university of groningen

University Medical Center Groningen

University of Groningen

# Opinion Dynamics in Online Social Media

Keijzer, Marijn

*DOI:*
10.33612/diss.196882523

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*
Publisher's PDF, also known as Version of record

*Publication date:*
2022

Link to publication in University of Groningen/UMCG research database

*Citation for published version (APA):*
Keijzer, M. (2022). *Opinion Dynamics in Online Social Media*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. https://doi.org/10.33612/diss.196882523

**Copyright**
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

**Take-down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

Download date: 28-05-2023

# Chapter 1

# The complex dynamics of opinions in online social media

This chapter includes sections of joint work with Michael Mäs that appeared in *Data Science* under the title 'The Complex Link Between Filter Bubbles and Opinion Polarization' (in press).

In this chapter I introduce the main themes of this dissertation on opinion dynamics in online social media. In particular, I argue that the complexity perspective can contribute critical insights into the social dynamics in these social environments. The chapters of this thesis are all linked to the complexity perspective in some way, and their (inter)relations are discussed. I conclude with some general remarks about and limitations of the chapters.

## 1.1 Prologue

It is almost midnight as the subway line bound westwards stops at station Coolhaven and opens its doors to let Nina in. The cabin is a little busier than usual, but has not been really busy anymore for a while now. Over a year has passed since the first COVID-19 case was discovered in Tilburg, not too far from here, and quite radically changed Nina's daily life and commute. She looks around and decides to take a seat opposite of an older Dutch man. He is wearing his facemask like a chin-diaper, slowly eating a kebab from a franchise chain located a few stops earlier. The man briefly looks up, but returns quickly to his original position as Nina sits down. His eyes rest on the phone in the palm of his right hand. She follows his example, pulling her new eco-friendly Fairphone from her vegan-friendly 'eat beans not beings' tote bag and unlocking it in one fluent motion.

Nina's social media feed is filled with comments and articles about last week's election. Disappointed by the result, she hadn't paid so much attention to the backlash that followed. The defeat of the political left had puzzled her. In the weeks leading up to the election, she felt that the left had gained momentum. Combating climate change would be the undeniable top-priority of politicians in the next term, and for almost everyone she knows, a thirst for more diversity in the parliament drove their vote for inclusive and progressive politics. Four years of Trump in the white house had certainly affected even Dutch politics. People saw what happened, right? They saw the havoc created by the world's first *Twitter-president*, whose term ended so dramatically with a violent attack on the Capitol. She would have never guessed that the Netherlands would take such a strong turn to the right. Didn't they see that the climate commercials of Sigrid Kaag's D66 were just a cheap trick to lure leftist yuppies into voting for what is actually a center-right party?

After a cinematographic showreel of incidents with mask-refusers, former late-night talk show host Robert Jensen opens today's episode of his vlog with a recap of the election. His disbelief is visible in his eyes. "I have enjoyed that evening. I had a good laugh. I have seen, even more clearly, how they play the game, and how they cheat it." Ben looks up from his phone to the woman who is sitting down opposite of him and takes another bite of his kebab. "This result is fake", Jensen continues, "because there is no way that this result could be real.". His fringe online show for 'the unheard voices in the Netherlands' was recently banned from YouTube, but still attracts a sizeable audience through his own platform

Ben was similarly perplexed by the outcome. Twenty-four seats for a party led by Sigrid Kaag, a woman who was, until a few months ago, virtually unknown in Dutch politics. Yet, she had been around the block in Europe, and had occupied a range of

positions in the UN. Someone with strong links to the political elite who made such an unlikely victory? It all seemed a little too suspicious for Ben.

A couple months before the election, Ben had started a hashtag criticizing Kaag after he'd seen her talking about the EU on national TV. The hashtag had taken on, and he received plenty of positive responses to it. His twitter following doubled in size. Getting noticed on social media gave him a sense of pride. Now, even national print media reported about the hashtag. On moments, he felt like an influencer, and noticed that the more extreme stuff he posted, the more engagement he received. A few times he was lucky with a scoop that he retweeted from some automated accounts he followed. Their messages seemed a bit gushy at first, but found a positive reception among Ben's fans.

The train comes to a bumpy stop as it approaches station Marconiplein. "Mathenesseweg, Spangen, Castle. Doors open to your left." The two strangers have been sitting opposite of each other this entire time. Silent. Absorbed in their own world that is being projected to them by their phone. Focused on the reality that social media created for them.

Online, they would never meet. Their tweets hidden by the different hashtags to which they link. Their messages never shared by the contacts in the handful of degrees that separate them. The content they consume invisible, through the algorithmic filter that refines their browsing experience.

But what if they talked? Nina could have addressed Ben's incorrect mask use. She would have corrected him, and an annoyed Ben would have started a heated discussion about the nonsense of wearing masks at all. Nina noticed the kebab, judging his meat-eating habits and ignorance to the climate crisis. She makes a pretentious remark underlining her wokeness, pissing off Ben who doesn't even consider the argument given by this wet-eared millennial. They might have tweeted about the experience with this political alien they just met. Their experiences would travel through their online social networks and trigger small changes in the views of their likeminded friends.

Or perhaps they would have just talked. They would have found common ground in their shared resentment of Kaag, or maybe they would have discussed a recent sports match, the weather, a joke. They shared a smile, largely hidden by the mask, but exposed by the crow's feet in the corners of their eyes. They would see that the other party isn't so bad and that we are all just trying to make sense of the world in our own ways, building on very different experiences, but with our heart in the right place.

## 1.2 Opinion dynamics in online social media

Online social media have changed the way we interact with and create an understanding of our social environment. The environment that, however slow or fast, shapes our views on the world, and defines the opinions we hold. This dissertation is concerned with these changes and contributes to the understanding of its causes and consequences, using a complexity approach and applying methods ranging from formal modelling to field experiments. Specifically, I focus on how the dynamics of opinions in large populations are affected by certain technological innovations we see in online social media, such as personalization algorithms or the spreading of misinformation by malicious social bots.

Pundits and scholars have frequently voiced concern about rising polarization and opinion fragmentation in Western societies (e.g. Abramowitz & Saunders, 2008; Vachudova, 2019; Pew Research Center, 2017; Finkel et al., 2020). Growing polarization—a high degree of differences between internally coherent groups of a certain ideology, groups of elites or sets of voters (McCoy, Rahman, & Somer, 2018)—is a phenomenon that a number of democracies have faced in recent years. The U.S. elections of 2016 and 2020, Brexit, and the rise of right-wing populism in many countries have demonstrated that elections leave many voices unheard. Stark contrast of opinions between winners and losers of these political battles do not only indicate rising polarization, they also leave both sides disillusioned, which, in turn, increases support for diverging forms of political activism (Maher, Igou, & van Tilburg, 2018). Moreover, the discontent of the unheard appears to grow, as does the bewilderment between groups in public debate (Iyengar, Lelkes, Levendusky, Malhotra, & Westwood, 2019).

The culprit is easily found as polarization rose in tandem with changing media consumption habits due to the growing popularity of the web. Particularly, online social networking websites changed the patterns of information diffusion, creating human and algorithmic informational filters, that lead to exposure bubbles and opinional echo chambers. Can we hold social media accountable for political polarization?

Understanding how communication in online social media affects processes of opinion formation has merit for three reasons. First, there are concerns about the impact of online communication on polarization (Bail et al., 2018). Online social media are used as an arena for informational warfare aimed at creating division and disagreement (DiResta et al., 2018). Brexit, the U.S. elections of 2016, the 2020 black lives matter protests, the revolutionary movements in the Arab spring, the 2019-20 Hong Kong protests, and the 2021 U.S. Capitol attack, to name a few, have all been marked as events where social media played an important role (Zhuravskaya,

Petrova, & Enikolopov, 2020; Munn, 2021; Howard et al., 2011). With perceptions of polarization in the U.S. steadily on the rise, worries exist that the digital revolution affected levels of polarization (Allcott, Braghieri, Eichmeyer, & Gentzkow, 2020). By now, CEOs of Twitter and Facebook have testified several times in the U.S. senate about social media's responsibility in the protection of democracy (Kang, 2020). Correlational evidence of rising popularity of social media platforms and political polarization, however, is not enough to prove that social media actually have a causal role in the emergence of polarization. Rigorous work aimed at understanding the mechanisms (Hedström & Ylikoski, 2010) that bring about polarization can fill this void.

