Computational and Mathematical Organization Theory (2021) 27:61–92
https://doi.org/10.1007/s10588-020-09315-8

MANUSCRIPT

Persuasion without polarization? Modelling persuasive argument communication in teams with strong faultlines

Thomas Feliciani[1,2] · **Andreas Flache**[1] · **Michael Mäs**[1]

Published online: 6 August 2020

Abstract
Strong demographic faultlines are a potential source of conflict in teams. To study conditions under which faultlines can result in between-group bi-polarization of opinions, a computational model of persuasive argument communication has been proposed. We identify two hitherto overlooked degrees of freedom in how researchers formalized the theory. First, are arguments agents communicate influencing each other's opinions explicitly or implicitly represented in the model? Second, does similarity between agents increase chances of interaction or the persuasiveness of others' arguments? Here we examine these degrees of freedom in order to assess their effect on the model's predictions. We find that both degrees of freedom matter: in a team with strong demographic faultline, the model predicts more between-group bi-polarization when (1) arguments are represented explicitly, and (2) when homophily is modelled such that the interaction between similar agents are more likely (instead of more persuasive).

Keywords Polarization · Work teams · Faultlines · Persuasion · Agent-based modeling · Social influence

1 Introduction

Demographic and cultural diversity is on the rise in many organizations. Labor forces diversify due to immigration; cultural minorities as well as women increasingly move upwards in occupational status; economic globalization gives rise to multi-national organizations; and pressures for more interdisciplinary work, especially in R&D and scientific research, increase disciplinary diversity in research teams (Meyer et al. 2014). Diversity can be an important asset for the performance

✉ Thomas Feliciani
thomas.feliciani@ucd.ie

[1] ICS/Department of Sociology, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands

[2] School of Sociology, University College Dublin, Belfield, Dublin 4, Ireland

of teams, especially when it comes to group tasks that require the combination of diverse sets of knowledge, skills, and experiences (Ellemers and Rink 2016).

Yet, diversity also has been characterized as a "double-edged sword" (Milliken and Martins 1996) "reducing social cohesion and increasing relationship conflict on one hand, and enhancing creativity and innovation on the other" (Carter and Phillips 2017, p. 1). Stereotypes and negative attitudes towards demographic or cultural outgroups have been found to fuel relational conflicts in a diverse team (Bowers et al. 2000; van Dijk et al. 2017; Van Knippenberg and Schippers 2007; Milliken and Martins 1996; Pelled 1996; Shemla et al. 2016; Stewart 2006; Webber and Donahue 2001; Williams and O'Reilly 1998). In addition, homophily, the well-documented tendency of people to preferentially link with similar others in informal networks (McPherson et al. 2001), can lead to between-group segregation of interpersonal relations in teams (Lau and Murnighan 1998; Reagans 2011), hampering the sharing and integration of the diverse pieces of knowledge and skills needed to master complex group tasks.

The notion that diversity may be a double-edged sword points to a complex process in which the effects on team-performance hinge on the interplay of multiple competing dynamics and contextual conditions. To unravel the complex social dynamics shaping consensus, cohesion or disagreement in organizations, researchers have employed the analytical power of computational modelling (Anzola et al. 2017; Harrison and Carroll 2002; Rouchier et al. 2014; Secchi and Gullekson 2016; Wang et al. 2017). A series of studies (Flache and Mäs 2008a, b; Fu and Zhang 2016; Grow and Flache 2011; Liu et al. 2015; Mäs et al. 2013; Mäs and Flache 2013; Pinasco et al. 2017; La Rocca et al. 2014) has focused in particular on formalizing and refining the theory of "demographic faultlines" (Lau and Murnighan 1998, 2005), highlighting the potential dangers of diversity for social cohesion that can arise from a `strong demographic Faultline'. "Group faultlines increase in strength as more attributes are highly correlated, reducing the number and increasing the homogeneity of resulting subgroups" (Lau and Murnighan 1998, p. 328). The core argument is that a strong faultline creates prominent subgroup distinctions, which may give rise to a 'group-split'. In the wake of Lau and Murnighan's seminal contribution, a range of empirical studies identified moderating conditions for this effect of faultline strength (Carter and Phillips 2017, p. 5; Leslie 2017). However, there is disagreement with regard to the theoretical assumptions explaining faultline effects, in that existing formal models are based on competing theoretical assumptions and generate opposing predictions about the conditions under which faultlines matter.

