Data Science 5 (2022) 139–166
DOI 10.3233/DS-220054
IOS Press

The complex link between filter bubbles and opinion polarization

Marijn A. Keijzer [a,b] and Michael Mäs [b,*]

[a] *Institute for Advanced Study in Toulouse, University of Toulouse 1 Capitole, France*
ORCID: https://orcid.org/0000-0002-7585-062X
[b] *Karlsruhe Institute of Technology, Department of Sociology, Institute of Technology Futures, Germany*
ORCID: https://orcid.org/0000-0001-9416-3211

Editor: Victor de Boer (https://orcid.org/0000-0001-9079-039X)
Solicited reviews: Miriam Fernandez (https://orcid.org/0000-0001-5939-4321); Mariken van der Velden (https://orcid.org/0000-0003-0227-9183); Catherine Faron (https://orcid.org/0000-0001-5959-5561); Jacco van Ossenbruggen (https://orcid.org/0000-0002-7748-4715); Ronald Siebes (https://orcid.org/0000-0001-8772-7904)

Received 19 April 2020
Accepted 7 April 2021

Abstract. There is public and scholarly debate about the effects of personalized recommender systems implemented in online social networks, online markets, and search engines. Some have warned that personalization algorithms reduce the diversity of information diets which confirms users' previously held attitudes and beliefs. This, in turn, fosters the emergence opinion polarization. Critics of this personalization-polarization hypothesis argue that the effects of personalization on information diets are too weak to have meaningful effects. Here, we show that contributions to both sides of the debate fail to consider the complexity that arises when large numbers of interdependent individuals interact and exert influence on one another in algorithmically governed communication systems. Summarizing insights derived from formal models of social networks, we demonstrate that opinion dynamics can be critically influenced by mechanisms active on three levels of analysis: the individual, local, and global level. We show that theoretical and empirical research on these three levels is needed before one can determine whether personalization actually fosters polarization or not. We describe how the complexity approach can be used to anticipate and prevent undesired effects of communication technology on public debate and democratic decision-making.

Keywords: Personalization, recommender systems, opinion polarizations, filter bubbles, complexity, opinion dynamics, online social networks

1. Introduction

Political events such as the Brexit referendum, the election of Donald Trump, and the success of populist politicians in European and Latin-American democracies have sparked an intensive public and scholarly discussion about the effects of online communication technology on public debate and collective decision-making. One of the most prominent warnings is that personalization algorithms installed

[*] Corresponding author: Michael Mäs, Department of Sociology, Institute of Technology Futures, Karlsruhe Institute of Technology; Douglasstrasse 24, 76133 Karlsruhe, Germany. E-mail: michael.maes@kit.edu.

in online social networks, search engines, and online stores contribute to the formation of so-called "filter bubbles" [107]. These bubbles create echo chambers, isolating users from information that might challenge their views and exposing them to online content that is in line with their views. Experts, pundits, and scholars have warned that biased information diets reinforce users' political opinions and contribute to opinion polarization, a dynamic where a population falls apart into subgroups with increasingly opposing opinions. Public attention is substantial. Newspapers regularly cover the topic (e.g. [23,71]; leading politicians echo the warning [105,124]; and various initiatives have been undertaken to fight filter bubbles and polarization [13].

Here, we summarize the key arguments underlying the hypothesis that personalization algorithms contribute to opinion polarization and reflect on existing empirical research testing this hypothesis. Next, we identify open empirical questions that need to be answered before one can conclude whether or not personalization affects polarization. To this end, we collected relevant insights from the literature on opinion dynamics in social networks, demonstrating that the direction and the strength of the effect of personalization on polarization may critically depend on aspects that have not been sufficiently studied by empirical research. Finally, we draw conclusions about future theoretical and empirical research needed to evaluate the hypothesis that personalization fosters polarization and to design personalization technology that prevents undesired effects on opinion dynamics on the Internet. We propose that empirically calibrated, computational models of opinion dynamics have the potential to advance public and political debate about personalization, allowing researchers to rigorously quantify the impact of this technology on polarization. In addition, computational models can be used to conduct what we call computational "crash-tests" to anticipate deleterious effects of new communication technology before it is implemented in online services.

Our analysis is inspired by complexity research [8,86,106], an interdisciplinary field concerned with systems that consist of multiple micro-entities that do not act in isolation but exert influence on each other. In online social networks, for instance, millions of individual users interact with a large number of network contacts, communicating content that can influence each other's opinions. Social influence between users can generate chains of reaction such that even rare idiosyncratic events can have profound impact on the system as a whole, making online social networks an ideal-typical example of complex systems [83,91,137].

The complexity arising from communication in networks has been on social scientists' research agenda since the 1950s and has attracted attention from fields as diverse as physics, computer science, mathematics, economics, philosophy, sociology, and political science [47,49,93]. In this literature, formal models of social influence in networks were developed where network nodes exert social influence on the opinions of their network contacts. These models allow one to understand the rich and intricate collective dynamics that arise from social influence and to identify the conditions under which repeated social influence fosters the formation of opinion consensus, the fragmentation of the network into multiple clusters with competing opinions, or even opinion polarization. This literature reveals aspects of communication in networks that according to formal models have critical impact on whether personalization affects opinion polarization but that have not been considered in the debate so far.

Social-influence models differ in important ways from models of diffusion in networks, a class of models that has been used to model, for instance the spreading of rumors and (mis)information in online social networks [102,139]. In both model classes, populations are represented as a set of nodes integrated in a network of social relationships. These connections allow nodes to pass information around or to exert influence on each other. In diffusion models, it is assumed that nodes receive content from their neighbors and subsequently share it with their network neighbors. Models of social influence, in contrast, often do not attempt to model this spreading of information explicitly, but focus on opinion values to describe nodes. When two connected nodes interact, they exert social influence on each other, influencing the opinion value of their network neighbor. The central difference between the two classes of models is that diffusion models assume that for instance a piece of fake news can be passed on from one node to another. However, a node that, for instance, has never been exposed to the piece cannot pass its unawareness on to its network neighbors. In social-influence models, in contrast, influence can be bi-directional in that nodes can push and pull each other's opinions in all possible directions independent of their current state. We focus here on social-influence models rather than diffusion models, as social-influence models have a direct representation of opinions and can, therefore, be used to study the conditions of opinion polarization.

In a nutshell, we demonstrate that models' predictions about the effects of personalization on polarization hinge on assumptions about (i) individual behavior, (ii) individuals' local information environment and local communication structure, and (iii) global characteristics of the whole communication network. We conclude that these aspects need to be studied both theoretically and empirically before one should draw conclusions about the effects of personalization and we criticize that prominent contributions to the debate have so far failed to do so. While we echo the warning that personalization might have serious effects on societal processes, we demonstrate that experts, politicians, and scientists leap to conclusions when they propose that personalization is responsible for increased polarization. Unlike other recent contributions to the debate [4,17,146], however, we do not conclude that personalization is an innocent technology, but point to gaps in the empirical literature that need to be filled before one can draw conclusions. Accordingly, we call for more research on communication in online environments, pointing to the potential of approaches that combine rigorous theoretical modeling with the emerging fields of data science and computational social science. While these fields provide innovative sources of data and powerful methods of data analysis, we argue that their potential may not be exploited if it is not combined with rigorous theoretical modeling of the complex dynamics emerging in online communication systems, an approach that is increasingly popular [34,133]. We also discuss how carefully calibrated formal models can inform the development of personalization technology that prevents undesired effects on opinion dynamics on the web. The remainder is organized as follows. In the following section, we summarize the central theoretical, empirical, and political arguments underlying the scholarly and public debate about the effects of personalization on polarization. Next, we identify gaps in these debates, reporting largely overlooked findings from the literature on social influence in networks. In the concluding section, we sketch an agenda for future research and the design of personalization technology that prevents opinion polarization.

2. The debate about the personalization-polarization hypothesis

Personalization is ubiquitous on the web. Providers of web services seek to tailor their products to the needs and interests of individual users. Search engines, for instance, rank the results of users' search queries according to the interests of the individual user. When the authors of the present article google the term 'polarization', for example, websites discussing political polarization should be ranked higher than websites of manufacturers selling 'polarized' sunglasses, even though both sets of websites contain the search term. Likewise, online markets recommend products based on the purchases of other customers who bought similar products in the past and online social networks sort incoming messages according to the similarity between the user and the source of the message. Personalization has tremendously improved online companies' services, making it easier for users to navigate the immense and rapidly growing amount of online content. Personalization has also turned into a multibillion-dollar business area, increasing engagement on online platforms using this technology, and allowing advertisers to directly target potential customers. Personalization algorithms have been developed for various online services including online social networks, search engines, and online markets [11,19,32,69,80,108,121]. What is more, for each of these services there is a vast number of different technical approaches to personalization. What these approaches share, however, is that they seek to infer individual users' interests from information they provided, from their earlier behavior, and from the behavior of other individuals who share relevant attributes with the respective user. For instance, if a YouTube user regularly watches a certain car show on the platform, its algorithms will recommend other car-related content and content that other users who watched the same car show have selected in the past. As a consequence, users are exposed to content that is in one way or the other similar to the content they chose to consume earlier. This tendency to provide users to similar content and to limit their exposure to content that deviates from users' interests and opinions is central to the debate about undesired social effects of the technology. Accordingly, we also focus on this central aspect, abstracting from the large variety of technical implementations of personalization.

Despite the immense social and economic advances 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 [123]. 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" [107]. 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 [107]. In other words, personalization intensifies what sociologists labeled "homophily", the notion that individuals tend to interact with likeminded individuals [73,95,141].

Homophily is a strong force in human interaction, also in the absence of personalization [36,79,127, 146]. There is a rich empirical literature documenting that humans tend to interact with others who hold similar demographic attributes, have similar social status, and hold similar opinions [20,94,95].[1] In addition, it has been proposed that the Internet makes it especially easy to find and contact to like-minded individuals, which allows in particular users with extreme opinions to form online enclaves that would be very difficult to establish and maintain offline [119]. Sunstein argued that this high degree of homophily is potentially harmful, as it intensifies processes of opinion polarization, the development of antagonistic groups, where opinion differences between groups intensify and positions between the two extremes of an opinion spectrum are increasingly sparsely occupied [129]. Informed by social-psychological research [100,136], he argued that strong homophily intensifies users' opinions, as they are mainly exposed to online content containing persuasive information that reinforces their initial opinions. As opinions of users from the left end of the political spectrum grow more leftist and users identifying with rightist

[1] The origins of individuals' homophilic tendencies in social relationships are much debated. Generally, the motivations for associating with like-minded others fall into one of two classes: opportunity and preference [73,95,141]. The first class refers to all similarity between connected individuals that is a result from their proximity in geography, (shared) family relationships, societal or organizational positions, and isomorphisms caused by the intersections of structures. All such opportunities for interactions increase the likelihood that individuals meet and connect. The second class refers to individual preferences to create ties with similar and dissolve ties with dissimilar individuals.

political views also grow more extreme, opinion differences between the political camps increase and the opinion distribution polarizes.

Scholars and experts have noted that personalization technology is yet another source of homophily, a hypothesis that also found empirical support in research on Facebook [6]. Accordingly, personalization algorithms might have further intensified opinion polarization and may even be responsible for the growing opinion polarization observed in many western countries [31,88]. Here, we refer to this conjecture as the "personalization-polarization hypothesis".