Second, online social media provide us with an interesting lens to view human behavior, and, in particular, the place where purposive individual behavior may have unexpected drawbacks at the macro-level (Ruths & Pfeffer, 2014; Edelmann, Wolff, Montagne, & Bail, 2020). The technologies that affect opinion dynamics are here, whether we like it or not. The short history of the Internet has produced a handful of platforms where people discuss politics and share and consume news-related content. Poor system design may create backlashes that cause mass retreat of a certain service, but new ones will quickly fill the vacant spot. Whether it is Facebook, Twitter, MySpace, Gab, TikTok, Instagram, or Weibo, all social media platforms create a user experience that depends on similar technologies. Understanding how the individual technologies and their interactions shape the behavior of users is a lens through which we can learn about social behaviors in general (Lazer et al., 2020). The good intentions of social media companies may serve individual needs, whilst still creating negative consequences for society as a whole. It is therefore an epitome of the problem of unintended consequences of purposive social action (Merton, 1936).

Thirdly, the online information ecosystem is a malleable system. Problems created algorithmically, will often have algorithmic solutions as well. Sociologists tend to be reluctant to engineer social systems that have emerged historically. What is more, designing feasible, non-invasive, evidence-based interventions to solve social problems is a difficult task. The Internet, however, is a fully engineered system. Online, society appears more adaptive and the rules that govern its interaction more versatile.

How communication via social media affects the process of opinion formation is not so obvious. The random encounter between Nina and Ben, described at the start of this section, happens in the subway everyday, yet becomes very improbable in online ecosystems. On the one hand, friendship-based and algorithmically fenced-off online social media favor connections between similar people. On the other hand, the set of possible encounters is incomparably large and extends factors beyond the number of people that could even fit into the subway train. At the start of the digital

revolution, the Internet received praise for its potential to break barriers and connect people globally (Hauben & Hauben, 1997). Nowadays, social media platforms are criticised for their propensity to sort people into bubbles of their own; ostracizing and extremizing communities of individuals that would not have connected so easily offline.

Likewise, too little is known about whether online encounters between individuals with different opinions generate mutual understanding and opinion convergence or not. In particular, will the likelihood that interaction leads to a positive encounter, in which parties create understanding for, and maybe adjust their opinions ever so slightly towards each other, be affected by whether this encounter takes place online or offline? The mere fact of having a computer mediate the communication could help to find consensus by removing cues that cause stigmatization (e.g. the tote bag and the kebab in the meeting described in the prologue) (Postmes, Spears, Sakhel, & de Groot, 2001). Yet, reducing the other party to a name and avatar may make it easier to dehumanize the opponent in an argument. Perhaps social desirability and conflict aversion ease face-to-face communication, or maybe the absence of social identity markers online will remove incentives to (dis-)align opinions with the out- or in-group. Are these incentives different in on- and offline contexts and does this affect how people are influenced by and spread new information? How do they actually respond to political argumentation? And what spill-over effects do these responses have on our friends, and the friends of our friends?

Finally, the context in which these encounters take place may be a substantial element in defining the topic, communication structure, and appropriate response of an interaction. Whilst it may be technically possible to upload your performance of the choreography to Katy Perry's latest song to LinkedIn, people seem to rarely use LinkedIn to share such content. Your LinkedIn contacts would probably be equally confused about an invitation to a work-life balance seminar posted on TikTok. Platforms structure interactions in more ways than just through (unwritten) social rules about expected behavior on a platform. Technological options available to facilitate interaction can meaningfully change behavior too. Messages on some platforms are restricted in size, medium (text, image, video or spoken word), or topic. Platforms also vary in their operationalization of a connection. Some require reciprocated contacts, others allow users to accumulate massive followings. Other platforms do not even require one-to-one connections, but allow people to connect to themes or topics. Structure of interaction is also important. Are interactions supposed to happen between individuals or in a whole group? What if, instead of talking to each other, Nina had to speak to the whole train and convince everyone of her views all at once, similar to sending a message to all her contacts on social media?

She might convince almost everyone, but create a lot of aversion from the ones who disagree.

Understanding opinion dynamics in systems of autonomous but interdependent individuals calls for much more than an understanding of meeting opportunities, effects and structure of the interactions. Polarization is a phenomenon that manifests at the macro-level and should therefore be understood as the result of the many interactions in the social system as a whole. Sometimes these behaviors aggregate to the macro-level with simple rules, like counting the number of votes in an election. But sometimes, the aggregation from the micro to the macro is more complex than adding up votes. When the behavior of an individual is not only dependent on the desires and preferences of that single individual, but also on the availability of information and behavior in an agent's (local) environment, the system overall can appear unpredictable or chaotic (Mäs, 2018). The choice of Nina and Ben to engage in their online realities is probably not motivated by a wish to extremize and polarize. Yet, the repeated interactions with recommended posts from likeminded friends who share convincing content could put them in the situation where they do. Moreover, their friends who are consuming, digesting and producing information on a continuous basis as well will create an impossible chicken-and-egg situation about the relationship between polarization and ideological segregation (Keijzer & Mäs, in press). The high degree of interdependence between actors in social media limits our ability to reason about the dynamics of the system as a whole by studying individuals in it as if they were independent from each other. Such systems can be studied using the lens of complexity: the study of emergent behaviors in systems of interdependent components (Page, 2015).

Models of social influence have been developed to understand opinion dynamics in complex systems (Flache et al., 2017). The literature of mathematical and computational models is rich and diverse, but little is known about the interrelations of, and empirical motivations for certain formalizations (Flache et al., 2017; Edmonds et al., 2019). In this young field, a plethora of models has been developed to describe similar processes or macro phenomena. While empirical research is the only method that will allow one to identify which explanations and models have verisimilitude (Popper, 2005), empirical research always departs from theory. Comparison of alternative theoretical explanations of a given phenomenon is needed to identify those aspects where competing theories differ. Empirical evidence can then serve to exclude problematic explanations (Carley, 2019; Keuschnigg, Lovsjö, & Hedström, 2017; Sobkowicz, 2009). For the sake of theory formation, insights in generalizability of documented mechanisms through model comparison is a useful frontier to be explored (Flache et al., 2017; Squazzoni, 2012). By learning more about our theoretical

models, we collectively build a body of knowledge (Axtell, Axelrod, Epstein, & Cohen, 1996).

This dissertation aims to contribute to filling that void. The chapters in this book explore the empirical basis of formal models, exploit their power to develop a theory of online communication, present explanations for polarization and the diffusion of beliefs at the level of the entities that constitute them, and touch on the generalizibility and validity of those explanations.

In this chapter, I will discuss further why and how the complexity perspective contributes to our understanding of opinion dynamics in online social media (Section 1.3) and what methods researchers typically employ to achieve that goal (Section 1.4). The studies included in this dissertation all revolve around complexity in online social media in some way, but their relation may not be obvious at first. In my discussion of the complexity approach—the common denominator in all contributions—I will therefore highlight the significance of each chapter to the overarching research problem. First, however, it is important to define some central concepts of this dissertation. What do we mean when we talk about social media, and what parts of it are interesting for social scientists to understand? Also, what is polarization exactly, and why would we need to improve communication tools to prevent it?

### 1.2.1 Social Media

It did not take long for developers of very early versions of what is now called the Internet, to take note of its potential disruptive effects in the democratization of knowledge. For some, the project started as an ideological one. Decentralization of the production and distribution of information would remove information monopolies and the competitive advantage of (with)holding knowledge. Freedom of information would create opportunities for anyone to contribute to and benefit from the collective body of knowledge (Hauben & Hauben, 1997). In relation to politics and opinion formation, free access to the same information should simplify discussions and have a positive impact on a population's ability to create consensus. The web would act as the public sphere in modern societies: a place for public discussion and the resolution of disputes (Gimmler, 2001).