Our paper contributes to the literature by deepening the analysis of existing computational models of strong faultlines in teams (Feliciani et al. 2017; Fu and Zhang 2016; Mäs et al. 2013; Mäs and Flache 2013). In these models, authors translated the informal theory proposed by Lau and Murnighan into a computational model, drawing closely on their central psychological assumptions of (i) persuasive-argument communication and (ii) homophily. First, Lau and Murnighan assumed that individuals exert influence on each other's opinion by communicating persuasive arguments that are in favor of or opposed to a given position on an issue (Myers and Lamm 1976; Vinokur and Burnstein 1978). In an interdisciplinary team of social scientists and computational modelers studying diverse organizations, for example,

computer scientists may try to persuade their colleagues to develop a formal model by making arguments for the analytical precision of a computational theory, while social scientists may express counterarguments pointing to the danger of oversimplifying a theory through formalization. This argument communication can entail reinforcing influence, as actors with more similar opinions are likely to reinforce each other's prevailing opinion tendency. As a consequence, a bi-polarized opinion division arises aligned with the demographic (i.e. disciplinary) team division.

The persuasive-argument model of faultline dynamics has implications that are intriguing for researchers of diversity in teams. First, it identifies new conditions under which faultlines have the effects predicted by informal theorizing (Mäs et al. 2013). Second, the persuasive-argument model offers a formal theoretical alternative compared to an earlier approach (Flache and Mäs 2008a, b; Grow and Flache 2011). Instead of persuasive argument communication, this previous work modelled the effects of faultlines assuming negative or, "repulsive" influence—that is, the tendency to increase one's opinion difference from the opinion of the outgroup (Flache et al. 2017). This assumption has recently been challenged in experimental research (von Hohenberg et al. 2017; Mäs and Flache 2013; Takács et al. 2016), which has raised modelers' interest in alternative theoretical accounts of faultline dynamics in teams.

While existing modeling work adopted the assumptions of persuasive-argument communication and homophily from Lau and Murnighan's informal theory, these two assumptions can be interpreted and formally implemented in various ways. Accordingly, in this paper we ask the theoretical questions: do the central predictions of these models depend on the exact formal elaborations of the micro-processes of (1) argument communication and (2) homophily? And if so, how? In Sect. 2, we review the existing modeling literature and identify two dimensions of variation between existing modeling approaches. In Sect. 3, we formalize the competing modeling approaches and discuss possible implications for model dynamics. Section 4 presents results from computational experiments testing how the different model versions affect bi-polarization between subgroups. Possible implications for future research on diverse teams are discussed in the concluding section.

2 Existing models of persuasive argument communication under a strong faultline

2.1 Reinforcing influence and homophily: dynamics of group split

The mechanism generating bi-polarization in models of persuasive argument communication can more generally be described as "reinforcing influence". Under reinforcing influence, communicating individuals who hold a similar opinion reinforce each other's opinion and jointly become more extreme in their views. Different formalizations of micro-processes have been proposed that can entail reinforcing influence such as "biased assimilation" (Dandekar, Goel, and Lee 2013) or social learning from approval of opinions by relevant peers (Banisch 2010; Banisch and Olbrich 2017; Mäs and Flache 2013). In the study of faultline dynamics in teams, modelers

(Flache and Mäs 2008a, b; Fu and Zhang 2016; Grow and Flache 2011; Liu et al. 2015; Mäs et al. 2013; Mäs and Flache 2013; Pinasco et al. 2017; La Rocca et al. 2014) have closely followed persuasive-argument theory (Myers 1978; Vinokur and Burnstein 1978), according to which individuals base their opinion about an issue on the relevant arguments they possess about that issue, and influence each other when they communicate these arguments. On the one hand, argument communication between two interacting individuals reduces opinion disagreement between them, as the communication of arguments increases the similarity between the sets of arguments on which they build their opinion. On the other hand, when individuals with similar opinions communicate arguments, they likely expose each other to new arguments supporting their current opinion. In this case, reinforcing influence shifts the opinions of both actors towards more extreme views, a process that can aggregate to 'extreme consensus', the emergence of consensus on an extreme opinion in a group. Social psychological studies of extreme consensus in what has been called "group polarization" (Myers 1982) have provided consistent empirical evidence in support of the persuasive argument theory as explanation for this outcome (for a review see Isenberg 1986).

Extreme consensus is fundamentally different from a group split that divides a team with a strong faultline into subgroups with strong mutual disagreement. However, following the informal reasoning of Lau and Murnighan (1998, 2005), the computational models discussed above showed how persuasive argument theory can provide an explanation also for group splits, if reinforcing influence is accompanied by homophily, the tendency of individuals to interact more likely with more similar individuals (McPherson et al. 2001). Homophily fosters influence between individuals who are demographically similar or hold similar views; and it discourages influence between dissimilar individuals. Homophily fosters group splits especially when a strong faultine is accompanied by initial "congruency" (Mäs and Flache 2013), a tendency of actors with the same fixed attributes, like gender or ethnicity, to hold similar opinions even prior to persuasive communication (Phillips 2003; Phillips et al. 2004). In this case, homophily decreases chances that individuals are exposed to arguments that contradict their prevalent conviction. Instead, in a team with a strong faultline and initial congruency individuals are mainly exposed to arguments that support their views and reinforce their opinions. As this happens simultaneously on both ends of the opinion spectrum, a divide can grow in groups with a strong demographic faultline. Homophily makes interactions between members of the different emergent opinion groups increasingly unlikely, inducing even more reinforcing influence of like-minded others. To the extreme, this generates perfect opinion bi-polarization, an outcome in which a group falls apart into subgroups with maximal disagreement between and maximal agreement within subgroups (Duclos et al. 2004; Flache and Macy 2011; Flache and Mäs 2008b), eventually resulting in a group-split aligned with the demographic faultline.