The warning that personalization fosters polarization needs to be taken seriously, as opinion polarization has been argued to endanger societal cohesion [18,35,38,38,39,42,100] or cause cultural conflicts [61,62]. Opinion polarization might also pose challenges for political decision making in general [16] as it impedes political consensus formation also on otherwise non-controversial issues [61,62].

Political decision makers have echoed these warnings. Very prominently, Barack Obama warned in his farewell address that "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. [..] And increasingly, we become so secure in our bubbles that we start accepting only information, whether it is true or not, that fits our opinions, instead of basing our opinions on the evidence that is out there." [105] Frank-Walter Steinmeier, Germany's president, took this argument even further, linking personalization with adverse societal outcomes. In his 2018 Christmas message, he argued that "more and more people are sticking with their own kind, living in self-made bubbles where everyone always agrees one hundred percent [..]. What happens when societies drift apart, and when one side can barely talk to the other without it turning into an all-out argument, is all too evident in the world around us. We have seen burning barricades in Paris, deep political rifts in the United States and anxiety in the United Kingdom ahead of Brexit. Europe is being put to the test in Hungary, Italy and other places" [124].

What is more, there are already initiatives to break filter bubbles. Software developers, for instance, proposed novel personalization algorithms ranking higher content that challenges the opinions of the user [80,126,143]. Bozdag and Van den Hoven [13] distinguish two types technological solutions: those that make the user aware of their own bias, and those that show the users the opinion diversity for a given topic. The first type includes online tools that help users quantify and visualize the degree to which their news consumption is biased. Awareness of the composition of their information diet should then make users more open to other views. The second type is supposed to make users aware of the existing opinion diversity not visible from the limited perspective from inside their bubble. Some of these tools use questionnaires to plot opinion distributions or allow users to list and share pro and con arguments they consider relevant for given issues, others alert users when they visit a website that has been disputed on the web. There have also been non-technical initiatives to break filter bubbles by fostering offline discussion between individuals with opposing views. In multiple national and international events, mycountrytalks.org motivated thousands of participants to first indicate their political opinions in online surveys to be then electronically matched for face-to-face discussion with users holding maximally opposite opinions.

While the public debate about the link between personalization and polarization is mainly based on anecdotal evidence, outcomes of scientific research also echoed the warnings. First, modelers of social-influence processes in networks have developed formal models mimicking communication on the web, showing that the theoretical reasoning underlying the personalization-polarization hypothesis is logically valid [31,87,88]. These models assume that individuals adjust their political opinions as a result of communication with network contacts. When two agents hold similar opinions, their opinions are reinforced because they provide each other with new persuasive arguments supporting their views. In line with the informal reasoning underlying the personalization-polarization hypothesis, these models show that opinion polarization is more likely to emerge when agents are mainly communicating with likeminded individuals. Recent modeling work based on alternative assumptions about communication found similar dynamics [50,122].

Second, researchers have collected ample empirical evidence for the central assumptions underlying the formal models. There is a rich empirical literature documenting that humans have a strong tendency to interact with similar others [20,95] and to selectively consume media that supports their own political views [64,99,127]. In search of evidence for the existence of echo chambers on the web, these tendencies have been observed in online settings too [2,6,65,104,111,118,135]. Online social networking platforms further promote homophilic interactions through personalization algorithms [6]. There is also ample empirical evidence for the second critical model assumption: opinion reinforcement by communication with likeminded individuals [63,88,100,130,136]. Recently, empirical research in online contexts also supported this assumption [55].

There is, however, also considerable skepticism about the personalization-polarization hypothesis. In an interview with the New York Times, Mark Zuckerberg, the CEO of Facebook, responded that it is a "good-sounding theory, and I get why people repeat it, but it's not true" [84]. More importantly, however, there is also empirical evidence that challenges the personalization-polarization hypothesis. For instance, analyzing users' browser histories, researchers found that a large part of online news is still being consumed on news websites that do not filter content on the personal level, which should temper the effects of personalization of other web services [48,98]. Some scholars even argue that "social media usage [. . .] reduces political polarization" [9]. Barberá's analyses, for instance, suggest that most Twitter users are still exposed to diverse content and that exposure to diverse content fosters moderate rather than polarized opinions. Similar observations led Axel Bruns to conclude that even if personalization did foster the creation of filter bubbles, "the disconnect [..] is too mild to create any deleterious effect" [17]. While these empirical observations cast doubt on the personalization-polarization hypothesis, the conclusion that personalization has no effect is too strong. Not observing clear echo chambers on one personalized network or on one political issue does not imply that personalization is unproblematic. Polarization can still generate echo chambers and polarization in other contexts or on longer time frames [26,75].

However, also research on social-influence processes on the web challenges the personalization-polarization hypothesis. A large-scale field experiment found that exposure to content from the opposite political spectrum did reduce participants' negative feelings towards the other political camp but failed to affect their actual political opinions [76]. Another study with a similar design even documented that participants who were exposed to online content that strongly deviated from their own views grew more extreme, an opinion adjustment that is called negative influence [4,5]. Since filter bubbles shield individuals from online content that causes negative influence, researchers concluded that personalization algorithms may actually prevent opinion polarization [4,76]. In a third large field experiment where Facebook users were paid to deactivate their accounts for about one month prior to the 2018 mid-term elections in the US, it was found that political opinions of participants who deactivated their accounts deviated more from the average opinions of users with the same political party identification [3]. While this supports the hypothesis that Facebook use affects political opinions and that opinions within political camps grow more homogenous as a result, the study did not find that Facebook use leads to significantly more extreme views, as the personalization-polarization hypothesis would imply. In fact, participants with deactivated accounts did not develop less extreme opinions about President Trump or their preferred Congress candidate in their district than study participants who did not deactivate their account.

What is more, empirical research on the collective level has not yet painted a clear picture. On the one hand, research has documented that opinion distributions have polarized in many western countries since the web has become a dominating communication platform [1,12,35,39]. On the other hand, it is debated whether increased web use is actually responsible for rising polarization. One could argue that the more time users spend on the web the easier it is for them to escape their filter bubbles. A Facebook user, for instance, who does not only read the top-ranked messages of their news feed will also consume online content challenging her views. In fact, a prominent study found that opinions amongst young people – the demographic subgroup that spends most time on the web and in social networks – are the least polarized of all age cohorts [12].

In sum, the personalization-polarization hypothesis has received a lot of attention but research has so far not been able to provide conclusive evidence supporting or falsifying it. “So far, the literature is inconclusive on this issue, providing arguments and evidence supporting both sides of the debate”[144]. In the following section, we reflect on reasons why studying this hypothesis is challenging, pointing to aspects of online communication that are highly complex but hardly understood. Subsequently, we sketch a research agenda combining formal theoretical and empirical research.

3. The complexity perspective on the personalization-polarization hypothesis

Answering the question whether personalization technology fosters polarization is an ideal-typical research problem requiring a complexity perspective, as it is concerned with the two defining ingredients of complexity. First, a complex system consists by definition of multiple levels of analysis [8,86]. In the case of the personalization-polarization hypothesis, there is the level of the individual user who consumes, shares, adjusts, and generates content; and there is the collective level, the web. Both personalization and polarization are collective phenomena. For instance, an individual user cannot be polarized, but the distribution of users’ opinions can be. The second defining ingredient of a complex system are interdependencies between the entities on the microlevel. 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, the core argument underlying the personalization-polarization hypothesis proposes that personalization manipulates who is interacting with whom, changing the structure of interdependencies between users. This suggests that the analytical tools developed by complexity researchers have the potential to generate critical insight into personalization effects.

Research in various fields has demonstrated that complex systems can generate so-called “emergent” phenomena, collective patterns that are a consequence of the behavior of the individual-level entities but that are incongruent with the motives of these individual-level actors [8,21,86,106]. In the social sciences, for instance, Schelling and Sakoda demonstrated that cities can segregate into black and white districts even when all inhabitants are tolerant [58,114,117]. 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 the 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 be an emergent phenomenon, according to theories underlying the personalization-polarization hypothesis [31,88]. These theories do not assume that web users intend to live in a polarized world or that personalization 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. However, personalization algorithms increase the degree to which they are communicating with likeminded individuals who likely expose them to information that reinforces their opinions. Thus, polarization is an unintended consequence of communication in a personalized world.

While complexity science appears to contribute a critical perspective on the personalization-polarization hypothesis, the public and scholarly debate about personalization largely ignores the complexity of online communication. We argue here that two typical characteristics of complex systems are largely overlooked. First, a typical characteristic of many complex systems is that even small and seemingly innocent aspects of a system can have immense impact on system behavior. In fact, theoretical as well as empirical research demonstrates that complex social systems can be in a state where even rare and random events can alter collective outcomes [83,91]. 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 [134]. In the remainder of this section, we will reflect on insights from complexity research on opinion dynamics that illustrate that seemingly small differences in models' assumptions about individual, local and global aspects of communication can have important implications for the effects of personalization technology on opinion polarization. Accordingly, we criticize contributions to the debate on the personalization-polarization hypothesis that draw conclusions without a careful consideration of the complexity arising from communication in networks.

A second typical 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 [25,140]. 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 [51]. There is also a debate about the linearity of the effect of personalization. In their study of Facebook users, Bakshy et al. [6] 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. This may suggest that personalization is an innocent technology, but in a complex system this may not be true [75,87]. 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 that homophily has a nonlinear effect on systems tendencies towards polarization [87]. 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 changes

Table 1

Levels of analysis on the personalization-polarization hypothesis

Level of analysis	Definition	Important open questions
Individual	The individual level relates to aspects of communication that affect processes internal to the sender and receiver of content.	- Who expresses their views online and do individuals express their opinions online in the same way as in offline interaction? - What is being communicated online and do individuals communicate different content online than offline? - Is content communicated differently in an online than in an offline setting? - How do individuals adjust their opinions after communication online and are opinions changed in the same way as after offline communication?
Local	The local level relates to aspects of communication that affect who is when encountering content emitted by whom.	- To which degree is polarization intensified when there is one-to-many communication rather than one-to-one communication? - To which degree is polarization decreasing when forwarding content allows individual to exert direct influence on users they are not directly connected to?
Global	The global level relates to the structural characteristics of the communication network that affect individuals' information diet	- How does personalization change the structure of the communication network? - How do these changes affect the diffusion of online content in the network?

in the information diets of these users will indirectly affect the information diets of their friends and the friends of their friends.

In the following subsections, we review central insights from complexity research on opinion dynamics in networks and conclude that the existing research on the personalization-polarization hypothesis is not sufficient. In particular, we show that the complexity of opinion dynamics can arise on three levels of analysis: the individual, the local, and the global level. We show that empirical and theoretical research on these levels is needed to test the personalization-polarization hypothesis. Table 1 summarizes the three levels of analysis.

3.1. The 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 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 web personalization increases or decreases polarization [87,88]. In particular, reinforcement models [7,31,88] and rejection models [45,82,85,116] imply competing predictions about the conditions under which polarization emerges.

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 a likeminded individual [7,31,88]. One theory supporting this assumption is Persuasive-Argument Theory [88,100, 136], a psychological theory assuming that humans communicate arguments underlying their opinions. Individuals may hold a nuanced opinion themselves, but still 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 [31] and reinforcement-learning theory [7].

Reinforcement of opinions is a central assumption underlying the personalization-polarization hypothesis [31,87]. As personalization of online services increases the exposure to likeminded users and content that is in line with one's own views Internet users with opinions leaning towards the left end of the opinion spectrum 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, in line with the personalization-polarization hypothesis.