Fast forward fifty years and we see that social media platforms have emerged on the web that fulfill the role of information democratizers. Broadly, we can define *social media* as "[...] a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation

and exchange of User Generated Content." (Kaplan & Haenlein, 2010, 61).[1] Social media are platforms that, according to this definition, solely provide the digital infrastructure and technology needed to add (publicly available) information to the web, and disseminate that information to a set of contacts. Due to the broadness of this definition, social media includes more than just *social networking websites* (e.g. Facebook or Twitter) that have popularized the term. Kaplan & Haenlein (2010) distinguish platforms by the degree to which they offer opportunities for self-presentation or demand self-disclosure—the means through which people present themselves and their attitudes, according to Goffman's influential theory of symbolic interactionism (1959). On one end, we might find Instagram, a social networking website of which the users post images or videos that are commonly representations of the users themselves. On the other end, there is websites like Wikipedia. Wikipedia demands unambiguous contributions substantiated with external sources, leaving very little room for self-expression.

In this dissertation, I focus on social media that allow for at least some degree of self-disclosure for the simple reason that self-expression is a necessary condition for the spreading of behaviors or beliefs.

Exploring the relationship between online social media and opinion formation empirically is a difficult task. Using observational data, it is not obvious how we can unravel the mechanisms that bring about the patterns we see on a system's global level. Observations about the net effect of social media use on all sorts of political outcomes, for example, is at risk of misattribution of causality based on spurious correlations (e.g. through omission of confounding factors) or fail to discriminate between competing or complementary explanations (for a review on such effects, see Zhuravskaya et al., 2020). For example, for a recent publication in a popular scientific journal, a team of researchers gathered the dissemination trajectories of over 100 million pieces of information on four different social media platforms (Cinelli, de Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021). They attempted to assess the degree to which information spreads not freely but in politically homogeneous echo chambers. The authors argue that, due to the variation in platform design of social media, comparative analysis, highlighting the differences of the four platforms is needed (Cinelli et al., 2021). The lack of systematic comparisons criticized by Cinelli et al. makes it problematic to attribute observed differences to the design of the platform alone. The platforms vary on too many dimensions like target audience, communication structure, algorithmic design, etcetera, all at once, prohibiting the

[1] The definition coined by Kaplan & Haenlein mentions *Web 2.0*—a commonly used term to describe the shift to a user-centered web. Its predecessor was an internet that was substantively more centralized. The first set of popular websites were mostly big publishing companies that simply used the Internet as an extension of their service. A burst of technological innovation in both hardware and software made it considerably easier for users to build websites, and make their own voice heard.

establishment of valid statistical relationships. Let there be no confusion about the value of studies that test and monitor population-level effects of social media. They explore new scientific territory and signal or soften concern about potential harm. For sociologists, however, social media present just another social sphere in which we can learn about social behavior. By combining rigorous theoretical work and clever empirical design, we can exploit the opportunities of digital interaction to contribute to scientific knowledge (Lazer et al., 2009; Salganik, 2018; Flache, Mäs, & Keijzer, in press). If we want to learn about the way in which opinion formation is affected by social media, we need to disentangle the features that social media have that alter communication, and make comparisons between systems with and without those features.

One feature of social media that is worth investigating because of its attracted concern in recent years is personalization technology (Keijzer & Mäs, in press; Pariser, 2011; Bruns, 2019). Personalization algorithms are the rules or processes that decide what a social media user is exposed to when entering the platform. At any point in time, the information available to each user is sorted and presented to the user based on recency, and on previously expressed preferences, behavior on the platform, or behavior of others who are similar to the given user (Papakyriakopoulos, Serrano, & Hegelich, 2020). While Nina and Ben, from the prologue of this chapter, meet in physical settings, personalization technology is likely to prevent their encounters in online settings. Perhaps unknowingly, they find themselves in niches of the Internet where information is shown based on their previously signaled interests.

The goal of personalization algorithms is not to deceive users by hiding information, but to show information that increases user engagement. Despite the user-level benefits generated by personalization technology, there is growing concern about unintended negative consequences. For many users, the web is an important source for information on political, social, and cultural topics (Smith & Anderson, 2018). Criticizing personalization in this context, observers of the web warned that users are less exposed to content that challenges their own political opinions. Being insulated from competing views, you get "stuck in a static, ever-narrowing version of yourself – an endless you-loop" (Pariser, 2011). Users of online social networks complained that their online communities have turned into cocoons consisting exclusively of likeminded friends, which makes online communication increasingly boring (Pariser, 2011). In other words, personalization intensifies what sociologists labeled homophily, the tendency to interact with relatively similar others (Lazarsfeld & Merton, 1954; McPherson, Smith-Lovin, & Cook, 2001). The conjecture that personalization technology is responsible for the creation of echo chambers, where opinion differences between chambers are intensified, has been labeled the *personalization-polarization hypothesis* (Keijzer & Mäs, in press).

On the level of the individual too, online social networking websites also create meaningful differences for the diffusion of beliefs and opinions. They provide a context in which they act and react slightly differently. Research on computer-mediated interaction, for instance, has shown that people are more open to messages they are exposed to in the absence of stereotyping the sender of the message based on social information (Postmes et al., 2001; Postmes & Spears, 2002). The tote bag, fairphone and appearance of Nina quickly planted an image in the mind of Ben, probably making him less open to her arguments. Computer-mediated interaction research suggests that he would be more willing to listen to the same argument in the anonymity of simple, online, text-based communication.

Finally, individuals respond differently to the context they find themselves in online. Most obviously, by posting content that fits the medium, like posting work related information on LinkedIn and selfies on Instagram, but worry exist that reactivity does not stop there. The relative success of emotional or extreme content creates incentives to post this kind of information (Munger, 2020; Munger & Phillips, 2020). Deceptive strategies to push the success of content is so common nowadays, they received their own well-known term: clickbait. What is more, reactivity to the platform can also happen in local neighborhoods, where posting or remaining silent about discussion topics depends on network positions (Gaisbauer, Pournaki, Banisch, & Olbrich, 2021). After launching his mean hashtag, Ben found himself in a bubble where many people responded positively to his post, stimulating Ben to continue on this path.

### 1.2.2 Polarization

Political polarization is a broadly used term that is not free from ambiguity. Various technical definitions and operationalizations exist in the literature (Bramson et al., 2016). The term polarization can be used to refer to a static distribution of opinions, as well as to a dynamic process of growing differences (DiMaggio, Evans, & Bryson, 1996). As a static concept, it refers to the existence of two or more groups that are internally homogeneous and externally heterogeneous (Esteban & Ray, 1994). In other words, a polarized distribution of opinions exhibits a clustered pattern where there is substantial agreement within and disagreement between clusters. Polarization as a dynamic concept indicates a trend towards such a distribution.

Technical definitions aside, polarization is probably best described not only as the degree to which opinion differences exist, but also by the amount of agreement that exist within opinion groups on a broader range of topics (McCoy et al., 2018). Though collective decision making can benefit from diversity (Habermas, 1998b), some argue that the U.S. and some European societies have reached levels of polarization at which

social fragmentation is undesirably high (Klein, 2020). This fragmentation could threaten social cohesion and political stability (Baldassarri & Gelman, 2008; McCoy et al., 2018).

Where traditional, uni-dimensional measurements of polarization fail to capture emerging cleavages between opinion groups in society, other indicators of increasing opinion differences are needed. Some capture polarization of the political elite or the discourse in the media (Hetherington, 2009; Wilson, Parker, & Feinberg, 2020). Polarization, surely, must trickle down from influential broadcasters and talking heads, onto the audience they reach with their messages (Druckman, Peterson, & Slothuus, 2013; Slothuus & Bisgaard, 2020). Others leave measurements of opinions altogether, and focus on feelings of affection towards someone of a different party (Iyengar et al., 2019) or identity (Hobolt, Leeper, & Tilley, 2020), or on political effects of polarized societies such as voting for radical parties or engaging in political protests (Zhuravskaya et al., 2020). Yet a different approach uses the coherence between positions on multiple opinion items to measure ideological segregation (Baldassarri & Goldberg, 2014; DellaPosta, 2020; Keijzer & Mepham, 2021). Its logic is simple: the more alignment of attitudes within clusters, the less common understanding there will be between opinion clusters.