Reinforcing influence and homophily have been shown to be conducive to bi-polarization in a team with a strong faultline. However, both the process of reinforcing influence via argument communication as well as the exact way how homophily moderates social influence allow for several "degrees of freedom" in their theoretical conceptualization and formal implementation. In the following we discuss these

degrees of freedom and their relation to the broader literature on modelling social influence processes.

2.2 Difference in the conceptualization of the argument-communication: explicit vs implicit variants

A controversial methodological debate in the agent-based modelling literature addresses the question how cognitively realistic agents should be. Some scholars call for "open[ing] the 'black box' of individual cognition" (Conte and Giardini 2016), arguing that modelers should identify and explicitly model the psychological mechanisms underlying social-influence processes. Others defend a more parsimonious, abstract definition of how to model behavioral micro-processes, arguing that cognitive realism can be progressively added to a minimalistic simple model until sufficient realism is met (Lindenberg 1992).

Reflecting these competing approaches, existing models of argument communication differ in the sophistication of the formal representation of arguments. On the one hand, sophisticated models represent arguments *explicitly*, assuming that an actor's opinion is a function of the arguments she considers relevant and that arguments are being shared with communication partners. Actors who are exposed to an argument they did not consider before adjust their opinions according to the argument. On the other hand, there are more parsimonious models that model arguments *implicitly*. That is, these models do not represent arguments but make assumptions about how communication would have adjusted their opinions had they communicated arguments.

2.3 Difference in the conceptualization of homophily: likelihood or effectiveness of interaction

An important conception of homophily in the sociological literature is that people more likely interact and communicate with similar others (Lazarsfeld and Merton 1954; McPherson et al. 2001; Wimmer and Lewis 2010). This form of homophily may be caused by the preference that "likes attract" (Byrne 1971) and thus reflect the outcomes of a choice people make among available interaction partners, but it can also result from structural patterns of social interaction that systematically sort similar people into similar "foci" (Feld 1982) where they meet and interact, like schools, neighborhoods, or workplaces. In both cases, the reason that more similar people influence each other more is that they interact more frequently than less similar people do. The view that similarity increases likelihood of interaction resonates in how homophily is conceptualized in large number of computational models of social influence (Axelrod 1997; Baldassarri and Bearman 2007; Chen et al. 2013; Dandekar et al. 2013; Mark 2003). In all of these models, actors select an interaction partner from a set of available agents such that more similar agents are selected with a higher probability.

An alternative conceptualization of homophily builds on the notion that similar people influence each other more effectively, because individuals are more open to arguments communicated by similar others. People can make more sense of input from sources with whom they have more in common. Furthermore, people might trust similar others more that dissimilar people would (Mark 1998). This view on homophily is likewise reflected in formal models of social influence where, for example, the similarity between two agents is expressed in terms of a weight that scales how much the opinion of a source of influence is taken into account for opinion changes of the target of influence (Deffuant et al. 2000; Duggins 2017; Flache and Macy 2011; Flache and Mäs 2008b; Hegselmann and Krause 2002; Kitts 2006; Kurahashi-Nakamura et al. 2016; Mäs et al. 2010).

While both competing conceptualizations of homophily have been adopted in a variety of formal models of social influence processes, it remains unclear how exactly variation between them affect the outcomes of formal models of faultline dynamics.

3 The model

In this section we introduce a generic formal and computational model which embeds different combinations of the degrees of freedom discussed above. Specifically, two of these combinations represent earlier formal models of faultline dynamics that will be systematically compared for the first time in the present study. One combination is to model reinforcing influence with explicit communication of arguments, and implement homophily via the likelihood of selection of interaction partners, hereafter referred to as X–S model. The X–S model has been adopted in earlier work for example by (Mäs et al. 2013; Mäs and Bischofberger 2015; Mäs and Flache 2013). Another combination, differing in both dimensions from the X–S model, is that reinforcing influence builds on implicitly represented arguments and homophily affects the effectiveness of arguments communicated, but not the choice of interaction partners, hereafter called the I–E model (adopted by Feliciani et al. (2017) in earlier work). Finally, we introduce with our framework a new model of persuasive argument influence differing from both the X–S and the I–E model in one of the two degrees of freedom. This is the model in which the communication of arguments is modeled implicitly like in I–E, but homophily is implemented via selection of interaction partners like in X–S, called I–S[1].

