2022). Opinion polarization in human communities can emerge as a natural consequence of beliefs being interrelated. *Entropy*, 24(9).

Zareie, A. and Sakellariou, R. (2023). Influence maximization in social networks: A survey of behaviour-aware methods. *Soc. Netw. Anal. Min.*, 13(1):78.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, 395(10225):676. Publisher: Elsevier.

Zavala-Rio, A., Soerensen, C. A. G., Navarro, I., and Matía, F. (2013). An introduction to swarm robotics. *ISRN Robotics*, 2013:608164.

Zhang, Y., Wang, L., Zhu, J. J. H., and Wang, X. (2020). Conspiracy vs science: A large-scale analysis of online discussion cascades. arXiv: 2006.00765.

Zollo, F., Bessi, A., Vicario, M. D., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., and Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, 12(7):e0181821. Publisher: Public Library of Science.