Empirical findings on the level of polarization in Western societies are mixed (Gentzkow & Shapiro, 2011; Pew Research Center, 2017). Studies that focused on unidimensional measures of polarization on a range of topics in the U.S. in the second half of the twentieth century show no sign of increasing polarization (Evans, 2003; DiMaggio et al., 1996). Among the elites in the U.S., polarization is growing, and spills over onto the general public, not so much in attitudes (Fiorina & Abrams, 2008), but clearly in affective polarization (Banda & Cluverius, 2018). With regards to affective polarization in general, there appears to be an increase in the U.S. in the past decades (Iyengar & Hahn, 2009; Iyengar et al., 2019), but trends are less clear in Europe (Boxell, Gentzkow, & Shapiro, 2020; Dekker & Den Ridder, 2019; Hobolt et al., 2020). Europe is, however, struck by increasing support for radical and populist parties (Winkler, 2019).

Theories pointing to potential drivers of polarization are abundant. For instance, growing inequality can lead to polarization because the have-nots are more likely to vote radical (Inglehart & Norris, 2016; Winkler, 2019), conceivably triggered by processes of relative deprivation (Rooduijn & Burgoon, 2018). Perhaps, over time, Nina and Ben will think more and more differently when they see that society is not fairly rewarding them for their own efforts, reducing their trust in the political center. Maybe their views will just follow those of opinion leaders. A party's attitude positions typically correlate with the positions of their voters, and appear to be able

to effectively pull their voters towards a new position when they need to (Slothuus & Bisgaard, 2020).

Polarization can also emerge from the bottom up. Many mechanisms exist that can produce polarization from a series of interactions at the micro-level without necessarily being intended by the actors (Flache et al., 2017). I will discuss these explanations in detail in Section 1.3.1, but want to stress the importance of homophily here. As noted by the 44th President of the United States, Barack Obama, in his farewell address: "For too many of us, it's become safer to retreat into our own bubbles, whether in our neighborhoods or on college campuses, or places of worship, or especially our social media feeds, surrounded by people who look like us and share the same political outlook and never challenge our assumptions. The rise of naked partisanship, and increasing economic and regional stratification, the splintering of our media into a channel for every taste—all this makes this great sorting seem natural, even inevitable. And increasingly, we become so secure in our bubbles that we start accepting only information, whether it's true or not, that fits our opinions, instead of basing our opinions on the evidence that is out there." (2017). Homophily is the tendency of individuals to interact with likeminded others (Lazarsfeld & Merton, 1954; McPherson et al., 2001). Most micro-level explanations of emergent polarization at the macro-level rely on homophily in some way. Homophily could determine openness to interaction (Deffuant, Neau, Amblard, & Weisbuch, 2000; Hegselmann & Krause, 2002), impact the effectiveness of the interaction (Jager & Amblard, 2005), dictate the direction of the opinion shift (Baldassarri & Bearman, 2007; Takács, Flache, & Mäs, 2016), or provide social cues about expected coherence of beliefs (Goldberg & Stein, 2018). Personalization algorithms contribute to this homophily principle in the same way, by promoting interactions based on similarity (Pariser, 2011; Bakshy, Messing, & Adamic, 2015; Keijzer & Mäs, in press). Former president Obama connects the increase in homophily to a tendency to accept information that fits our beliefs. Ideological segregation, as indicator of polarization, would thus be a direct consequence of exposure to our personalized social media feeds. Whether strongly homophilous systems can actually be linked to increasing polarization through promotion of interactions between individuals who hold similar beliefs, is a recurring theme in the chapters of this dissertation.

## 1.3 A complexity perspective

The study of opinion dynamics in online social media is the ideal-typical research problem for a complexity perspective, as it is concerned with two defining ingredients of complexity. First, a complex system consists by definition of multiple levels of analysis (Mäs, 2018). In online social media, there is the level of the individual user

who consumes, shares, adjusts, and generates content; there is the local level, the immediate informational neighborhoods in which individuals reside; and there is the collective level, the network of users. To explain the relationship between the dominance of social media platforms and polarization—a collective phenomenon emerging from the actions of many individuals—we need to learn how actors respond to and affect the context they are in. The research problem then becomes an archetypal question about the *micro-macro link* (Squazzoni, 2008). This conception of our research problem relates a famous argument by Coleman (1990) who argued that macro-level propositions should be derived from the micro-level actions of the units that make up the system. The micro-macro link is concerned with the way in which all action of individuals combine or aggregate to create a (by the micro-level entities unanticipated) macro-level phenomenon (Raub, Buskens, & van Assen, 2011). The second defining ingredient of a complex system are interdependencies between the entities on the micro-level. On the web, users do not act in isolation but they share information, respond to each other, and exert influence on each other's opinions. In fact, concern about social media related to individual filter bubbles or effects of echo chambers should be seen as concern about a change in the macro-level structure of interdependencies between users on polarization.

The complexity approach is quintessentially sociological. Sociology—the study of societal phenomena—deals with the problem of unintended, macro-level consequences of individual actions (Merton, 1936; Hedström, 2005). Understanding social systems where such unanticipated outcomes obtain is particularly challenging, as the units that comprise them act inconsistently, make errors, and have limited capacity of foreseeing the consequences of their actions (Merton, 1936). What is more, actors are reactive to the actions of others, and, in turn, alter the conditions for those others. In systems where the impact of such interdependencies exceeds the predictability of individual responses, the effects of individual error and behavioral inconsistencies are amplified, making the system hard to predict (Van de Rijt, 2019; De Matos Fernandes & Keijzer, 2020).

Sociological research that has to deal with emergent phenomena—collective patterns that are a consequence of the behavior of the individual-level entities but that are external to the behavioral patterns of these individual-level actors—is ubiquitous (Page, 2015). For instance, Schelling and Sakoda famously demonstrated that cities can segregate into black and white districts even when all inhabitants are tolerant to the presence of members of another ethnicity in their proximity (Schelling, 1971; Sakoda, 1971). In their models, agents accept to live in neighborhoods where their own ethnic group is in the minority. They leave their homes only when, for example, more than seventy percent of their neighbors belong to the other ethnic group. Cities segregate, despite this high degree of tolerance, because agents do not act in isolation.

Whenever an agent moves, it affects its old and new neighborhood, making its own group less represented in its old and more represented in its new neighborhood. These changes in the composition of its neighborhoods might convince its old and new neighbors who used to be satisfied with their neighborhood's composition, to also move away. Thus, every moving has the potential to spark chains of reaction that intensify the ethnic homogeneity of neighborhoods and foster differences between neighborhoods to a degree that is not intended by the individuals that give rise to this pattern.

Also opinion polarization can emerge from hard-to-predict processes of influence among interdependent individuals (Dandekar, Goel, & Lee, 2013; Mäs & Flache, 2013). Theories of individual opinion formation do not assume that people intend to live in a polarized world or that the use of online social media increases their motivation to intensify opinion differences to other users. In contrast, these models assume that users seek to be positively influenced by their communication partners. Yet, one possible mechanisms generating polarization is that their tendency to communicate with similar others creates a loop of influence and selection through which they reinforce their opinions until they find themselves with rather extreme opinions (Sunstein, 2002a; Dandekar et al., 2013). Thus, polarization can be an unintended consequence of exclusive communication among similar others.

While complexity science appears to contribute a critical perspective on opinion dynamics in online social media, the public and scholarly debate about personalization largely ignores the complexity of online communication. More than just a missed opportunity, the unawareness of this perspective in the literature can lead to inappropriate extrapolation of individual behavioral observation or misinterpretation of macro-level relationships. Complex systems are particularly vulnerable to such threats for two reasons. First, a typical characteristic of many complex systems is their capacity to trigger cascades of behavior. Seemingly innocent actions of individuals affect the conditions of the individuals that depend upon them, who, in turn, alter the conditions of others around them. If Ben and Nina did talk, their agreement might spill over to both their networks. They may tell others about their encounters, or change their social media presence. The small changes they make can trigger others to behave differently, who trigger others, who trigger others. In fact, theoretical as well as empirical research demonstrates this point, and shows that complex social systems of opinions or coordination can be in a state where even rare and random events can alter collective outcomes (Macy & Tsvetkova, 2015; Mäs & Helbing, 2020). The segregation models by Schelling and Sakoda, for instance, generate higher segregation when small amounts of randomness are added to the behavior of the agents. That is, it is added that also agents who are satisfied with their neighborhood may move and that the agents who are dissatisfied happen

to refrain from moving. It turns out that this randomness increases segregation, because every random moving by an agent has the potential to motivate further moving decisions by its old and new neighbors, potentially sparking a new cascade of segregation-increasing moving sequences (Van de Rijt, Siegel, & Macy, 2009).