Box 1 provides the pseudo-code of the ABM; the code itself can be found in a public GitHub repository (https://github.com/thomasfeliciani/persuasive_argum

[1] It is worth noting that a fourth model version is theoretically possible, with explicit representation of arguments (as in X–S) and where homophily affects the effectiveness of the arguments communicated (as in I–E). However, this fourth possible model would require additional assumptions (e.g. on the weighting of arguments) that would set it apart from the other three models (X–S, I–E and I–S). This makes the fourth model version unsuitable for our study.

```
Let agentset
Fill agentset with N agents with attributes group, opinion
Create interaction network
Let t = 1
While t ≤ 10^4:
        For each agent i in agentset: (random order)
                Calculate similarity vector
                Let o = opinion of i

                If model_version is "X-S" then:
                        Let j = randomly selected interaction partner (probability weight = similarity)

                        Let i_memory_vector = memory vector of i
                        Let j_memory_vector = memory vector of j
                        Remove the least recent argument from i_memory_vector
                        Let x = randomly drawn argument from j_memory_vector (uniform probability)
                        Make x the new most recent argument in i_memory_vector
                        Update opinion of i = compute(i_memory_vector)

                If model_version is "I-E" then:
                        Let j = randomly selected interaction partner (uniform probability)

                        Let forgotten_argument_type = pick (pro, con) via binomial trial
                        Let communicated_argument_type = pick (pro, con) via binomial trial
                        Let a = compute (forgotten_argument_type, communicated_ar gument_type)
                        Update opinion of i = compute (o, a, similarity between i and j)

                If model_version is "I-S" then:
                        Let j = randomly selected interaction partner (probability weight = similarity)

                        Let forgotten_argument_type = pick (pro, con) via binomial trial
                        Let communicated_argument_type = pick (pro, con) via binomial trial
                        Let a = compute (forgotten_argument_type, communicated_argument_type)
                        Update opinion of i = compute (o, a)

        If system has converged then:
                Exit while loop
        Else:
                Set t = t + 1
Calculate outcome measures
Terminate simulation
```

Box 1 ABM pseudo-code

ent_model_NetLogo). Table 1 (in Sect. 4.1) contains an overview of the variables presented.

3.1 General modeling framework

We assume a population of N agents in all three models. Typically, we assume $N=10$, approximating the size of teams in many organizations, but we will also explore effects of a bigger population size. Agents have two main attributes: their group identity, and their opinion on an issue. A maximally strong faultline is implemented in terms of a dichotomous group identity $g_i \in \{-1,+1\}$ that is a fixed attribute randomly assigned to every agent i at the outset[2]. We assume that the two groups have equal size: exactly half of the population belongs to group -1, and the other half to group $+1$.

The opinion of an agent i at time point t is denoted $o_{i,t}$, and is a continuous variable in the range $[-1,+1]$. Opinions are an aggregation of the positive ('pro') and negative ('con') arguments an agent considers relevant. More precisely, pro arguments are in favor of $o_{i,t}=+1$ and con argument support the opinion $o_{i,t}=-1$. At

[2] This is the setup that was adopted in the I–E, and is equivalent to the persuasive argument model as in the E–S for one demographic dimension, and maximal faultline strength.

the outset of the simulation, agents' opinion is initialized as follows. Agents hold S arguments (S is a model parameter, set to 4 by default). When arguments are represented implicitly, the initial arguments are only used to induce an initial opinion, otherwise they are explicitly assigned to agents' memories. The model further contains a parameter w for the degree of congruency between the demographic attribute and the opinion. Specifically, for each of the S memory "slots", agents with $g_i = 1$ will receive a positive argument with probability w[3], and a negative argument otherwise. Conversely, the other group ($g_i = -1$) will receive a negative argument with probability w, and a positive argument otherwise[4]. Finally, opinions are calculated as the number of positive arguments over S, scaled to range from -1 and $+1$. Formally, if we define $P_{i,t}$ as the number of pro arguments held by agent i, then:

$$o_{i,t} = 2 \cdot \frac{P_{i,t}}{S} - 1 \tag{1}$$

This implies that parameter S also defines how much an argument can impact the opinion of an agent. For $S = 4$, for example, an agent can only know 4 arguments, and every argument accounts for one quarter of the agent's opinion.

A congruency $w = 0.5$ yields an opinion distribution without any correlation between group and opinion; for $0.5 < w \leq 1$, higher values of w produce stronger correlation between group and initial opinion.

After the initialization, time elapses in discrete steps. For each time step t, all agents are selected for initiating an interaction. The sequence in which they are selected is randomly shuffled at the beginning of each time step. When agent i is selected for this, we first select an interaction partner j, and then simulate the interaction between j and i.