Rejection models, on the other hand, make alternative assumptions and imply markedly different macro-predictions [82,85,116]. Similar to the reinforcement models, rejection models also assume that individuals generally tend to grow more similar to likeminded individuals, an assumption that is usually implemented as averaging [49]. These models typically assume that users convey their position on an opinion continuum rather than communicating arguments as 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, an opinion shift that is labeled "rejection" [41,131].

Rejection models contradict the personalization-polarization hypothesis [87]. As personalization leads to fewer encounters between users who hold opposing views, rejection 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. Rejection models thus predict that an increase in web personalization will decrease opinion diversity over time.

Figure 1 illustrates the contradicting predictions of reinforcement and rejection models, showing the distribution of opinions over time in two scenarios; with and without personalization. The figures in the top row show typical simulation runs with a reinforcement model and were generated with a model assuming persuasive-argument communication [88]. In the bottom row of the figure, we show two typical runs with a rejection model [45].

In a nutshell, depending on whether one assumes rejection models or reinforcement models, one will come to the conclusion that personalization either decreases or increases polarization. Empirical research on social influence, however, is inconclusive in that it does not inform about which of the two models or which combination of the two models is empirically more accurate. On the one hand, social-psychological research suggests that online communication should reduce rejection between members of different demographic subgroups or different political camps. As group memberships are not observed in many forms of online communication, group boundaries that might cause rejection effects in offline settings could turn irrelevant online [112]. On the other hand, there is also research pointing in the opposite direction. In qualitative studies, it has been observed that online communication is often "unregulated by social context cues" [96]. In e-mails, users therefore use various tactics to allow receivers to better understand the meaning of their messages. Online social networks, however, restrict communication to relatively short messages, which makes communicating meaning and nuance more complicated. This,

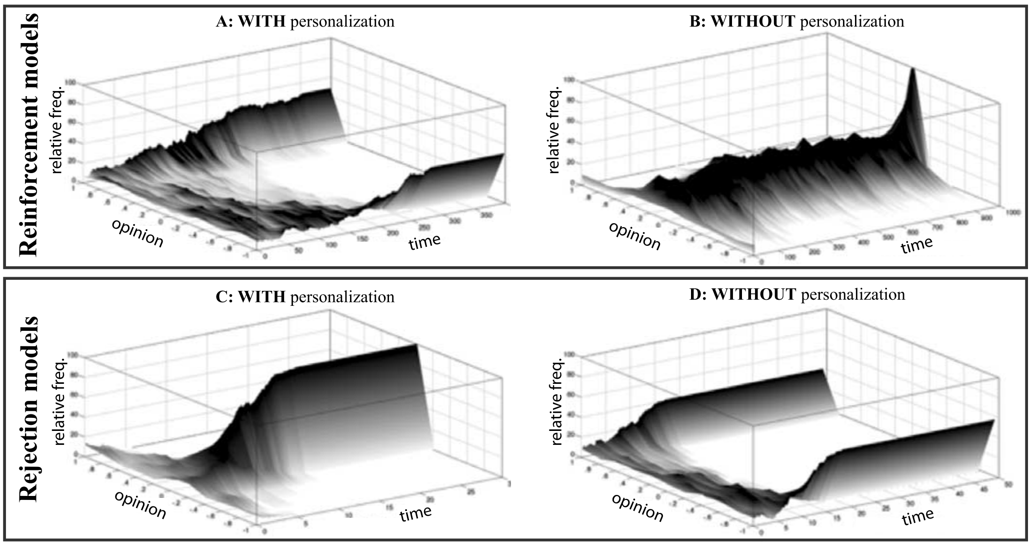

Fig. 1. Predictions of reinforcement and rejection models.

it has been observed, can cause confusion and rejection when receivers misinterpret messages [78,96]. Also experimental research on online social networks provided competing evidence for rejection [5,55]. Research on the persuasive-argument communication did provide ample of empirical support for reinforcement models, but this research has been conducted in offline settings [100,136]. In sum, it remains an open empirical question whether users of online social networks emit and receive persuasive arguments as described by reinforcement models, in particular because communication in these settings is often restricted to very short messages.

In addition to individual responses to political messaging, personalized online environments may also affect senders' communication decisions. Recently, researchers reported that the personalized design of online platforms contributes to political outrage, rather than actual opinion shifts within individuals [14]. Predominantly communicating with likeminded contacts, users may experience outrage when content challenging their views enters their filter bubble [78]. Furthermore, users may misrepresent their opinions to obtain credibility among likeminded others, communicating more extreme views than they actually hold [30,66]. Since, in addition, extreme, moral, and emotional content tends to spread more easily on online social media [15] and since computer mediated communication decreases empathy on the sender's side [30], political debate within filter bubbles can grow more heated than users' actual opinions would suggest.

In sum, alternative theories of the individual-level processes in communication network make opposing predictions about whether the personalization-polarization hypothesis is true or false. In reality, online communication may be best described by a hybrid of assumptions from rejection and reinforcement models, but without empirical information about which theory is true under what conditions, there are too many possible ways to combine assumptions of the competing theories into a single model. Furthermore, there are additonal individual-level factors that may have critical effects on whether polarization emerges or not. For instance, the described models abstract from possible heterogeneity between individuals. Some individuals may be more open to influence from online contacts than others, and some may exert stronger influence than others. To our knowledge, such heterogeneity has not been studied in the context of the two models, but for alternative opinion-dynamics models researchers documented critical effects [24,68,77].

In conclusion, without reliable empirical information about how individual users exert influence on each other's opinions in online contexts in conjunction with a rigorous theoretical analysis of the macro-level consequences of these micro-processes, it seems hardly possible to derive reliable predictions about the consequences of web personalization.

3.2. *The local level*

The local level of observation covers 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.

The difference between one-to-one and one-to-many communication 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 [43,67]. Intuitively, the difference between one-to-one and one-to-many communication may seem to be trivial, as a one-to-many communication-event is the same as a sequence of one-to-one communication events. Yet, modeling work with Axelrod's seminal model of cultural dissemination demonstrated that one-to-many communication fosters opinion fragmentation and social isolation in opinion systems [67]. This demonstrates that a simple change on the system's local-level can have serious effects on the whole system, without making more additional assumptions.

Figure 2 illustrates why one-to-many communication fosters the emergence of distinct network clusters with fragmented opinions according to Axelrod's model. Assume that there are four users who follow each other on Twitter. Each user has a stance on three issues illustrated by their color (black or white), shape (circle or box), and letter (A or B). In Panel a of Fig. 2, the number of lines connecting two users corresponds to the number of issues where users agree at the outset of the communication process. In a personalized system, this overlap will affect how likely an emitted piece of information will be consumed by the other user. The two users on the left, for instance, have zero opinion overlap and are, therefore, not exposed to each other's tweets. Axelrod's model takes this homophily into account, as it

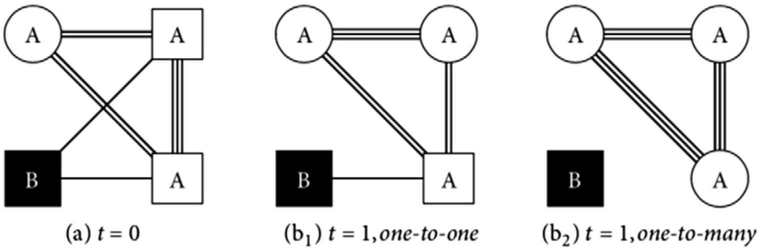

Fig. 2. Illustration of the intuition that one-to-many communication fosters isolation [67].

includes the assumption that the probability that an agent adopts a trait communicated by another agent is equal to the opinion overlap between the two agents.

Next, assume that the top-left user communicates her shape. Under the one-to-one communication regime, this trait may, for instance, be received by the top-right user, who adopts it and grows more similar to the sender as Panel b_1 of the figure illustrates. This instance of communication also changed the overlap between the receiver of the communication and the two remaining users, as a side effect. Nevertheless, the network remains connected and further communication between the two users on the right or the two users on the bottom can increase similarity between these users again.

Panel b_2 shows what happens under one-to-many communication when the top-left user emits her shape trait to all of her followers and all followers with a non-zero overlap adopt this shape. As the bottom-left user does not share a trait with the sender, the personalization algorithm will not expose the bottom-left user to the message. This form of communication has two effects, as Panel b_2 shows. First, a homogenous cluster formed because the communication did not only increase overlap between sender and each receiver. In addition, the overlap between the two receivers increased. Second, the bottom-left user ended up isolated, as they no longer shares any trait with the three others. Communicating her shape, the sender did not only increase the overlap between themselves and the two users on the right. In addition, the sender pulled these two users away from the bottom-left user. As a consequence, they will not interact with the isolated agent anymore.

The case shown in Fig. 2 is the simplest scenario where the difference between one-to-one and one-to-many communication can be illustrated. Modeling work, however, 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 when actors have many network ties [67]. One-to-many communication increases the chances that individual agents are isolated and that multiple internally homogenous but mutually distinct subgroups form.

The increased tendency to generate opposing clusters under the one-to-many regime is relevant for the personalization-polarization debate, because the difference between the two communication regimes is greater when homophily is increased. To demonstrate this, we conducted a simulation experiment with Axelrod's model of cultural dissemination, extending the analyses of [67]. In this model, all agents are characterized by a vector of F nominal features (beliefs, tastes or opinions) with Q traits. Agents adopt nodes in a network, being linked to other agents who they can influence and who can influence them. A sequence of discrete events is then initiated in which an agent is picked at random to be the sender of a message, and a randomly picked feature of this agent is shared with (one of) its neighbors. The receiver(s) of the message then decides whether to adopt or reject the trait, depending on the total number of traits that the sender and receiver have in common. This process of selection and influence is then repeated until all connected agents have either perfectly similar or dissimilar feature vectors.

We implemented one-to-one and one-to-many communication in Axelrod's model and varied the intensity of homophily in an experiment. Homophily was implemented with Eq. (1), which controls the probability P_{ij} that an agent i adopts the trait that one of its contacts j emitted. Variable overlap_{ij} describes the degree to which the two agents agree and is measured as the share of traits that the two agents share. When the homophily parameter h is set to a value of 1, then Eq. (1) implements exactly the model proposed by Axelrod where the probability that an agent adopts a communicated trait is equal to the overlap between the sender and the receiver of the trait. Furthermore, Eq. (1) does not affect the equilibria of Axelrod's model, since an overlap of zero always implies a zero probability that the receiving agent adopts the communicated trait. This holds for all values of parameter h. The function also implies that a perfect overlap translates into an adoption probability of one, but this event is actually ineffective,

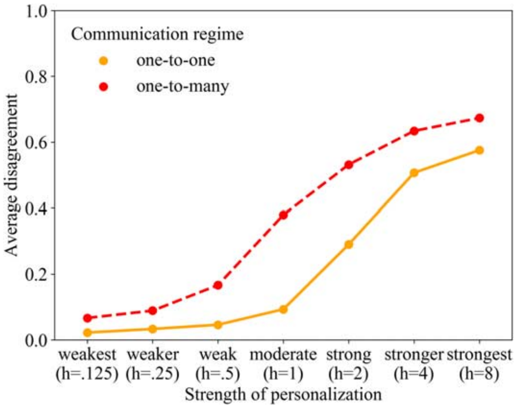

Fig. 3. Average disagreement when increasing homophily in the Axelrod model under two different communication regimes (averaged over 200 replications per homophily condition and regime).

as the two agents already share all traits. When h adopts a value close to zero the probability that the receiving agent adopts the communicated trait is close to 50% for all values other than zero and one, implying that similarity has only a very small impact on the probability that the receiving agent will adopt the trait. This implements that personalization is very weak. When h, in contrast, adopts values above one, then Eq. (1) turns into an S-shaped function, where overlaps below 0.5 translate into a low probability of trait adoption and overlaps above 0.5 most likely lead to adoption. This represents strong personalization.

$$P_{ij} = \begin{cases} (\frac{1}{2})^{1-h}\text{overlap}_{ij}^{h}, & \text{if overlap}_{ij}^{h} \leqslant 0.5 \\ 1-(\frac{1}{2})^{1-h}(1-\text{overlap}_{ij})^{h}, & \text{if overlap}_{ij}^{h} > 0.5 \end{cases} \tag{1}$$

To explore the relationship between personalization and average disagreement, we conducted a simulation experiment. We studied seven different values of personalization strength h and conducted 200 independent simulation runs per condition with one-to-one and with one-to-many communication. In all runs, we assumed a population of 100 agents interacting on a wrapped lattice network with Moore neighborhoods, which is a setup that is similar to the one studied by Axelrod. Every agent could exert social influence on their eight closest neighbors. We assumed that agents hold five cultural features ($F = 5$) that each contain one of 15 traits ($Q = 15$).