A second characteristic of complex systems is that dynamics can be highly nonlinear. A typical example of a nonlinear dynamic on the web is the phenomenon that sometimes information goes viral (Weng, Menczer, & Ahn, 2013; Goel, Anderson, Hofman, & Watts, 2016). In such an event, content is suddenly shared by a huge number of users and diffuses through the network at exponential rates, creating bursts of attention that are notoriously hard to predict (Goel et al., 2016). There is also a debate about the linearity of the effect of personalization. In their study of Facebook users, Bakshy, Messing, and Adamic (2015) found that the homophily generated by Facebook's personalization algorithms is considerably smaller than the homophily resulting from users' own tendency to select content that supports their political orientation. We could dismiss the effect as a negligibly small change, but in a complex system this may not be true (Lazer, 2015; Mäs & Bischofberger, 2015). Increasing the temperature of water by one degree, for instance, usually does not have meaningful consequences, but it can trigger of a transition from liquid to gas when the temperature increases from 99 to 100 degrees Celsius. Likewise, it has been demonstrated theoretically that homophily can have a nonlinear effect on systems' tendencies towards polarization (Mäs & Bischofberger, 2015). A slight increase in the already high degree of homophily on the web may be enough to tip the system over, and cause polarization. This is because algorithmically increasing homophily has an effect on many users. What is more, even when only a few users were directly affected by personalization algorithms, the change in the information diet of these users will indirectly affect the information diet of their friends and the friends of their friends.

Here, I argue that in order to employ the complexity perspective, empirical and theoretical work needs to consider where and when emergent phenomenon may arise by recognizing the processes that may be triggered in the system. This is best achieved by analysis and reasoning at three levels of the system: the individual, the local and the global level. In the upcoming sections, I show how these three levels are relevant for understanding opinion formation, and discuss how the individual studies of this dissertation fit into that framework. Table 1.1 summarizes the three levels of analysis.

### 1.3.1 Individual level

The level of analysis that has certainly received most attention in the literature is the individual level. It is concerned with all processes that act within the sender and the receiver of communication in online social-networks. That is, it is focused on who is

**Table 1.1** Levels of analysis on the personalization-polarization hypothesis

| Level of analysis | Definition |
|---|---|
| Individual | The individual level relates to aspects of communication that affect how individual states change in response to changes in the local environment as perceived by the individual. |
| Local | The local level relates to aspects of communication that affect who is when encountering content emitted by whom. |
| Global | The global level relates to the structural characteristics of the communication network, elements that affect the parts of the system universally, and invariable conceptualizations of system components. |

emitting what content, to whom, and when. In addition, it matters who is exposing themselves when to online content and how this content affects the opinions of the target of communication.

Models of opinion dynamics demonstrate that alternative assumptions about how users update their opinions can lead to markedly different conclusions about whether communication through online social media increases or decreases polarization (Flache & Macy, 2011b; Flache et al., 2017; Mäs & Bischofberger, 2015). In particular, reinforcement models (Mäs & Flache, 2013; Dandekar et al., 2013; Banisch & Olbrich, 2019) and negative influence models (Macy, Kitts, Flache, & Benard, 2003; Flache & Mäs, 2008; Salzarulo, 2006) imply competing predictions about the conditions under which polarization emerges in systems with a high degree of personalization.

The central assumption of reinforcement models is that individuals with opinions leaning towards one of the poles of the opinion scale will develop more extreme views after communication with likeminded individuals (Myers, 1978; Sunstein, 2002a; Dandekar et al., 2013). Following the well-documented *group polarization* phenomenon (Myers, 1978), modelers proposed the Persuasive-Argument Theory, a psychological theory assuming that humans communicate arguments underlying their opinions (Mäs & Flache, 2013). Individuals may hold a nuanced opinion themselves, but can only convey arguments that support or oppose an issue. During communication with likeminded individuals, users of online social networks will be mainly exposed to arguments in line with their own opinions. This, it is argued, reinforces their views and, thus, leads to more extreme opinions. Communication with users holding opposing opinions, in contrast, leads to opinion shifts in the opposite directions, as users are exposed to arguments challenging their opinions. The reinforcement of opinions also follows from biased-assimilation theory (Dandekar et al., 2013) and reinforcement-learning theory (Banisch & Olbrich, 2019).

Reinforcement of opinions is a central assumption underlying the personalization-polarization hypothesis (Pariser, 2011; Mäs & Bischofberger, 2015; Keijzer & Mäs, in press). As personalization of online services increases the exposure to likeminded users and content that is in line with one's own views, web users with opinions leaning towards the left end of the opinion spectrum would develop more leftist opinions and users with rightist opinions shift further towards the right. On the global level, this aggregates to increasing levels of opinion polarization.

Models assuming negative influence (sometimes called rejection or repulsion models) on the other hand, make alternative micro-assumptions and imply markedly different macro-predictions (Macy et al., 2003; Jager & Amblard, 2005; Salzarulo, 2006; Mark, 2003). Similar to the reinforcement models, negative influence models also assume that most of the time individuals tend to grow more similar to likeminded individuals. These models typically assume that users convey their position on an opinion continuum rather than exchanging arguments as is assumed by reinforcement models. Furthermore, it is added that individuals tend to dislike communication partners holding very distant views. Seeking to increase dissimilarity to persons they dislike, individuals adjust their opinions away from their communication partner to decrease cognitive dissonance (Festinger, 1964), or to signal their possession of a certain social identity (Tajfel & Turner, 1986).

Negative influence models contradict the personalization-polarization hypothesis (Mäs & Bischofberger, 2015). As personalization leads to fewer encounters between users who hold opposing views, rejection of distant opinions is an increasingly unlikely event. Over time, users who hold the most extreme opinions engage in interactions with communication partners who are similar, but a bit less extreme, little by little pulling even the most extreme agents towards consensus. Negative influence models thus predict that an increase in web personalization will decrease opinion diversity over time. Closely related arguments have been made for effects of network or spatial segregation (Flache & Macy, 2011a; Feliciani, Flache, & Mäs, 2020; Flache, 2019), whose effects are akin to personalization technology.

In a nutshell, depending on whether one assumes negative influence models or reinforcement models, one will come to the conclusion that personalization either decreases or increases polarization. Thus, empirical research is needed to identify the most plausible combination of micro-assumptions (for a given context), calibrated to the behavior that individuals display.

Chapter 2 is concerned with responses to argumentation in online interaction. I investigate *how the perceived distance to a sender moderates how individuals adjust their opinion after exposure to an argument*. The inverse predictions that negative influence and reinforcement models make under weak and strong levels of personalization

are demonstrated in Figure 2.1 of Chapter 2, and motivate the design of a formal framework for distinguishing between positive and negative influence.

Empirically, previous attempts to find whether a threshold for negative influence exists often ran into (one of) two important methodological problems. First of all, many studies used students samples or set the topic of discussion to a fictional or strongly uncontroversial discussion topic. In both cases, the expected opinion distribution on the outset has little variance, resulting in small pre-interaction opinion distances. Undersampling of dyads with larger opinion distances—precisely the dyads where we would expect negative influence to occur—leads to less power to find negative influence. I avoided those issues by sampling politically interested individuals on Facebook, and exposed them to topics where I expected them to be opinionated already. Secondly, most statistical models do not account for censoring at the extremes. While the formal models I use to model negative influence do model negative influence as a process where individuals can feel repulsion that pushes them beyond the extreme of the opinion bound, our statistical models have not followed suit. Using Bayesian models for the censored data I account for the bounded nature of opinion measurement.[2]

I then recruited over four hundred respondents on Facebook for two experiments and exposed them to arguments about two political topics. I find that, on average, argument exchanges lead to opinion shifts that increase agreement, though effects of the stimuli are small. When confronted with an opinion that was considered by the respondent to be very far from the respondent's own opinion, the respondents did sometimes adjust their opinion negatively. In other words, their opinion became more extreme, increasing distance to the opinion of the sender of the message. What is more, these negative shifts occurred more frequently when they perceived the sender to be of a different political orientation or when the sender used language that signaled moral values of the political opposites. This suggests that if there was little personalization technology preventing interactions between users who disagree, online social media might be a much more hostile place.