How interaction partners are selected (i.e. the implementation of homophily), and how interactions take place (the argument-communication mechanics) are the two main differences between the X–S and the I–E model versions, which we describe in the following two sections.

3.2 Difference in the formalization of the argument-communication

3.2.1 Explicit argument-communication

Under explicit argument communication, we assume that the pool of available arguments consists of 10 pro and 10 con arguments. To mirror the limits in human capability to retain and process information (Cowan 2001; Miller 1956), the X–S assumes that, at any point in time, agents can only memorize a subset of S arguments. Furthermore, the memory vector allows to implement the recency (or salience) of an argument: the first argument of the vector is the most recent—the last,

[3] Since all arguments of the same sign (i.e. "pro" or "con") are equivalent, we do not need to sample from a population of available arguments; instead, we rely on a Bernoulli trial to determine whether each given argument in i's memory S will be "pro" or "con".

[4] This is the opinion initialization method used in the E–S, and for $w = 0.5$ is equivalent to the initialization method in the I–E.

the least. The underlying assumption is that recently acquired arguments linger in agents' memory for longer[5] (Mäs et al. 2013).

During interaction, agent j communicates one of the arguments she considers salient to i. An argument from j's memory is picked at random and then becomes the first and thus most recent argument in i's memory. Due to the limited memory size S, agent i also drops the most dated argument[6]. If the argument communicated by j is already present in i's memory (read: if i already considers it), then it shifts from the current location in i's memory to the first position. A known argument, if encountered again, thus becomes more recent and remains salient longer. The interaction event is then terminated by updating the opinion of agent i as defined in Eq. 1.

3.2.2 Implicit argument-communication

Under implicit argument-communication, the interaction is simplified by mimicking only the opinion change induced by argument communication, without actually representing the arguments. To achieve this, the probabilities are calculated that an explicit argument communication would result in each of the possible outcomes of shifting the opinion upwards, downwards or not at all on the opinion scale. First, we determine the likelihood that agent j would communicate a pro argument according to Eq. 2. With one minus this probability, j would communicate a con argument to her communication partner i.

$$\text{Probability of } j \text{ communicating a pro argument} = \frac{1}{2} \cdot \left(o_{j,t} + 1\right) \tag{2}$$

Next, it is determined how likely i drops a pro or con argument at the end of the interaction. Like in Eq. 2, the probability that i drops a pro argument is:

$$\text{Probability of } i \text{ dropping a pro argument} = \frac{1}{2} \cdot \left(o_{i,t} + 1\right) \tag{3}$$

Based on these two probabilities, the algorithm of implicit argument-communication conducts a random experiment that selects a combination of the two events "j communicates a pro or a con argument" and "i drops a pro or con argument" as outcome of the interaction. Next, we compute the resulting shift of i's opinion as follows. The magnitude of the opinion adjustment $a_{i,t}$ is a function of S, as S determines how much a new argument can affect an agent's opinion:

[5] The assumption of argument recency and its effects on the model mechanics is inherited from the original model, where the authors discuss its theoretical motivation from psychology literature. The authors also show that the model results are overall robust to an alternative implementation without argument recency, where the argument to be forgotten is picked at random instead of based on recency (Mäs et al. 2013, Online Appendix).

[6] Selection and forgetting of arguments can be implemented in different ways in this process. Mäs et al (2013) found that bi-polarization dynamics are robust across different specifications.

$$a_{i,t} = \begin{cases} 2/S, & \text{if j picks a pro and i drops a con argument} \\ -2/S, & \text{if j picks a con and i drops a pro argument} \\ 0, & \text{if j pics and i drops the same kind of argument} \end{cases} \quad (4)$$

3.2.3 Implications

Explicit and implicit versions of the argument-communication seem equally plausible modeling approaches to model the underlying theory, and we want to test whether or for which conditions they yield consistent results. We know of a crucial difference between the two versions: the implicit argument-communication cannot reproduce all of the opinion outcomes generated by the explicit communication version. Specifically, in the model with explicit argument-communication, a population of agents might develop consensus over a moderate opinion, where all agents have the same number of pro and con arguments. If all agents have the exact same set of arguments, there are no arguments that can be communicated that would change an agent's opinion: this means that the agents would be locked in consensus.

With implicit argument-communication, in contrast, a consensus on a moderate opinion is not an equilibrium. Since arguments are not explicitly represented, two agents can always influence each other as long as they have not agreed on the same extreme opinion. That is, Eqs. 2 and 3 always yield a positive probability for an outcome in which the argument communicated by j has the opposite sign than the argument dropped by i, unless $o_i = o_j = \pm 1$—i.e. when the dyad is in the equilibrium state of consensus over an extreme opinion. The other possible equilibrium under implicit argument-communication is that two agents are maximally dissimilar and thus do no longer interact. If a strong faultline aligns with a group-split in opinions, this situation can arise for all pairs of agents in a team. Two agents then either fully agree and are maximally similar, or they are maximally dissimilar and maximally disagree. In this situation, implicit argument-communication can settle into the outcome of stable bi-polarization.