Figure 3 visualizes the findings from our simulation experiment, reporting how personalization affects the formation of fragmented subgroups in the simulated populations. To this end, we measured the average amount of disagreement between all pairs of connected agents in the network when the dynamics had reached a state of equilibrium. This macro-outcome measure adopts its minimal value of zero when all agents coordinated on the same set of traits. The maximal value of one represents the extremely unlikely event that every agent disagrees on all F dimensions with their eight network neighbors. Higher values, thus, indicate that the network consists of multiple subgroups with zero opinion overlap.

Figure 3 shows two central effects. First, it replicates the main findings from [67]: One-to-many communication generates stronger separation into subgroups than one-to-one communication. Second, the difference between the two regimes is particularly big when personalization is strong. Only when personalization is very strong, the difference between the regimes starts to level off again because of a ceiling effect. Here, the number of distinct clusters can hardly rise more and as a consequence, the difference between the two regimes declines. In the absence of the ceiling effect, the figure shows that the difference between the two communication regimes is bigger when personalization is stronger.

Personalization affects the observed difference in average disagreement within the two communication regimes for the following reason. One-to-many communication intensifies the emergence of different clusters, because an agent communicating a trait to multiple network neighbors pulls these neighbors away from other agents who are connected to them. Due to the homophily principle, the influence that these other agents can exert on their joint neighbors will decrease, which in turn increases the chances that differences grow even bigger in subsequent events. Under the one-to-one regime, in contrast, an agent can only pull one neighbor into its direction at a time. As a consequence, it is likely that the neighbor is subsequently influenced by another neighbor and, therefore, switches back to the traits adopted by the other agent. With one-to-many communication, such events are less likely, because the communicating agent influences all its neighbors at the same moment and, thus, pulls them all away from the other agent. Homophily is a central mechanism in this dynamic, as it is homophily that makes agents refuse influence from the agent they have been pulled away from. As a consequence, increased homophily makes the emergence of distinct clusters more likely.

In sum, the one-to-many communication effect illustrates that local-level aspects can impact opinion dynamics in social networks. It also shows that the effect of personalization technology on opinion polarization can depend on local-level aspects. We conclude that a thorough analysis of the personalization-polarization hypothesis needs to consider relevant local-level aspects. Unfortunately, researchers are only starting to explore local-level aspects, which suggests that it is too early to draw conclusions about the truth of the personalization-polarization hypothesis.

3.3. The global level

The global level refers to all structural elements of the communication network as a whole. For example, one characteristic of a network's structure that has been shown to have strong effects on opinion dynamics is network clustering, the degree to which connected nodes in a graph share other connections forming densely connected groups [22,40,43,109]. For illustration, Fig. 4 shows two networks with 120 nodes and different degrees of clustering [138]. To generate them, we arranged nodes in a circle and created undirected links between each agent and their five nearest neighbors to the right and the five nearest neighbors to the left. The resulting network has 600 edges and is shown in Panel a of Fig. 4 . It is characterized by very high clustering because this method of generating a network ensures a high number of triads, sets of three connected nodes. The transitivity coefficient – the number of realized triads over all possible triads – in this network amounts to .67. In contrast, the network shown in Panel b of Fig. 4 has a much lower degree of clustering. To generate it, we departed from the same circle network, but randomly rewired 35% of the links [92]. As a consequence, the number of links in the network and the number of links each agent has remained unaffected, but the transitivity coefficient dropped to .22.

In order to illustrate that network clustering affects opinion dynamics, we studied the dynamics generated by one of the most prominent social-influence models, the bounded-confidence model [33,56]. We chose this model, as it already has been used to derive hypotheses about the effects of personalization on opinion dynamics [50,122]. However, unlike earlier implementations of the bounded-confidence model, we assumed one-to-many communication, because this communication regime better mimics communication in online social networks [67].

To implement the bounded-confidence model, we assigned to every agent a random initial opinion drawn from a uniform distribution ranging from zero to one. Dynamics were then broken down into a sequence of discrete events. At every event, a randomly picked agent exerted influence on each of its network neighbors. That is, the program selected always one agent i who then communicated its opinion to all of its network neighbors j. When the opinion difference between the source of communication and the respective target was smaller than the so-called "bounded-confidence threshold" ϵ, then the opinion of the target agent was updated according to Eq. (2). Parameter μ represents how open agents are to social influence and was set to a value of .5.

$$o_{j,t} = o_{j,t} + \mu(o_{i,t} - o_{j,t}) \tag{2}$$

This model assumes that agents can exert only positive influence on each other, which is implemented as opinion averaging [47,49]. However, two agents can only exert influence on each other when two conditions are met. First, the two agents need to be directly connected by a network link. Second, agents' opinions must be sufficiently similar, a simple representation of personalization [50,122]. Small values of the bounded-confidence threshold ϵ imply that agents are only influenced by very similar network contacts, which represents that the influence from network neighbors with dissimilar views is suppressed by a personalization algorithm. Higher values represent that agents are also exposed to influence by neighbors who hold relatively different opinions. This represents that personalization algorithms have a weaker effect. We ran all simulations until a state of equilibrium was reached in that further communication would not have led to opinion adjustments because all connected agents either held identical opinions or held opinions that were too different to result in social influence. All simulations were implemented in defSim [72] and the code is made available as supplementary material to this article.

In Fig. 4, the four panels below the two network graphs show typical opinion dynamics in networks with high and low clustering and with strong or weak personalization. In each panel, we plot the share of agents' opinions. Initially, all four opinion distributions were uniform, but dynamics led in all four runs to the formation of subgroups. Comparison of the panels on the left-hand side with those on the right-hand side shows that opinion dynamics resulted in the formation of a higher number of subgroups when network clustering was high. That is, highly clustered networks tend to fall apart into a larger number of homogenous but mutually distinct subgroups. Agents belonging to a subgroup hold identical opinions but the opinion differences to their network neighbors who do not belong to the same subgroup are too high to allow for more influence. Note that the bounded-confidence model, unlike the models studied in Section 3.1, fails to generate increasing opinion differences between subgroups if no assumptions other than positive influence are added [57]. The model does, however, allow one to study the conditions of opinion fragmentation as the emergence of multiple subgroups.

Network clustering promotes opinion fragmentation because network clusters hamper the growth of subgroups. If, for instance, three agents are connected by two links and, thus, form a line network, then social influence will lead to opinion convergence if their opinions do not differ too much. A third link that closes the triad will in most cases not affect opinion dynamics in this small group. However, if the triad is not closed but a third link is added that connects any of the three agents in the line network to a fourth agent, they can deliberate with this new agent too and possibly make it join the subgroup. Network clustering thus hampers the growth of subgroups because every tie to an already included agent is a tie less to other agents who could join the group, ceteris paribus.

Figure 4 also suggests that personalization fosters the formation of opinion subgroups, according to the bounded-confidence model. This effect obtains because personalization decreases the number of neighbors that agents exert influence on. Those neighbors who do influence each other form homogenous groups, pulling agents who could have acted as bridges between groups towards the group's opinion

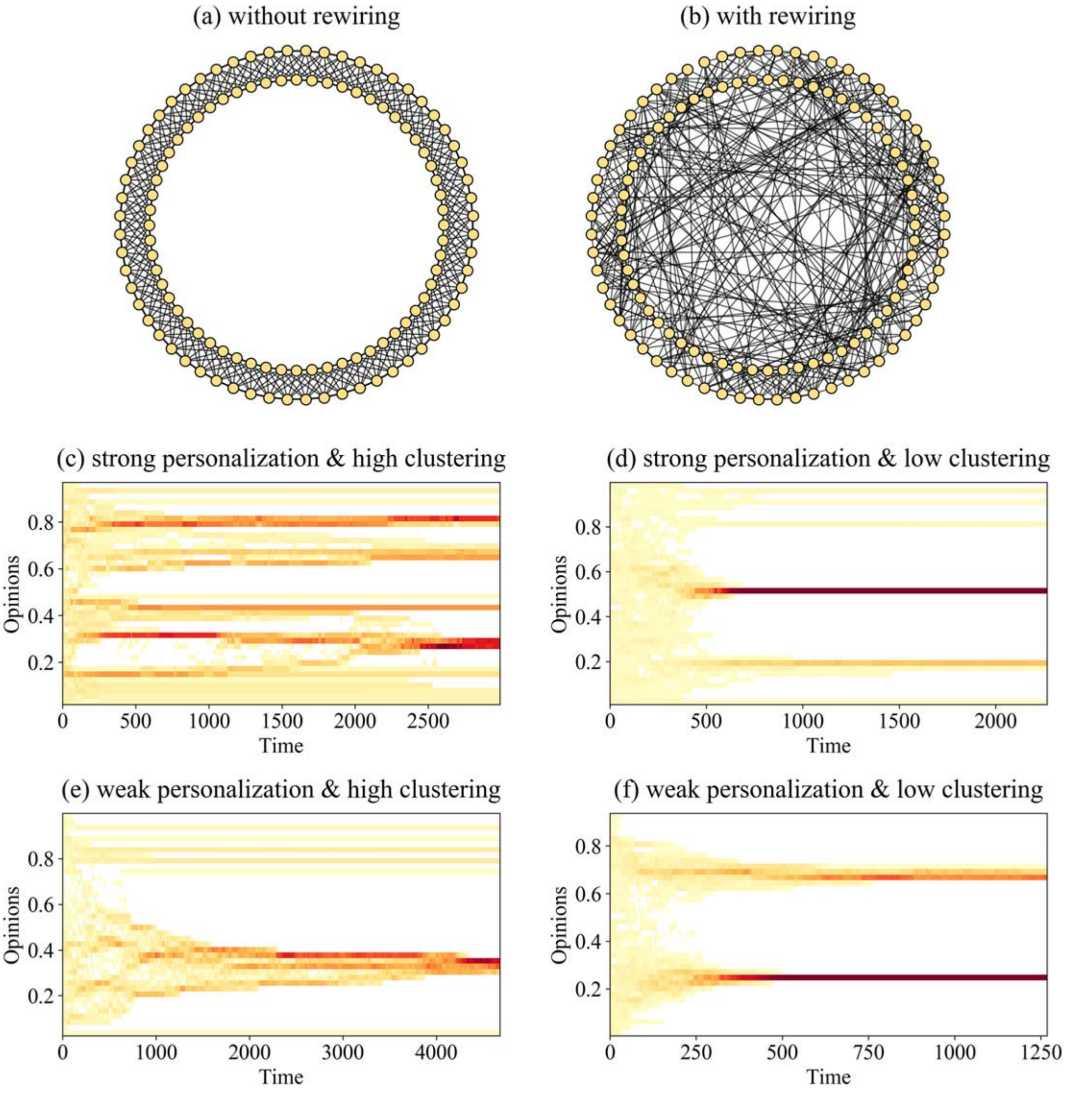

Fig. 4. Effect of network clustering and personalization on opinion fragmentation in typical simulation runs with the bounded-confidence model. Darker areas contain more agents.

average until they have grown too different from other groups to exert influence on them. When personalization is strong, agents exert influence on fewer neighbors. As a consequence, the network falls apart into a larger number of subgroups.