Opinion shifts in response to argumentation are surely not the only interesting aspect to be understood at the individual level. In recent years, there has been plenty of work done on the individual level of politics and social media such as, for example, effects of informational overload on acceptance of misinformation (Pennycook & Rand, 2019), aggravation and outrage (Crockett, 2017), friendship formation and political disagreement (Yang, Barnidge, & Rojas, 2017), and moralization of discussion topics (Mooijman, Hoover, Lin, Ji, & Dehghani, 2018). Though they fall outside of the scope of this dissertation, I would like to note here that the insights from these valuable contributions would—where they do not do so already—benefit from adopting a

[2] Appendix A.1 includes an in-depth analysis of both issues.

complexity perspective. Just like the dynamics of opinion change, the individual decisions create externalities that affect the decision of others. Whether it would be spreading misinformation, moral sentiment, or outrage, or whether it is direct modification of the friendship network, these individual decisions will interact with each other and with processes at the local and global level of the system.

### 1.3.2 Local level

The local level of observation is concerned with all mechanisms that govern the sharing of information in individuals' direct network neighborhoods. In the context of online social networks, this refers mainly to the technical implementation of communication and personalization. Unlike individual-level factors, local-level aspects are external to the individual sender or receiver. That is, these technical aspects do not affect how senders of communication emit online content and how receivers respond to communication. Local-level aspects change who is when encountering online content emitted by another user. It turns out that even seemingly small technical aspects can have strong effects on collective opinion dynamics and can change the effects of personalization technology.

Chapter 3 of this dissertation shows how a minor variation on a local-level characteristic of opinion dynamics models has major implications for the system as a whole. The chapter was motivated by an interest in *how technological decisions on structure of communication affect the dynamics of the spreading of beliefs*. Implementing a different communication rule, or regime, illustrates how local-level factors might interfere with the effects of personalization technology on polarization dynamics.

On many online social media platforms users emit messages to all of their friends or followers at the same time. This so-called *one-to-many communication* differs from the one-to-one communication implemented in most opinion-dynamics models developed for offline contexts (Flache & Macy, 2011a; Keijzer & Mäs, 2021). Intuitively, the difference between one-to-one and one-to-many communication may seem to be trivial. At first glance, a one-to-many communication-event is not much more than a sequence of one-to-one communication events. Yet, modeling work drawing on and extending Axelrod's seminal model for the dissemination of culture (1997) demonstrated that one-to-many communication fosters opinion fragmentation and social isolation in opinion systems (Keijzer, Mäs, & Flache, 2018). One-to-many communication is only one aspect, but its strong and counter-intuitive effects illustrate that even seemingly innocent local aspects can have substantial effects on the polarization dynamics.

The one-to-many effect obtains because it introduces negative consequences of disagreement. These negative consequences are at the core of the argument that is

formalized and analyzed in Chapter 3. When agents interact on a one-to-one basis, any decision not to accept influence of a sending agent does not change the state of the system, hence the between-agent dissimilarity remains unchanged for all pairs of agents in the model. One-to-many interaction, on the other hand, makes agents who do not accept the message more dissimilar from those contacts who do accept the message.

Consider the example from the prologue. If Nina were to speak up to Ben, to try to get him to wear his mask properly, a dismissive response would not alter his (experienced) social distance from the others in the metro carriage. However, if Nina stood up and announced to everyone who would hear her that people should wear their mask, she introduces a (small) division between those who accept the message and those who would not. This externality may still seem like a small change, but its effect amplifies in complex systems. The combination of the emergence of small divisions between and ample positive influence within clusters can transform these small divisions into large cleavages over time.

I analyzed the simplest, four-actor case using Markov Chain analysis, and proved formally that isolation is substantially more likely under one-to-many communication. Simulation work demonstrated robust differences between one-to-one and one-to-many communication also in much bigger networks, in particular in networks characterized by high transitivity and high node degrees (Keijzer et al., 2018). One-to-many communication increases the chances that individual agents are isolated and that multiple internally homogeneous but mutually distinct subgroups form.

Another mechanism that nicely demonstrates a local-level mechanism is the *strength-of-weak-bots* effect described in Chapter 4. In this chapter, I attempted to *explain the contradiction that social bots appear to be effective on the global-level while the direct effects on the sparse contacts they have seem feeble*. Social bots—automated social media accounts devised to influence opinion dynamics—have been identified as a threat to democracy in digitalizing societies (Ruths, 2019). Empirical studies show a strong presence of bots (Williams et al., 2020), and considerable effectiveness on, for example, the spreading of misinformation (Vosoughi, Roy, & Aral, 2018). Yet, bot accounts tend to differ from human accounts in important ways (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Most importantly, their connectedness to human social media users is limited (González-Bailón & De Domenico, 2020), as is their direct effect on individual beliefs (Bail et al., 2019).

The social influence literature offers an interesting explanation for the bot effectiveness puzzle (Hegselmann & Krause, 2015; Keijzer & Mäs, 2021). Again, using Axelrod's model for the dissemination of culture, I demonstrate in Chapter 4 that bots are more effective when they are less active and weakly connected. By adding a bot agent—an agent who cannot be influenced, but stubbornly sends the same

messages—to a simple social-influence model, I could measure the effectiveness of the bot as the share of agents who grew similar to the bot in equilibrium. The analyses expose a mechanism that highlights the significance of indirect social influence. The strong bot, it turned out, is very effective at convincing a small cluster of contacts who already agreed with the bot quite a bit, and pulls them close rapidly. Once this cluster is formed, there is little interaction possible between the agents who reside in that cluster, and their contacts who do not. The weak bot, on the other hand, slowly nudges his contacts, leaving opportunities for them to propagate the bot's message to their indirect contacts, who propagate it even further.

Both examples demonstrate that a simple change to theoretical models of opinion dynamics affecting local-level dynamics can have serious effects on the whole system, without introducing additional assumptions on agent-heterogeneity or behavioral complexity. By no means do these explanations eliminate the need for explanations on the individual or global-level. Existing explanations for the spreading of bot-infused content or isolation and extremization are not invalidated by the work here. Rather, the strength-of-weak-bots effect and isolation from one-to-many communication are parsimonious relational alternatives to some of the explanations known to date (DellaPosta, Shi, & Macy, 2015). What is more, they may interact with processes at the individual or global level. For example, individual differences in likelihood to accept misinformation will affect bot effectiveness, particularly when more gullible users are clustered in the network (Aral & Walker, 2012; Pennycook & Rand, 2020). Both effects also interact in interesting ways with global level characteristics, as I will sketch in the next section and show in more detail in Chapters 3 and 5.

### 1.3.3 Global level

The global level refers to all structural elements of the communication network as a whole. Global level characteristics are characteristics that do not vary between individuals, or between local neighborhoods in the network, but describe peculiarities of a certain complete social system or describe commonly shared beliefs about the state of the world more generally. With regards to opinion systems in online social media, they may reflect system design decisions, such as the decision to implement undirectional ties (like on Facebook), directional ties (Twitter), or no ties at all (Reddit), but they can also reflect our beliefs about what is discussed (e.g., a rich cultural profile or a uni-dimensional opinion).