In sum, extreme consensus and perfect bi-polarization are the only stable equilibria in the implicit version, whereas the explicit version can generate moderate consensus as a third possibility. Based on this consideration, we expect that the implicit version of argument-communication is more likely to generate extreme opinion outcomes than the explicit version, all other things being equal.

3.3 Difference in the formalization of homophily

3.3.1 Homophily as likelihood of interaction

This implementation of homophily mirrors the implementation of the X–S model version. When i is selected for an interaction, her potential interaction partners are the remaining team members. The likelihood that the interaction takes place

between i and j is a function of their similarity modeled as a combination of similarity in group identity and opinion similarity. Formally, the similarity between i and j at time point t is:

$$sim_{ij,t} = 1 - \frac{\left(\left|g_i - g_j\right| + h_o \cdot \left|o_{i,t} - o_{j,t}\right|\right)}{2 + 2 \cdot h_o} \tag{5}$$

The similarity $sim_{ij,t}$ can vary between 0 (no similarity) to 1 (perfect similarity). Parameter h_o defines the relative weight of group identity and opinion on the similarity between two agents. For $h_o = 1$, group similarity and opinion similarity have the same impact on the overall similarity between i and j. For $h_o > 1$, opinion dissimilarity weighs more than group similarity, while group similarity is more important if $0 < h_o < 1$.

Throughout this paper we explore two different values for this parameter, $h_o \in \{0.3,3\}$. The lower value, $h_o = 0.3$, represents the assumption that demographic differences weigh much more for defining similarity than opinion differences do. This is similar to the X–S model version, with three different demographic attributes and one opinion attribute, where all attributes have the same weight. By contrast, in the I–E (baseline condition) it is assumed that opinion difference weighs 3 times more than group identity for the similarity between agents. This scenario is, thus, replicated with $h_o = 3$ in our study. Finally, the probability that j is selected as interaction partner is computed for all the network neighbors of i, as a function of the similarity of a particular network partner relative to all other network partners, and is defined as:

$$P_{ij,t} = \frac{\left(sim_{ij,t}\right)^{h_s}}{\sum_{j=1}^{N-1}\left(sim_{ij,t}\right)^{h_s}} \tag{6}$$

In every interaction of an agent i, exactly one of the potential interaction partners j is selected with the probability given by Eq. 6. The strength of homophily is represented by parameter h_s (not to be confused with the parameter h_o): higher values of h_s make the relative similarity between i and j have a bigger effect on the probability that they interact.

3.3.2 Homophily as effectiveness of influence

In this approach (reproducing the setup in I–E), the chances of interaction are independent of the similarity between potential interaction partners. When i is selected to carry out an interaction, an interaction partner j is randomly picked from her teammates, ignoring Eq. 6. Next, the similarity between i and j is calculated according to Eq. 5. The similarity between i and j, however, impacts the magnitude of the opinion change that such an interaction can bring about. More precisely, the similarity $sim_{ij,t}$ moderates the effect that the communicated argument a has on the opinion of agent i as formalized in Eq. 7.

$$o_{i,t+1} = o_{i,t} + a_{i,t} \cdot \left(sim_{ij,t}\right)^{h_p} \tag{7}$$

Mirroring the formalization of the effect of similarity on the probability of interaction (Eq. 6), we include a parameter h_p that scales the impact of the relative similarity between i and j on the effectiveness of the interaction ($h_p > 0$). The larger the value of h_p, the stronger the effect that similarity has on the impact that a communicated argument has on the opinion of its recipient.

Finally, a truncating function ensures that the updated opinion $o_{i,t+1}$ stays within the limits of the opinion scale $[-1, +1]$. That is, whenever an agent's opinion is outside the range of the opinion scale after argument communication, the opinion is set to the value of the closest pole of the scale[7].

3.3.3 Implications

The two notions of homophily are similar in that they both imply that actors who hold similar opinions influence each other more than those who are more dissimilar—this is the core of the reinforcing influence that can drive a group towards bi-polarization or extreme consensus. However, it remains unclear how the model differences exactly affect the chances that the persuasive argument model generates extreme opinion outcomes.

A possible clue lies in the number of interactions that are needed for an agent to develop an extreme opinion, depending on the two versions of homophily. When homophily affects the likelihood of interaction, every interaction can modify the agent's opinion by a fixed amount (i.e. $\pm 2/S$, see Eq. 4). Following previous studies (i.e. both X–S and I–E), here we assume $S = 4$. This means that agents are always respectively *at most* two interactions away from potentially developing an extreme opinion.