Panel a of Fig. 5 shows that network clustering intensifies the effects of personalization on the emergence of subgroups according to the bounded-confidence model. The figure is based on a simulation experiment in which we experimentally varied network clustering and the strength of personalization. We studied the same circle networks as shown in Fig. 4, including networks without rewiring (clustering = .67), networks with moderate clustering (105 rewiring iterations, average clustering = .38, sd = .02), and networks with strong clustering (210 rewiring iterations, average clustering = .22, sd = .02). In addition, we studied three levels of personalization, simulating dynamics under $\epsilon = .1$

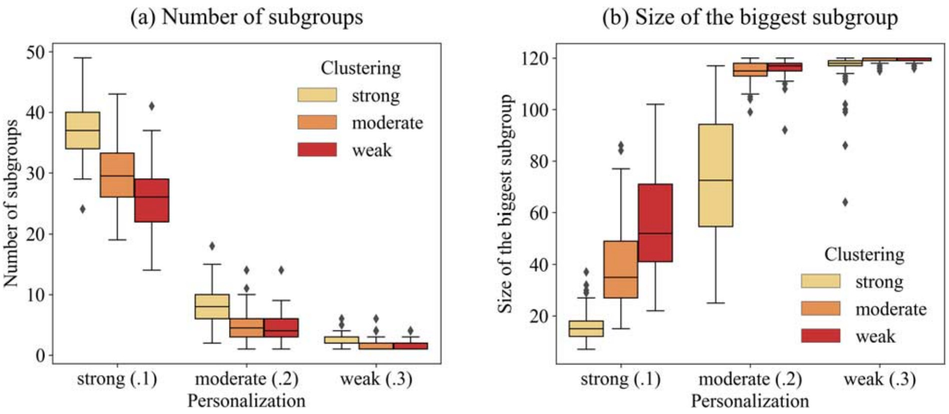

Fig. 5. The effect of network clustering and personalization on opinion fragmentation measured by the number of subgroups and by the size of the biggest subgroup.

(strongest personalization), $\epsilon = .2$, and $\epsilon = .3$ (weakest personalization). For each of the nine experimental treatments, we studied 100 independent simulations runs and always counted the number of distinct opinion subgroups in equilibrium.

Panel a of Fig. 5 shows for all three personalization treatments that more distinct subgroups formed when the network was characterized by higher clustering. Poisson regressions revealed that the effect of the number or rewiring iterations on the number of subgroups observed in equilibrium was statistically significant in each personalization treatment (minimal z-value was -6.38). In addition, the effect of network clustering was strongest in the treatment with strong personalization. In fact, in a Poisson regression, there is a strong and significant interaction effect between the number of rewiring iterations and personalization on the number of subgroups in equilibrium ($b = -3.49$, SE $= 1.20$, $p = .004$, full model in Appendix A).

Panel b of Fig. 5 shows results from the same simulation experiment but reports the size of the biggest subgroup in the network as the outcome variable, revealing another interesting difference between the moderate and the weak-personalization treatment. While panel a of Fig. 5 depicts that the number of subgroups formed was relatively similar, Panel b of Fig. 5 shows that under weak personalization there tends to be one very big subgroup and a number of smaller subgroups. Under moderate personalization, the average number of subgroups increases from 1.98 to 5.69, but the size of the biggest group tends to be considerably smaller than under the low personalization treatment, showing that multiple bigger groups of more similar size had formed.

The presented analysis of the effects of network clustering illustrates, in a nutshell, that the structure of the communication network can affect the degree to which personalization technology affects the outcomes of opinion-dynamics processes. Network clustering was taken here as an example and is just one of many potentially important global aspects. Other potentially relevant global aspects that have been shown to influence-model dynamics are demographic diversity [45,46,90], network segregation [40,53], the number of bridges connecting otherwise disconnected network clusters [90], and the existence of agents with many connections [22,128]. Considerable empirical and theoretical research is needed to understand whether and under what conditions global aspects affect how personalization technology affects opinion polarization. Without this research, however, it is not possible to evaluate whether or not online communication systems are a setting where personalization could affect polarization or whether global aspects prevent any desired or undesired effects.

Researchers are only starting to understand the structure of online communication networks, which makes it hard to evaluate the personalization-polarization-hypothesis. There are three main roadblocks. First, gathering data about online social networks is very challenging [113] and the information needed to quantitatively describe the structure and the evolution of communication networks is available only for very few networks [37,101,125]. Second, too little is known about the overlap between different networks. Critics of the personalization-polarization hypothesis do admit that online communication network can be segregated into clusters, but they also point to the fact that users tend to be active in various online and offline networks [17]. This, it is argued, creates crosscutting ties, allowing information and arguments to travel from one cluster to the other and decreasing opinion polarization. Whether this network multiplexity contributes to the diffusion of the same information or arguments over the composite graph is an empirical question that requires more research. For instance, users may use Twitter to communicate about political issues and focus on Facebook on entertainment and leisure. As a consequence, different clusters may be connected, but have only limited impact on opinion dynamics in the network overall. Third, personalization can also affect the structure of the interaction network. For instance, if personalization algorithms intensify the degree to which users are exposed to other users holding similar views, then they can also increase the degree to which the social network is clustered [54]. Assume, for illustration, that user A and user B are friends on Facebook and hold similar opinions. If Facebook's algorithms tend to propose creating links to users who hold similar views, then they may propose to both A and B to create a link to the same user C. While both links would result from the intention to create ties to likeminded users, an unintended consequence would be that A, B, and C form a triangle and, thus, contribute to network clustering. The analyses presented in this section have demonstrated that an increased degree of network clustering can further intensify processes of opinion polarization.

4. Conclusion

There is public and scholarly debate about the hypothesis that the personalization technology of online services contributes to the polarization of political opinions. On the one hand, experts, scholars, and political decision makers warn that personalization creates echo chambers where users' opinions are reinforced as they are mainly exposed to content that does not challenge their views. On the other hand, there are skeptical contributions arguing that the homophily generated by personalization may be too mild to generate these undesired effects.

We have shown that proponents of both positions in this debate leap to conclusions. We summarized insights from research on social-influence dynamics in networks to demonstrate that more empirical and theoretical research needs to be conducted before one can arrive at reliable predictions about the effects of personalization. Modeling work demonstrated that the opinion dynamics caused by personalization can critically dependent on assumptions about aspects on the system's individual, local, and global level. That is, if false assumptions about these aspects are made or if these aspects are not taken into account at all, then there is a good chance that one arrives at false conclusions about personalization effects. To date, there is insufficient empirical and theoretical research into these aspects, which makes it impossible to reliably conclude whether or not personalization breeds polarization. To be sure, we do echo the warning that personalization may have detrimental effects on public opinion formation and democratic decision making. These warnings need to be taken very seriously as democratic societies rely on an open public debate and a population's ability to find collective consensus. Although so far based on informal reasoning and anecdotal evidence, it would be dangerous to simply neglect the warnings.

The current state of the debate is worrisome for two reasons. First, the fact that there are theoretical arguments for and against undesired effects of personalization allows stakeholders to cherry-pick arguments that support their interests. In his 2017 community address, for instance, Mark Zuckerberg referred to the rejection assumption (see Section 3.1), arguing that "ideas, like showing people an article from the opposite perspective, actually deepen polarization by framing other perspectives as foreign" [145]. In fact, Zuckerberg might be correct but so far research has not demonstrated whether rejection models or reinforcement better describe online opinion dynamics. Second, there are already various attempts to break filter bubbles with the help of sophisticated technology and international events creating debate between individuals holding opposite views (see Section 2). The problem is that designing a successful intervention requires a proper understanding of the opinion dynamics on personalized communication networks. If, for instance, opinion dynamics are better described by rejection models than reinforcement models, then interventions trying to expose users more to content challenging their views might increase rather than decrease opinion polarization [4,76]. Interventions that are based on a false theory about how users exert influence on each other's opinions can have unintended consequences.

We advocate an approach that combines formal theoretical modeling with empirical research, because conclusions drawn from research applying just one of the two methods are likely problematic. On the one hand, a purely empirical approach to testing the personalization-polarization hypothesis can lead to false conclusions. Assume, for instance, that an empirical study quantified the degree of personalization-induced homophily in various settings and found no correlation with opinion polarization in these settings. This finding certainly challenges the personalization-polarization hypothesis. In complex systems, however, effects can take long to unfold and can then be very abrupt and strong. In Panel a of Fig. 1, for instance, polarization remained low for a long time, until it grew rapidly [87]. In addition, personalization algorithms are still being developed and refined. The observation that they have not contributed to opinion polarization so far, does not imply that further advances in personalization will also remain without negative effects [75]. This suggests that the empirical observation that personalization so far appears to be relatively mild and its effects on opinions modest [6,17], should not lead one to conclude that personalization will remain an innocent technology in the future. On the other hand, also a purely theoretical approach will fail to generate reliable predictions about personalization effects, even when analytical and computational tools are used to derive predictions. Our review of the opinion-dynamics literature provided several examples of modeling decisions that can have big impact on model predictions. As a consequence, models relying on assumptions that have not been backed up by rigorous empirical research in the context of online social networks may fail to make true predictions and, in addition, cannot be considered reliable tools for anticipating future opinion dynamics.

From the perspective of complexity research, the most promising approach to deriving predictions about the future effects of personalization on opinion polarization is to develop empirically calibrated models, an endeavor that requires empirical and theoretical research from various disciplines [47]. Theoretical research is needed to identify those assumptions that have a critical impact on model predictions, as these assumptions need to be put to the test by empirical research. Our summary of relevant complexity research has covered several aspects that require empirical investigation, but this list is not conclusive. To identify the most important mechanisms, modelers should invest more into comparing the predictions of alternative models [44,70,87–89]. Unfortunately, a recent review of the literature concluded that many contributors fail to highlight the similarities and differences between the model underlying their work and existing models [47], hampering the field's ability to accumulate knowledge and move forward. To improve, modelers should invest more into identifying these critical model assumptions, understanding why their model generates outcomes that other models do not. Furthermore, theoretical work should not only derive predictions about when a given model generates certain outcomes, but should find conditions under which different models provide different predictions. These insights will point empirical researchers to the empirical settings where competing models can be tested against each other, which in turn will help modelers develop validated models.