In Chapter 5, I present a comparison of metric and nominal opinion models by evaluating the robustness of the one-to-many communication mechanism to our beliefs about the nature of opinion models. I wondered *whether the effects of one-to-many communication on consensus formation generalize to other opinion dynamics*

*models*. Models of social influence can be categorized into models where agents hold nominal or metric beliefs (Flache et al., 2017). Some argue that metric models are more appropriate for modeling opinions, as they naturally order agents on a (left-right) continuum (Lorenz, 2007), express a degree of trust in a considered opinion (Chatterjee & Seneta, 1977), or represent some belief about an unknown quantity or probability (DeGroot, 1974). Agents can favor or oppose a certain political solution, and those preferences can be expressed in degrees of conviction. Models with nominal opinions cannot order agents in that way. They are better suited for situations where agents can differ, but the extent of their difference cannot be expressed numerically (Flache & Macy, 2006), for example, when something could either be true or not true, or when multiple solutions for a problem exist and individuals can only propose one. There is no obvious reason to favor the nominal over the metric conceptualization, or vice versa. It may just be that some models are more appropriate for the study of some systems, and others for others. In some cases, whether an opinion space is nominal or metric could be imposed by global-level institutional features, such as differences between voting systems. Many voting systems allow only a single choice out of a distinct set of options and can hence be considered 'nominal'. Some others allow 'metric' expression of support by ranking options (e.g. positional voting, as it is, for example, used for jurors in Eurovision Song Contest), or ask voters to express their degree of approval for a list of options (e.g., score voting).

The analyses show that the effects of one-to-many communication observed in the nominal opinion world generalize to models with metric opinions, though some interesting differences exist between the models. For example, the metric models showed much less agreement between agents in equilibrium, overall, and the proportion of runs that end in complete consensus was negatively related to the size of the network. A surprising result, considering that scaling has the exact opposite effect in nominal opinion models with one-to-one communication, such as the model for the diffusion of culture (Axelrod, 1997; Flache, Macy, & Takács, 2006). In Axelrod's model, the size of the network is directly related to the degree of diversity on the outset, as well as to the number of interactions that need to take place before reaching equilibrium. Both of these factors increase the probability that, at some point, a trait will propagate through the whole network, convincing many along the way and making them ever so slightly more similar. These increases in similarity, then, manifest in equilibrium as increased likelihoods to converge to a single culture (Flache et al., 2006).[3] The metric, bounded confidence models do not exhibit the same kind of behavior. Bounded confidence models are models with a continuous opinion dimension, where interaction only occurs when the opinion difference between agents

[3] An interesting exception to this regularity occurs in models with many-to-one influence, where the relationship between network size and consensus flips around (Flache & Macy, 2011a).

is smaller than a *confidence bound* value. There, opinion changes propagate only locally, until they reach the sharp boundaries that divide clusters of different opinions. Scaling, it seems, has less pronounced effects in metric opinion models with bounded confidence thresholds. The one-to-many communication effects observed in nominal models, thus, do not replicate unconditionally. The assumptions I make at the global level about the cultural profile of agents, as well as about the size of the network or complexity of the opinion profile under study, interact with the expectations from one-to-one versus one-to-many communication.

Another global-level aspect that creates interesting interactions with effects at other levels is personalization technology. As discussed earlier in this chapter, the algorithmic sorting of relevant information that is so distinctive of online social media could be seen as a global-level characteristic, adding another layer of homophily to the system (Geschke, Lorenz, & Holtz, 2019). We've already discussed how personalization technology may prevent the emergence of polarization through avoidance of negative influence ties, but its impact on systems of social influence does not stop there. For example, the strength-of-weak-bots effect generally disappears with strong homophily (Flache et al., in press). As higher between-agent similarity is required for agents to interact successfully, the bot's ability to persuade large populations through indirect social influence gradually disappears.

Structural elements of the system may interact with individual or local-level processes as well. For instance, in Chapter 3 I argue that the amount of clustering in the network is positively related to the strength of the one-to-many effects. Here, clustering is defined as a global-level characteristic that expresses the proportion of closed triads over all connected groups of three nodes. Colloquially, transitivity expresses the probability for each node to have two of their friends sharing a tie as well. Since isolation occurs as a negative externality of rejection in social influence situations where others around you do accept a message, the amount of transitivity locally explains the degree to which an agent who rejects a message experiences alienation from their social environment (Keijzer & Mäs, in press). Clustering in online social media happens partly as bottom-up process—reflecting the clustering of real-world social networks—and partly by design. The latter explains why differences exist in the amount of clustering on different types of social media (Malik & Lee, 2020), opening the door to fruitful between-platform comparisons.

The examples in this section illustrate interactions that may occur between global-level, and individual and local-level processes. They highlight the relevance of thinking about contextual factors that serve as necessary requirement for, or have a direct effect relationship with the proposed mechanism, even when the ability to test those relationships empirically is limited. Empirical studies of global-level phenomena such as, for example, social media use, fall short of the counterfactuals

needed to study global-level aspects directly. Causal inference is rarely possible with observational data. The studied differences between various causes that appear 'in the wild' are typically not completely at random. That is to say that between cause and effect there is often an (unobserved) confounder explaining variance in both. The inferred relationship then remains one where we can only identify *causes of effects* rather than *effects of causes* (Pearl, 2015). For that reason, studying the effects of particular features of social media on global-level outcomes is a hard problem. We simply do not have multiple versions of Facebook that vary in the degree to which their users create transitive ties. In those cases, theoretical explorations using, for example, agent-based modeling could be helpful. I will discuss this, and other methods for studying opinion dynamics, in the next section.

## 1.4 Methods for studying opinion dynamics

The popularity of the world wide web did not only attract the attention of social research for its profound effects on social life, but also for its ability to record and trace social behavior that is normally hard to observe. The relatively new field of computational social science now flourishes, exploiting the possibilities offered by "data with an unprecedented breadth and depth and scale" (Lazer et al., 2009). Not only can social scientists nowadays use online settings to easily scale up traditional research methods like experiments or surveys, they can also profit from digital traces of human behavior through existing databases (Lazer et al., 2020).

The data-driven approaches in computational social science will yield most scientific fruit when combined with rigorous approaches to theory construction (Flache et al., in press). To understand human behavior, a purely empirical approach can sometimes be problematic. Secondary observational data obtained on the web is vulnerable to incompleteness, nonrepresentativity, system drift (i.e., changes in user population or to a platform's design over time), or algorithmic confounding (Salganik, 2018; Goldthorpe, 2021). Establishing causality in complex systems—a hard problem in empirical social science anyway—can become tricky when not all variables are controlled by the researcher.

Where empirical analysis falls short, formal modeling can provide insight by formulating competing expectations derived from multiple internally coherent and logically consistent explanations. Traditionally, theory building in sociology is done using inductive—starting from the observation and generalizing to an internally coherent theory—or deductive methods—through falsification of existing theory. Formal modeling for the social sciences offers a third approach: theory construction using generative models (Epstein, 1999). Formal modeling tools are used to provide parsimonious explanations that start from actors who are responsive to other actors

around them as well as to the environment or greater social context in which they are situated. They allow the formulation of relational alternatives to factor-based explanations (DellaPosta et al., 2015; Macy & Willer, 2002). Researchers build theories inductively, focusing on the mechanisms that bring about a certain macro-level regularity, and provide hypotheses that can falsify or discriminate models deductively.

Agent-based models (ABMs) are a tool well suited to studying complex systems of interdependent actors (Squazzoni, 2012; Macy & Flache, 2009). ABMs are computer simulations centered around autonomous, interdependent, heterogeneous, and embedded agents (Macy & Flache, 2009). These agents have both a social and a cognitive structure (Gilbert & Troitzsch, 2005). Stripped to the bare necessities for the process under study, ABMs are typically used as a simple representation of humans in a social and institutional context. Rather than focusing on society-level factor explanation, ABMs are inherently actor-based (Macy & Willer, 2002). Following the method of decreasing abstraction (Lindenberg, 1992), the field of agent-based modeling collectively attempts to create increasingly realistic models of social behavior, while keeping touch with the mechanisms through which certain social outcomes can be generated (Epstein, 1999).

Creating common ground in a field that knows a great diversity in model variations is challenging. Models of social influence processes have been around since as early as the 1950s (some early contributions include French, 1956; Harary, 1959; Abelson, 1964). The field generated rich insights into complex systems, but building on each other's insights has not been easy. To be able to compare findings, researchers need to align their models precisely (Axtell et al., 1996). The process of aligning models requires the researcher to keep all the elements between a well-understood model and the novel extension the researcher wishes to study unchanged, except the extension of interest. This is easier said than done, as there are many degrees of freedom when programming, and it may not at all be obvious when a small difference of assumptions matters and why. Yet, model comparison of aligned models has been identified as a major frontier in social-influence modeling research (Flache et al., 2017). In a recent review of the field, prominent social simulation researchers stated that "modelers need to identify the critical assumptions and predictions of their models, and need to compare these assumptions as well as their formal implementation to existing models." (Flache et al., 2017).