By contrast, when homophily affects the effectiveness of the influence, the opinion change is weighted by the similarity between interaction partners. Here, agents are always (not at most but) *at least* two interactions away from potentially becoming extremists. Influence will have relatively little effect especially in the early steps of a cascade of mutually reinforcing influence, when two interacting agents are still relatively dissimilar. This means that conceptualizing homophily as influence effectiveness might make it more difficult for agents to reach an extreme opinion. If agents need more interactions to develop extreme opinions, we can expect two things: first, that extreme consensus or bi-polarization are less likely to emerge within a given time frame; second, that when they do emerge, they do so after a higher number of interaction events compared to the other conceptualization of homophily. In other words, we expect that—all other things being equal—a team with a strong demographic faultline is less likely to develop a group-split or group-polarization within a

[7] Truncation is necessary for the model variant I–E, because under some conditions some interactions may push agents' opinions outside of the range $[-1, +1]$. This can happen when agents with a very positive (or very negative) opinion, e.g. $o_i = \pm 0.9$, receive an opinion push of as big as ± 0.5, according to Eq. 4 with $S = 4$ and $sim_{ij} = \pm 1$.

given time frame if similarity affects the effectiveness of interaction rather than the likelihood of it.

4 Simulation experiments

4.1 Experiment design

In our experiments, we compared the X–S, the I–E and the I–S model. To have a defined point of comparison, we assume a baseline parameter configuration for all three models that represents the scenario we are most interested in. This is a team of a size plausible for real organizations, with a maximally strong demographic faultline. Moreover, in this team group identity is important as a source of similarity, and homophily as well as initial congruency are sufficiently strong so that the emergence of a group-split is possible, but not trivial. The parameter space we explored is tailored to accurately replicate the setup used in earlier work with the X–S and I–E model. Accordingly, we define a baseline scenario with $N=10$, strength of homophily $h_s=h_p=4$, impact of opinion differences on similarity $h_o=3$, congruency $w=0.8$ and agents' memory capacity $S=4$. To assess the robustness of the effects of implementation of reinforcing influence and homophily, these effects will also be explored and reported for some alternative parameter settings. Table 1 provides an overview of the parameter space that was explored in this study.

For every condition inspected in our simulation experiment, we conducted 100 independent simulation runs using NetLogo (Wilensky 1999), where each run was initialized with a different random seed. If not reported otherwise, simulations are run at least 10^4 interaction events per agent, unless the model converged to equilibrium before[8].

Model outcomes at the end of a simulation were measured in two different ways. First, we tested whether an outcome fell into one of the categories of moderate consensus, extreme consensus or bi-polarization. Second, as we are interested in the degree to which the model generates a group-split, we also measured between-group polarization, defined as the absolute value of the distance between the average opinions within group -1 and group $+1$, respectively.

The classification of model outcomes into one of the three categories is most meaningful if model dynamics have converged, i.e. reached a state in which no further change of the distribution of opinions is theoretically possible. This was not feasible in all of the conditions we inspected: Exceptions will be discussed in more detail below (see also endnote vi).

[8] The choice for a limit of 10^4 maximum time steps is arbitrary but motivated by the aim of our study, which is to identify how model outcomes vary across conditions. During our work in preparation for this study we ran explorative simulation runs for a longer number of iterations (up to 5×10^5 iterations). We have found that typically model runs show patterns that are distinctive across conditions and converge to equilibrium in most cases much earlier than after 10^4 iterations. Runs not converging by that time display a distinctive pattern of erratic dynamics, which we elaborate upon at the end of Sect. 4.3. Thus, 10^4 time steps seemed a conservative choice for a threshold.

Table 1 Overview of variables and parameters

Variables	Values range	Description
i, j		Agent identifiers
g_i	$\{-1,+1\}$	Group
o_i	$[-1,+1]$	Opinion
P_i	[0, S]	Number of pro arguments held by agent *i*.
a_i	[− 2/S, + 2/S]	Argument—or 'opinion push'—received by agent *i* from *j* (Eq. 4)
t	$[1, 10^4]$	Time steps
sim_{ij}	[0,1]	Similarity between two agents (Eq. 5)
P_{ij}	[0,1]	Probability that agent i selects j as interaction partner (Eq. 6)
Parameters	**Values**	**Description**
N	{10,100}	Population size
S	{4, 7}	Memory size
w	{0.5, 0.6, 0.7, 0.8, 0.9}	Congruency—the correlation between agents' group and initial opinion.
Argument-communication	(Explicit or implicit)	Explicit as in X–S; implicit as in I–E.
Homophily	(Via likelihood of interaction of influence effectiveness)	Via likelihood of interaction as in X–S; via influence effectiveness as in I–E.
h_o	{0.3, 3}	Relative weight of group identity and opinion in the similarity between two agents
h_s	{1,2,3,4,5}	Strength of homophily in X–S
h_p	{1,2,3,4,5}	Strength of homophily in I–E