The emerging fields of data science and computational social science provide novel measurement tools, sources of data, and methods of analysis to study opinion dynamics in online environments [29,52,74] and to calibrate formal models. Informing research on the individual level, many online services offer application programming interfaces (APIs) that provide researchers with information about the content that users share online. In tandem with novel methods of sentiment analysis and topic modeling, this may allow testing theoretical model's assumptions about who is communicating what content to whom on the web [28,78]. In addition, controlled online experiments shed light on how users adjust their opinions as a result of online communication [5,27,55,120]. On the local level, models need to be enriched with empirical information on how often users are exposed to online content on different online platforms and when they decide to contribute to online debates. Finally, there have been advances in gathering, storing, and analyzing detailed information about global-level factors [37,101,125]. In particular, there is considerable research on the structure of online communication networks, which makes it possible to directly implement or regrow realistic communication networks in models of opinion dynamics [59,97,103]. When this empirical information is fed into a formal model of opinion dynamics, it will be possible to predict the collective dynamics arising from social influence and to study whether and to which degree personalization technology affects model dynamics.

In addition, the complexity perspective can inform the design of interventions combatting undesired effects of online media on public discourse and democratic decision making. First, the complexity perspective contests approaches that point to the individual level and advocate that educating individual users and enhancing their so-called digital literacy will prevent undesired effects of online media. Finkel et al., for instance, argued that one should be "encouraging them [social media users] to deliberate about the accuracy of claims on social media, which causes them to evaluate the substance of arguments and reduces their likelihood of sharing false or hyperpartisan content" [42]. They also proposed that users should contribute to the identification of false or hyperpartisan content and, thus, augment professional factchecking. To be sure, we do not doubt that it is useful to educate users about dynamics unfolding on online media and about how their own behavior can contribute to problematic outcomes. However, it is hard to imagine how individuals can be put in the position to reliably evaluate the truth value of online content that may have reached them via a long and usually invisible paths through the network [81]. What is more, fact-check labels on some but not all content in social media can backfire. According to the so-called implied-truth effect, unchecked content appears more truthful in the presence of labels on checked content [110]. Furthermore, common indicators of trustworthiness such as the number of times a piece of content has been shared or liked can be manipulated and can affected by social dynamics that have the potential to make them unreliable [10,115]. From the perspective of complexity research, it is naïve to assume that individuals are able to gain the literary needed to evaluate the effects of a complex system on them and vice versa.

Pointing to the individual level has also been used to immunize online social networks against criticism. When asked why he refused to "at least admit that Facebook played a central role or a leading role in facilitating the recruitment, planning, and execution of the attack on the Capitol", Zuckerberg pointed to "the people who spread that content, including the President but others as well, with repeated rhetoric over time saying that the election was rigged and encouraging people to organize. I think that those people bear the primary responsibility as well." [60] From a complexity perspective, this reasoning is problematic. On the one hand, complexity research allows one to demonstrate how individual behavior aggregates to collective outcomes and sometimes even the behavior of a single individual can have considerable impact on the dynamics of the overall system. On the other hand, we have shown that aspects on the local and the global level can have decisive impact too. In less abstract terms, local and global characteristics of online social systems can be designed in such a way that the behavior of the same individuals does not generate undesired effects. Thus, the designers of this communication technology can influence what dynamics emerge on their platforms. From this perspective, adjusting the design of online communication systems is an important contribution to combatting excessive opinion polarization.

So, how can complexity research contribute to a better design of online communication systems? The wide application of formal models in other fields where the complexity approach is used shows that empirically validated models of social influence dynamics are a potential game changer in the public and political debate about the effects of communication technology on opinion dynamics. To inform the climate-change debate, for instance, climate models are used to quantify the impact of specific economic sectors on climate change [142]. To this end, modelers compare temperature rise predicted by models considering the emissions of a given sector (e.g. air travel) with predictions of the same model assuming no emissions from this specific sector. Similarly, one could compare the levels of opinion polarization in online social networks predicted by social-influence models assuming strong and weak personalization. Such an analysis would provide a rigorous measure of the contribution of a specific web technology on the degree of polarization in a society. Obviously, companies will question the assumptions of the models when findings attribute undesired dynamics to their services. However, this would elevate the public and political debate from a discussion about untested theories and anecdotal empirical evidence to a scientific debate about the processes causing polarization. What is more, companies would be given a strong incentive to conduct the empirical research needed to test model assumptions and further improve models.

Validated social-influence models will also make it possible to conduct computational crash tests for online communication systems. In medicine, for instance, simulation software is used to predict which route a given substance will take in the human body after administration, which organ will break it down, and where it will affect the organism. Such virtual tests are conducted in early stages of drug development - before exposing humans or animals to an unexplored product. This approach is also used in varioius industries to optimize production and supply chains. NASA, for instance, is using so-called “digital twins” of spacecrafts to monitor and anticipate the behavior of their vehicles when exposed to the harsh environment of space [132]. Likewise, tech companies could use social-influence models to conduct virtual crash tests before they install new algorithms on their platforms, predicting whether and under what conditions they have undesired effects. In some industries, companies are legally required to conduct such tests. Like car manufacturers who are required to put their cars through a battery of tests before putting them on the road, tech companies could be forced to demonstrate with virtual crash tests that new algorithms have no undesired effects before implementing them in a system as complex as a communication network. Since digital communication technology has the potential to interfere with democratic processes on a global scale, implementing it without rigorous pre-testing is reckless.

Supplementary materials

Code for all models and analyses was implemented in Python using defSim, a software package designed for discrete-event social influence models [72]. All code is available at: doi.org/10.5281/zenodo.5948824.

Acknowledgements

We would like to thank the members of the Norms and Network Cluster at the department of sociology at the University of Groningen for their helpful feedback on earlier versions of the paper. Special thanks to Marion Hoffman for her suggestions on the implementation of homophily in the model. Marijn Keijzer acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

References

[1] A.I. Abramowitz and K.L. Saunders, Is polarization a myth?, *J Polit* **70**(2), 542–555. doi:10.1017/S0022381608080493.

[2] L.A. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election: Divided they blog, in: *3rd International Workshop on Link Discovery*, Chicago, Illinois, August 21–25, Association of Computing Machinery, 2005, pp. 36–43. doi:10.1145/1134271.1134277.

[3] H. Allcott, L. Braghieri, S. Eichmeyer and M. Gentzkow, The welfare effects of social media, *Am Econ Rev* **110**(3), 629–676. doi:10.1257/aer.20190658.

[4] C.A. Bail, *Breaking the social media prism: How to make our platforms less polarizing*, Princeton University Press, Princeton, New Jersey, 2021. 9780691203423

[5] C.A. Bail, L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.B.F. Hunzaker et al., Exposure to opposing views on social media can increase political polarization, *Proc Natl Acad Sci U S A* **115**(37), 9216–9221. doi:10.1073/pnas.1804840115.

[6] E. Bakshy, S. Messing and L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, *Science* **348**(6239), 1130–1132 (80-). doi:10.1126/science.aaa1160.

[7] S. Banisch and E. Olbrich, Opinion polarization by learning from social feedback, *J Math Sociol* **43**(2), 76–103. doi:10.1080/0022250X.2018.1517761.

[8] Y. Bar-Yam, *Dynamics of Complex Systems*, Westview Press, 2003, 848 p. ISBN 9780813341217.

[9] P. Barberá, How social media reduces mass political polarization. Evidence from Germany, Spain, and the U.S., 2015, http://pablobarbera.com/static/barbera_polarization_APSA.pdf.

[10] S. Bikhchandani, D. Hirshleifer and I. Welch, A theory of fads, fashion, custom, and cultural-change as informational cascades, *J Polit Econ* **100**(5), 992–1026. http://www.jstor.org/stable/2138632.

[11] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez, Recommender systems survey, *Knowledge-Based Syst* **46**, 109–132. doi:10.1016/j.knosys.2013.03.012.

[12] L. Boxell, M. Gentzkow and J.M. Shapiro, Greater Internet use is not associated with faster growth in political polarization among US demographic groups, *Proc Natl Acad Sci* **114**(40), 10612–10617. doi:10.1073/pnas.1706588114.

[13] E. Bozdag and J. van den Hoven, Breaking the filter bubble: Democracy and design, *Ethics Inf Technol* **17**(4), 249–265. doi:10.1007/s10676-015-9380-y.

[14] W.J. Brady, M.J. Crockett and J.J. Van Bavel, The MAD model of moral contagion: The role of motivation, attention and design in the spread of moralized content online, *Perspectives on Psychological Science*, 2020. doi:10.1177/1745691620917336.

[15] W.J. Brady, J.A. Wills, J.T. Jost, J.A. Tucker and J.J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks, *Proc Natl Acad Sci* **114**(28), 7313–7318. doi:10.1073/pnas.1618923114.

[16] P.R. Brewer, Polarisation in the USA: Climate change, party politics, and public opinion in the Obama era, *Eur Polit Sci* **11**(1), 7–17. doi:10.1057/eps.2011.10.

[17] A. Bruns, *Are Filter Bubbles Real?* John Wiley & Sons, 2019. ISBN 978-1-509-53644-3.

[18] B. Bryson, 'Anything but heavy metal': Symbolic exclusion and musical dislikes, *Am Sociol Rev* **61**(5), 884–899. doi:10.2307/2096459.

[19] R. Burke, Hybrid recommender systems: Survey and experiments, *User Model User-adapt Interact* **12**, 331–370. doi:10.1023/A:1021240730564.

[20] D. Byrne, *The Attraction Paradigm*, Academic Press, New York, London, 1971. ISBN 9780121486501.

[21] S. Camazine, J.L. Deneubourg, N. Franks, J. Sneyd, E. Bonabeau and G. Theraulaz, *Self-Organization in Biological Systems*, Princeton University Press, Princeton, New Jersey, 2001. ISBN 9780691116242.

[22] C. Castellano, S. Fortunato and V. Loreto, Statistical physics of social dynamics, *Rev Mod Phys* **81**(2), 591–646. doi:10.1103/RevModPhys.81.591.

[23] S. Chapin, Who's living in a 'bubble'? – The New York Times [Internet], 2018. https://www.nytimes.com/2018/12/11/magazine/whos-living-in-a-bubble.html.

[24] X. Chen, Z. Wu, H. Wang and W. Li, Impact of heterogeneity on opinion dynamics: Heterogeneous interaction model, *Complexity* **2017**, 1–10. doi:10.1155/2017/5802182.

[25] P. Cihon and T. Yasseri, A biased review of biases in Twitter studies on political collective action, *Front Phys* **4**, 34. doi:10.3389/fphy.2016.00034.

[26] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi and M. Starnini, The echo chamber effect on social media, *Proc Natl Acad Sci* **118**(9). doi:10.1073/pnas.2023301118.

[27] B. Clemm von Hohenberg, M. Mäs and B.S.R. Pradelski, Micro influence and macro dynamics of opinion formation [Internet], (SSRN), 2017. doi:10.2139/ssrn.2974413.

[28] R. Cohen and D. Ruths, Classifying political orientation on Twitter: It's not easy!, in: *Seventh International AAAI Conference on Weblogs and Social Media*, 2013. https://ojs.aaai.org/index.php/ICWSM/article/view/14434.

[29] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz et al., Manifesto of computational social science, *Eur Phys J Spec Top* **214**, 325–346. doi:10.1140/epjst/e2012-01697-8.

[30] M.J. Crockett, Moral outrage in the digital age, *Nat Hum Behav* **1**, 769–771. doi:10.1038/s41562-017-0213-3.