Chapter 5 introduces a novel piece of software specifically designed for the creation and systematic comparison of ABMs of social influence. The discrete-event framework for social-influence modeling, or *defSim* for short, is a Python-based software package that can be used to design, replicate and compare models (Laukemper, Keijzer, & Bakker, 2019). The software is designed based on seven

principles that all discrete-event models of social influence follow, and has been implemented using principles of modular programming (Martin, 2003; Gamma, Helm, Johnson, & Vlissides, 1995). Assumptions that adhere to the principles formulated in defSim can be mixed and matched to create hybrid models to assess the generalizability of insights generated by those models, and to validate and calibrate models against empirical data.

To enrich and calibrate models of collective behavior, running experiments on the behavior of individuals—the primary source of agency in research on micro-macro problems—can be a fruitful avenue. Experiments to observe individual behavior in isolation (in lab-like settings) or in the field (using field experiments) are tools at the disposal of those who aim to establish clean, causal relationships. Chapter 2 takes on the challenge of analyzing the behavior of individuals in isolation using a lab-on-the-web experiment. I introduce a novel design that combines the strengths of lab-experiments—with pre- and post-treatment measurement and full control over the stimuli—and field experiments—by recruiting our respondents on social media. The results from the experiment were then fed into the individual behavioral part of a simple ABM. This model illustrates how polarization or consensus emerges from the interplay of personalization technology and repeated interactions of social media users.

## 1.5 Conclusion & outlook

In this chapter I have laid out the social and scientific relevance of the thesis overall, and defined the core concepts used throughout the book. We have seen that social media affects the dynamics of opinions and beliefs in various ways, and I have argued that truly understanding the problem demands the disentanglement of those processes one by one. Moreover, understanding the processes in the system as a whole, requires a multilevel approach—looking at the individual, local and global level of complex systems—and awareness of all mechanisms at play in the system at once.

In the chapters that follow, I will zoom into specific research questions around the central theme of the thesis, but I would like to address their main conclusions and limitations, in short, here.

In Chapter 2, I present a framework for modeling opinion shifts dependent on the receiver's perceived opinion distance to the sender of a message. This framework can be used to distinguish between various notions of influence responses, and is translated to a statistical model using Bayesian models for the analysis of censored data to account for the bounded nature of opinion scales. In two experiments with Facebook users, I find that perceived distance indeed affects the opinion shifts

reported by the respondents. The estimated model is a direct representation of the models used in ABM research on social influence, and can, hence, be used to empirically calibrate those models directly. Those estimates indicate that negative influence—adjustment of the opinion away from the source of influence—is plausible, though the magnitude of the experimental effects overall is small.

The two experiments prompt an array of questions for future research. For example, what contextual factors trigger negative influence? Or how does the choice of topic moderate the shape of the influence function? The chapter introduced a statistical method for the analyses of response functions, which leaves a degree of uncertainty in the obtained results. More applications of the same method in other contexts will contribute towards the understanding of the method, as well as towards the understanding of the substantive questions raised about social influence on political topics. A more detailed and precise representation of the cognition of actors involved in opinion exchange, will benefit mathematical modeling work of opinion dynamics too, as the method employed here maps precisely onto the behavioral rules of agents in common representations of opinion formation in agent-based modeling. Furthermore, the calibration of such models to network structures of online social media will yield ABMs that would be able to generate hypotheses about macro-level consequences, testable with observational data from social media platforms themselves. Lastly, with regard to the discrimination of negative influence and reinforcement models, the development of an experimental scenario where the model from Chapter 2 and the persuasive argument variation are compared empirically can be considered low-hanging fruit. Following the framework I proposed here, such a comparison is possible, and would provide an important contribution to the modeling literature.

Chapter 3 is concerned with understanding an important difference between communication in online versus offline social networks. The simple addition of communicating to all network neighbors at once—resembling the sharing of information so typical of many social media websites—rather than the conventional one-to-one interaction procedure, dramatically changes the dynamics of models of social influence. Using Markov-chain analysis, a proof is provided for the heightened propensity of one-to-many communication to generate isolation and polarization. Agent-based social simulations are used to show that those results persist in larger populations. Furthermore, the generalizibility of those results to models of metric opinions is established in Chapter 5.

Although the proposed one-to-many effects show a high degree of internal validity and generalizability, whether they occur in the wild is an open question. Most likely, the strict isolation we observe in the model's equilibria is not an observable outcome in real life. However, a small tendency to alienate contacts who disagree

with an emerging consensus in their local neighborhood would be possible to observe experimentally, and could have profound effects on social fragmentation in online social media in the long run. Experimental tests of repeated opinion exchange under both communication regimes would strengthen confidence in the conjecture that the one-to-many effect is a force to consider in online communication.

Another local-level phenomenon of interest to the dynamics of opinions in online social media is the presence of social bots, discussed in Chapter 4. Social bots—automated social media accounts that try to influence public debate—have attracted concern from pundits and scholars due to their overwhelming online presence. Yet, their direct effects on the few contacts they have appears to be limited. I show in this chapter that bots can actually be more effective when their direct influence is small, as this, seemingly paradoxically, gives their connections more opportunity to influence those who do not receive messages from the bot directly. The strength-of-weak-bots effect highlights the importance of the complexity approach, as the direct effects of the bot are strictly at odds with its effectiveness in the long run through indirect influence.

The strength-of-weak-bots effect is an interesting, counter-intuitive perspective on the spreading of misinformation, but its purely theoretical nature demands empirical validation. The use of large-scale digital trace data on spreading of information in systems tainted with malicious bots might allow for comparative research that could test the proposed mechanism. What is more, the assumptions about openness to influence from bots related to the perceived distance to the bot, could be tested in experimental settings too. For example, verification of the assumption that individuals accept information from bots and from human connections at a comparable rate would strengthen the conclusions from the theoretical exploration in Chapter 4.

Finally, besides presenting a discussion and test of the generalization for the one-to-many effect, the main contribution of Chapter 5 is the presentation of foundational principles of social-influence models, and an accompanying software tool to create and explore such models. This chapter answers the call from social simulation researchers for a holistic view on the literature, and a deeper understanding of the literature's insights through rigorous comparison and empirical calibration of models. The chapter discusses the motivation for and principles of those models. The main functionality of the entirely open-source software package is discussed.

Though the software is functional, available, and provides a rich set of models and assumptions to choose from, there are many ambitions for future development. What is more, there are plenty of social-influence models conceivable that do not strictly follow the seven principles outlined in Chapter 5 (e.g. models that use synchronous updating). For those communities that use agent-based modeling to understand

complex systems other than opinion systems, I hope the software and the underlying motivations can be of inspiration in the pursuit of a common body of knowledge.

A deeper understanding of the way social media shape our beliefs and opinions can yield important insights for developers of those systems, as well as to responsible authorities that wish to intervene. Translating scientific findings to practical applications is generally not social science's strongest suit. In malleable systems like the Internet, however, the possibilities to intervene or nudge are abundant and are typically not perceived as unethical social engineering. What is more, the formal models that produced predictions about the facets of social media used in the past, can also be exploited to provide predictions about interventions or techniques that stakeholders wish to introduce. A rich and calibrated but parsimonious model of online social media can, thus, be used as an ex ante *crash-test* model of online social media.

The study of opinion dynamics on the Internet is an exciting field for sociologists. The availability of fine-grained social behavioral data and popularity of computational social science techniques open up possibilities for rigorous testing of sociological theory and statistical modeling of social mechanisms (Flache et al., in press). Used in conjunction with rigorous theorizing, the cocktail of rich data and advanced technical abilities can be very fruitful (Keuschnigg et al., 2017; Molina & Garip, 2019). Sociology is a field that can be characterized by a low-degree of scientific agreement (Goldthorpe, 2021), through lack of a truly shared body of knowledge. An optimist might claim, however, that the Internet could be to sociologists what the telescope was to astronomers (Watts, 2011). Let's hope that we can exploit its possibilities such that, at least among sociologists, the Internet helps us to find common ground.