[31] P. Dandekar, A. Goel and D.T. Lee, Biased assimilation, homophily, and the dynamics of polarization, *Proc Natl Acad Sci* **110**(15), 5791–5796. doi:10.1073/pnas.1217220110.

[32] A.S. Das, M. Datar, A. Garg and S. Rajaram, Google news personalization, in: *Proceedings of the 16th International Conference on World Wide Web – WWW '07*, ACM Press, New York, New York, USA, 2007, p. 271. doi:10.1145/1242572.1242610.

[33] G. Deffuant, S. Huet and F. Amblard, An individual-based model of innovation diffusion mixing social value and individual benefit, *Am J Sociol* **110**(4), 1041–1069. doi:10.1086/430220.

[34] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli et al., The spreading of misinformation online, *Proc Natl Acad Sci* **113**(3), 554–559. doi:10.1073/pnas.1517441113.

[35] P. DiMaggio, J. Evans and B. Bryson, Have Americans' social attitudes become more polarized?, *Am J Sociol* **102**(3), 690–755. doi:10.1086/230995.

[36] G. Eady, J. Nagler, A. Guess, J. Zilinsky and J.A. Tucker, How many people live in political bubbles on social media? Evidence from linked survey and Twitter data, *SAGE Open* **9**(1). doi:10.1177/2158244019832705.

[37] H. Efstathiades, D. Antoniades, G. Pallis, M.D. Dikaiakos, Z. Szlavik and R.-J. Sips, Online social network evolution: Revisiting the Twitter graph, in: *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, 2016, pp. 626–635. doi:10.1109/BigData.2016.7840655.

[38] J.-M. Esteban and D. Ray, On the measurement of polarization, *Econom J Econom Soc* **62**(4), 819–851. doi:10.2307/2951734.

[39] J. Evans, Have Americans' attitudes become more polarized?-an update, *Soc Sci Q* **84**(1), 71–90. doi:10.1111/1540-6237.8401005.

[40] T. Feliciani, A. Flache and J. Tolsma, How, when and where can spatial segregation induce opinion polarization? Two competing models, *J Artif Soc Soc Simul* **20**(2), 6. doi:10.18564/jasss.3419.

[41] L. Festinger, *A Theory of Cognitive Dissonance*, Row, Petersen and Company, Evanston, White Plains, 1957. ISBN 9780804709118.

[42] E.J. Finkel, C.A. Bail, M. Cikara, P.H. Ditto, S. Iyengar, S. Klar, L. Mason, M.C. McGrath, B. Nyhan, D.G. Rand, L.J. Skitka, J.A. Tucker, J.J. Van Bavel, C.S. Wang and J.N. Druckman, Political sectarianism in America, *Science* **370**(6516), 533–536. doi:10.1126/science.abe1715.

[43] A. Flache and M.W. Macy, Local convergence and global diversity: From interpersonal to social influence, *J Conflict Resolut* **55**(6), 970–995. doi:10.1177/0022002711414371.

[44] A. Flache, M.W. Macy and K. Takács, What sustains cultural diversity and what undermines it? Axelrod and beyond, in: *Advancing Social Simulation: Proceedings of the First World Congress on Social Simulation*, S. Takahashi, ed., Springer, Kyoto, Japan, 2006, pp. 9–16. https://arxiv.org/abs/physics/0604201.

[45] A. Flache and M. Mäs, How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams, *Comput Math Organ Theory* **14**(1), 23–51. doi:10.1007/s10588-008-9019-1.

[46] A. Flache and M. Mäs, Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion, *Simul Model Pract Theory* **16**(2), 175–191. doi:10.1016/j.simpat.2007.11.020.

[47] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet et al., Models of social influence: Towards the next frontiers, *Jasss-the J Artif Soc Soc Simul* **20**(4). doi:10.18564/jasss.3521.

[48] S. Flaxman, S. Goel and J.M. Rao, Filter bubbles, echo chambers, and online news consumption, *Public Opin Q* **80**(S1), 298–320. doi:10.1093/poq/nfw006.

[49] N.E. Friedkin and E.C. Johnsen, *Social Influence Network Theory*, Cambridge University Press, New York, 2011. ISBN 9781107002463.

[50] D. Geschke, J. Lorenz and P. Holtz, The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers, *Br J Soc Psychol* **58**(1), 129–149. doi:10.1111/bjso.12286.

[51] S. Goel, A. Anderson, J. Hofman and D.J. Watts, The structural virality of online diffusion, *Manage Sci* **62**(1), 180–196. doi:10.1287/mnsc.2015.2158.

[52] S.A. Golder and M.W. Macy, Digital footprints: Opportunities and challenges for online social research, *Annu Rev Sociol* **40**(1), 129–152. doi:10.1146/annurev-soc-071913-043145.

[53] A. Grow and A. Flache, How attitude certainty tempers the effects of faultlines in demographically diverse teams, *Comput Math Organ Theory* **17**(2), 196–224. doi:10.1007/s10588-011-9087-5.

[54] T.U. Grund and J.A. Densley, Ethnic homophily and triad closure: Mapping internal gang structure using exponential random graph models, *J Contemp Crim Justice* **31**(3), 354–370. doi:10.1177/1043986214553377.

[55] D. Guilbeault, J. Becker and D. Centola, Social learning and partisan bias in the interpretation of climate trends, *Proc Natl Acad Sci*. doi:10.1073/pnas.1722664115.

[56] R. Hegselmann and U. Krause, Opinion dynamics driven by various ways of averaging, *Comput Econ* **25**, 381–405. doi:10.1007/s10614-005-6296-3.

[57] R. Hegselmann and U. Krause, Opinion dynamics and bounded confidence models, analysis, and simulation, *J Artif Soc Soc Simul* **5**(3). http://jasss.soc.surrey.ac.uk/5/3/2.html.

[58] R. Hegselmann, Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model, *J Artif Soc Soc Simul* **20**(3), 15. doi:10.18564/jasss.3511.

[59] P. Holme and B.J. Kim, Growing scale-free networks with tunable clustering, *Phys Rev E* **65**, 026107. doi:10.1103/PhysRevE.65.026107.

[60] House of Representatives, Disinformation nation: Social media's role in promoting extremism and misinformation, 2021 (Washington, D.C.) https://www.congress.gov/event/117th-congress/house-event/111407.

[61] J.D. Hunter, *Culture Wars: The Struggle To Control The Family, Art, Education, Law, And Politics In America*, Basic Books, New York, 1991. ISBN 0684867478.

[62] J.D. Hunter, Covering the culture war: Before the shooting begins, *Columbia J Rev* (July/August), 29–32. link.gale.com/apps/doc/A13192945/AONE.

[63] D.J. Isenberg, Group polarization: A critical review and meta-analysis, *J Pers Soc Psychol* **50**(6), 1141–1151. doi:10.1037/0022-3514.50.6.1141.

[64] S. Iyengar and K.S. Hahn, Red media, blue media: Evidence of ideological selectivity in media use, *J Commun* **59**(1), 19-U6. doi:10.1111/j.1460-2466.2008.01402.x.

[65] T.J. Johnson, S.L. Bichard and W.W. Zhang, Communication communities or "CyberGhettos?": A path analysis model examining factors that explain selective exposure to blogs, *J Comput Commun* **15**(1), 60–82. doi:10.1111/j.1083-6101.2009.01492.x.

[66] J.J. Jordan, M. Hoffman, P. Bloom and D.G. Rand, Third-party punishment as a costly signal of trustworthiness, *Nature* **530**(7591), 473–476. doi:10.1038/nature16981.

[67] M.A. Keijzer, M. Mäs and A. Flache, Communication in online social networks fosters cultural isolation, *Complexity*, 1–20. doi:10.1155/2018/9502872.

[68] G. Kou, Y. Zhao, Y. Peng and Y. Shi, Multi-level opinion dynamics under bounded confidence. Holme P, editor. *PLoS One* **7**(9), e43507. doi:10.1371/journal.pone.0043507.

[69] M. Kunavera and T. Požrl, Diversity in recommender systems – a survey, *Knowledge-Based Syst* **123**, 154–162. doi:10.1016/j.knosys.2017.02.009.

[70] T. Kurahashi-Nakamura, M. Mäs and J. Lorenz, Robust clustering in generalized bounded confidence models, *J Artif Soc Soc Simul* **19**(4), 7. doi:10.18564/jasss.3220.

[71] I. Lapowsky, How'd the Cohen hearing go? That depends on your filter bubble [Internet], *WIRED*, 2019. https://www.wired.com/story/cohen-hearing-filter-bubbles/.

[72] A.L. Laukemper, M.A. Keijzer and D.M. Bakker, defSim (v0.1). GitHub; 2019. https://github.com/defSim/defSim.

[73] P.F. Lazarsfeld and R.K. Merton, Friendship and social process: A substantive and methodological analysis, in: *Freedom and Control in Modern Society*, M. Berger, T. Abel and C.H. Page, eds, Van Nostrand, New York, Toronto, London, 1954, pp. 18–66. ISBN 9780374906085.

[74] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer et al., Computational social science, *Science* **323**(5915), 721–723 (80-). doi:10.1126/science.1167742.

[75] D.M.J. Lazer, The rise of the social algorithm, *Science* **348**(6239), 1090–1091 (80-). doi:10.1126/science.aab1422.

[76] R. Levy, Social media, news consumption, and polarization: Evidence from a field experiment, *Am Econ Rev* **111**(3), 831–870. doi:10.1257/aer.20191777.

[77] H. Liang, Y. Yang and X. Wang, Opinion dynamics in networks with heterogeneous confidence and influence, *Phys A Stat Mech its Appl* **392**(9), 2248–2256. doi:10.1016/j.physa.2013.01.008.

[78] T.Z. Lin and X. Tian, Audience design and context discrepancy: How online debates lead to opinion polarization, *Symb Interact* **42**(1), 70–97. doi:10.1002/symb.381.

[79] F. Loecherbach, J. Moeller, D. Trilling and W. van Atteveldt, The unified framework of media diversity: A systematic literature review, *Digit Journal* **8**(5), 605–642. doi:10.1080/21670811.2020.1764374.

[80] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang and T. Zhou, Recommender systems, *Phys Rep* **519**(1), 1–49. doi:10.1016/j.physrep.2012.02.006.

[81] B.A. Lyons, J.M. Montgomery, A.M. Guess, B. Nyhan and J. Reifler, Overconfidence in news judgments is associated with false news susceptibility, *Proc Natl Acad Sci* **118**(23). doi:10.1073/pnas.2019527118.

[82] M.W. Macy, J.A. Kitts, A. Flache and S. Benard, Polarization in dynamic networks: A Hopfield model of emergent structure, in: *Dyn Soc Netw Model Anal*, R. Breiger, K. Carley and P. Pattison, eds, January 2003, pp. 162–173. ISBN 0-309-08952-2.

[83] M.W. Macy and M. Tsvetkova, The signal importance of noise, *Sociol Methods Res* **44**(2), 306–328. doi:10.1177/0049124113508093.

[84] F. Manjoo, Can Facebook fix its own worst bug? *The New York Times Magazine*. https://www.nytimes.com/2017/04/25/magazine/can-facebook-fix-its-own-worst-bug.html.

[85] N.P. Mark, Culture and competition: Homophily and distancing explanations for cultural niches, *Am Sociol Rev* **68**(3), 319–345. doi:10.2307/1519727.

[86] M. Mäs Analytical sociology and complexity research, in: *Research Handbook on Analytical Sociology*, G. Manzo ed., 2021, pp. 100–118. ISBN 978-1-78990-685-1.

[87] M. Mäs and L. Bischofberger, Will the personalization of online social networks Foster opinion polarization?, *SSRN Electron J*. http://papers.ssrn.com/abstract=2553436.

[88] M. Mäs and A. Flache, Differentiation without distancing. Explaining opinion bi-polarization without assuming negative influence, *PLoS One* **8**(11). doi:10.1371/journal.pone.0074516.

[89] M. Mäs, A. Flache and J.A. Kitts, Cultural integration and differentiation in groups and organizations, in: *Perspectives on Culture and Agent-Based Simulations*, V. Dignum, F. Dignum, J. Ferber and T. Stratulat, eds, Springer, New York, 2013. doi:10.1007/978-3-319-01952-9_5.

[90] M. Mäs, A. Flache, K. Takács and K. Jehn, In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization, *Organ Sci* **24**(3), 716–736. doi:10.1287/orsc.1120.0767.

[91] M. Mäs and D. Helbing, Random deviations improve micro–macro predictions: An empirical test, *Sociol Methods Res* **49**(2), 387–417. doi:10.1177/0049124117729708.

[92] S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, *Science* **296**(5569), 910–913 (80-). doi:10.1126/science.1065103.

[93] W.A. Mason, F.R. Conrey and E.R. Smith, Situating social influence processes: Dynamic, multidirectional flows of influence within social networks, *Personal Soc Psychol Rev* **11**(3), 279–300. doi:10.1177/1088868307301032.

[94] M. McPherson and L. Smith-Lovin, Homophily in voluntary organizations – status distance and the composition of face-to-face groups, *Am Sociol Rev* **52**(3), 370–379. doi:10.2307/2095356.

[95] M. McPherson, L. Smith-Lovin and J.M. Cook, Birds of a feather: Homophily in social networks, *Annu Rev Sociol* **27**(1), 415–444. doi:10.1146/annurev.soc.27.1.415.

[96] D.A. Menchik and X. Tian, Putting social context into text: The semiotics of E-mail interaction, *Am J Sociol* **114**(2), 332–370. doi:10.1086/590650.

[97] A. Mislove, H.S. Koppula, K.P. Gummadi, P. Druschel and B. Bhattacharjee, Growth of the Flickr social network, in: *Proceedings of the First Workshop on Online Social Networks*, 2008, pp. 25–30. doi:10.1145/1397735.1397742.

[98] J. Möller, R.N. van de Velde L. Merten and C. Puschmann, Explaining online news engagement based on browsing behavior: Creatures of habit?, *Soc Sci Comput Rev*, 1–17. doi:10.1177/0894439319828012.

[99] J.S. Morris, The Fox News factor, *Harvard Int J Press*, **10**(3), 56–79. doi:10.1177/1081180X05279264.

[100] D.G. Myers, Polarizing effects of social interaction, in: *Group Decision Making*, H. Brandstätter, J.H. Davis and G. Stocker-Kreichgauer, eds, Academic Press, London, 1982, pp. 125–161. ISBN 0121258203.

[101] S.A. Myers, A. Sharma, P. Gupta and J. Lin, Information network or social network?: The structure of the twitter follow graph, in: *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web*, ACM Press, New York, New York, USA, 2014, pp. 493–498. doi:10.1145/2567948.2576939.

[102] A. Nematzadeh, E. Ferrara, A. Flammini and Y.-Y. Ahn, Optimal network modularity for information diffusion, *Phys Rev Lett* **113**(8), 088701. doi:10.1103/PhysRevLett.113.088701.

[103] M.E.J. Newman, The structure and function of complex networks, *Siam Rev* **45**(2), 167–256. doi:10.1137/S003614450342480.

[104] D. Nikolov, M. Lalmas, A. Flammini and M.F. Menczer, Quantifying biases in online information exposure, *J Assoc Inf Sci Technol* **70**(3), 218–229. doi:10.1002/asi.24121.

[105] B. Obama, Farewell adress [Internet], 2017. https://obamawhitehouse.archives.gov/farewell.

[106] S.E. Page, What sociologists should know about complexity, *Annu Rev Sociol* **41**(1), 21–41. doi:10.1146/annurev-soc-073014-112230.

[107] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press HC, New York, 2011. ISBN 9780141969923.

[108] D.H. Park, H.K. Kim, I.Y. Choi and J.K. Kim, A literature review and classification of recommender systems research, *Expert Syst Appl* **39**(11), 10059–10072. doi:10.1016/j.eswa.2012.02.038.

[109] N. Perra and L.E.C. Rocha, Modelling opinion dynamics in the age of algorithmic personalisation, *Sci Rep* **9**(1), 7261. doi:10.1038/s41598-019-43830-2.

[110] G. Pennycook, A. Bear, E.T. Collins and D.G. Rand, The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings, *Manage Sci* **66**(11), 4944–4957. doi:10.1287/mnsc.2019.3478.

[111] E. Peterson, S. Goel and S. Iyengar, Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign, *Polit Sci Res Methods*, 1–17. doi:10.1017/psrm.2019.55.

[112] T. Postmes, R. Spears, K. Sakhel and D. De Groot, Social influence in computer-mediated communication: The effects of anonymity on group behavior, *Personal Soc Psychol Bull* **27**(10), 1243–1254. doi:10.1177/01461672012710001.

[113] M.A. Russell and M. Klassen, *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and more*, O'Reilly Media, Inc., 2018. ISBN 9781491985045.

[114] J.M. Sakoda, The checkerboard model of social interaction, *J Math Sociol* **1**(1), 119–132. doi:10.1080/0022250X.1971.9989791.

[115] M.J. Salganik, P.S. Dodds and D.J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, *Science* **311**, 854–856. doi:10.1126/science.1121066.

[116] L. Salzarulo, A continuous opinion dynamics model based on the principle of meta-contrast, *J Artif Soc Soc Simul* **9**(1). https://www.jasss.org/9/1/13.html

[117] T.C. Schelling, Dynamic models of segregation, *J Math Sociol* **1**, 143–186. doi:10.1080/0022250X.1971.9989794.

[118] A.L. Schmidt, F. Zollo, A. Scala, C. Betsch and W. Quattrociocchi, Polarization of the vaccination debate on Facebook, *Vaccine* **36**(25), 3606–3612. doi:10.1016/j.vaccine.2018.05.040.

[119] P. Seargeant and C. Tagg, Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum, *Discourse, Context Media* **27**, 41–48. doi:10.1016/j.dcm.2018.03.005.

[120] C.R. Shalizi and A.C. Thomas, Homophily and contagion are generically confounded in observational social network studies, *Sociol Methods Res* **40**(2), 211–239. doi:10.1177/0049124111404820.

[121] Y. Shi, M. Larson and A. Hanjalic, Collaborative filtering beyond the user-item matrix, *ACM Comput Surv* **47**(1), 1–45. doi:10.1145/2556270.

[122] A. Sîrbu, D. Pedreschi, F. Giannotti and J. Kertész, Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model, Gargiulo F, editor. *PLoS One* **14**(3), e0213246. doi:10.1371/journal.pone.0213246.

[123] A. Smith and M. Anderson, Social media use in 2018, *Pew Res Cent* (March), 1–17. https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/.

[124] F.-W. Steinmeier, Christmas message [Internet], 2018. https://www.bundespraesident.de/SharedDocs/Reden/EN/Frank-Walter-Steinmeier/Reden/2018/12/181225-Christmas-message.html.

[125] M. Stella, E. Ferrara and M. De Domenico, Bots increase exposure to negative and inflammatory content in online social systems, *Proc Nat Acad Sci* **115**(49), 12435–12440. doi:10.1073/pnas.1803470115.

[126] J. Stray, Designing recommender systems to depolarize, *Arxiv*. http://arxiv.org/abs/2107.04953.

[127] N.J. Stroud, Media use and political predispositions: Revisiting the concept of selective exposure, *Polit Behav* **30**(3), 341–366. doi:10.1007/s11109-007-9050-9.

[128] K. Suchecki, V.M. Eguíluz and M. San Miguel, Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution, *Phys Rev E* **72**(3), 036132. doi:10.1103/PhysRevE.72.036132.

[129] C.R. Sunstein, *Republic.com 2.0*, Princeton University Press, Princeton, New Jersey, 2007. ISBN 9780691143286.

[130] C.R. Sunstein, The law of group polarization, *J Polit Philos* **10**(2), 175–195. doi:10.1111/1467-9760.00148.

[131] H. Tajfel and J.C. Turner, The social identity theory of intergroup behavior, in: *Psychology of Intergroup Relations*, S. Worchel and W.G. Austin, eds, Nelson-Hall Publishers, Chicago, 1986, pp. 7–24. doi:10.1177/053901847401300204.

[132] F. Tao and Q. Qi, Make more digital twins, *Nature* **573**, 490–491. doi:10.1038/d41586-019-02849-1.

[133] A. van de Rijt, Self-correcting dynamics in social influence processes, *Am J Sociol* **124**(5), 1468–1495. doi:10.1086/702899.

[134] A. van de Rijt, D. Siegel and M. Macy, Neighborhood chance and neighborhood change: A comment on bruch and mare, *Am J Sociol* **114**(4), 1166–1180. doi:10.1086/588795.

[135] A. Vespignani, Modelling dynamical processes in complex socio-technical systems, *Nat Phys* **8**(1), 32–39. doi:10.1038/nphys2160.

[136] A. Vinokur and E. Burnstein, Depolarization of attitudes in groups, *J Pers Soc Psychol* **36**(8), 872–885. doi:10.1037/0022-3514.36.8.872.

[137] M.M. Waldrop, News Feature: Modeling the power of polarization, *Proc Natl Acad Sci* **118**(37). doi:10.1073/pnas.2114484118.

[138] D.J. Watts and S. Strogatz, Collective dynamics of “small-world” networks, *Nature* **393**(6684), 440–442. doi:10.1038/30918.

[139] L. Weng, A. Flammini, A. Vespignani and F. Menczer, Competition among memes in a world with limited attention, *Sci Rep* **2**(335), 1–9. doi:10.1038/srep00335.

[140] L. Weng, F. Menczer and Y.-Y. Ahn, Virality prediction and community structure in social networks, *Sci Reports* **3** (2013), 2522. doi:10.1038/srep02522.

[141] A. Wimmer and K. Lewis, Beyond and below racial homophily: ERG models of a friendship network documented on Facebook, *Am J Sociol* **16**(2), 583–642. doi:10.1086/653658.

[142] P. Zhai, B. Zhou and Y. Chen, A review of climate change attribution studies, *J Meteorol Res* **32**(5), 671–692. doi:10.1007/s13351-018-8041-6.

[143] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling and Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, *Proc Natl Acad Sci* **107**(10), 4511–4515. doi:10.1073/pnas.1000488107.

[144] E. Zhuravskaya, M. Petrova and R. Enikolopov, Political effects of the internet and social media, *Annu Rev Econom* **12**(1), 415–438. doi:10.1146/annurev-economics-081919-050239.

[145] M. Zuckerberg, Building global community [Internet], 2017. https://www.facebook.com/notes/3707971095882612/

[146] F. Zuiderveen Borgesius, D. Trilling, J. Möller, B. Bod, C.H. de Vreese and N. Helberger, Should we worry about filter bubbles?, *Internet Policy Rev J Internet Regul* **5**(1). doi:10.14763/2016.1.401.