Criteria for model convergence and classification of outcomes needed to be tailored to the different model types. If the communication of arguments is modelled implicitly, convergence occurs if and only if dissimilarity between every pair of agents is either maximal ($sim = 0$) or when they agree on the same extreme opinion (± 1). In the former case, influence is impossible because the probability or effectiveness of it is zero. In the latter case, influence cannot alter their opinion because no arguments with a different valence can be adopted. If argument communication is modelled explicitly, convergence occurs when no agent can receive an argument that would change her opinion. Technically, this can happen in two cases. First, all team members hold exactly the same set of arguments and are thus in perfect consensus (extreme or moderate)[9]. Second, in all pairs of agents, they either have exactly the same set of arguments and thus the same opinion, or their similarity is zero, making interaction impossible or influence ineffective.

The conditions for convergence are only met by the three qualitatively different outcomes of moderate consensus, extreme consensus or maximal between-group bi-polarization (equivalent to a group-split). *Moderate consensus* occurs when all agents hold the same opinion and this opinion is neither -1 nor $+1$. This can only be a stable state when argument communication is explicitly modelled. If all agents hold the same vector of arguments containing both pro and con arguments, then no agent has an extreme opinion and no argument can be circulated that would change an agent's opinion. *Extreme consensus* is possible when all agents agree on the same opinion coinciding with one of the poles of the opinion spectrum $[-1, +1]$. When the communication of arguments is explicit, this does not require that agents agree on the same set of arguments, but only that they all possess only arguments of the same valence. In this case no agent can receive an argument with a different valence from an interaction partner and further change is precluded. Similarly, if the argument communication is implicit, no agent has a positive probability of giving a positive (or negative) opinion push to their interaction partner. Either way, when extreme consensus emerges no further influence is possible: extreme consensus is thus a converged outcome. *Maximal between-group bi-polarization*, finally, occurs when both demographic groups have internally reached extreme consensus on the opposite poles of the opinion spectrum. When this occurs, the absolute difference between the average opinions of the two groups equals 2, since 2 is the span of the opinion scale. It is worth noting that the team could bi-polarize also along other lines of division than group identity, but only if the opinion divide overlaps with the group divide neither outgroup agents nor ingroup agents can further influence a focal agent.

[9] This is a softer definition of convergence than adopted in previous implementations (E–S). Here, runs were flagged converged only if all interacting agents would have the same argument set. This allowed the possibility that perfect consensus or bi-polarization were in equilibrium for a long time frame, before agents agreed on the same set of arguments and the run met the convergence criterion.

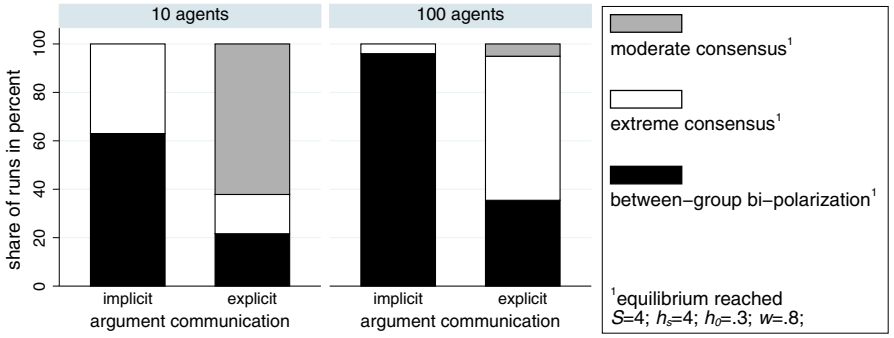

Fig. 1 Effect of implicit vs. explicit argument communication for the baseline scenario and a ceteris-paribus replication with $N = 100$

4.2 Results 1: Effects of explicit or implicit argument communication

We expect that the implicit version of argument communication is more likely to generate extreme opinion outcomes than the explicit version, all other things being equal. To test this expectation, we compared the two versions of the models that conceptualize homophily as the likelihood of interaction but differ in assuming implicit vs. explicit argument-communication. Figure 1 shows results of 100 independent realizations of running these two models under the baseline condition. In addition, we conducted a robustness test with a ceteris-paribus replication with $N = 100$. All simulation runs reached convergence within the limit of 10,000 interaction events per agent so that all outcomes could be classified in the three categories of moderate consensus, extreme consensus and between-group bipolarization. Figure 1 shows how implicit vs. explicit argument communication affected the share of runs ending in each of the three categories.

To begin with, the results show that both model versions can generate a group-split with bi-polarization. Beyond that, Fig. 1 reveals three main findings. First, as anticipated, we did not observe moderate consensus in any run with implicit argument-communication: moderate consensus emerges an outcome only with the explicit version of the argument-communication.

Second, as a consequence of the above and in line with our expectations, we found that the implicit argument-communication generated more extreme opinion outcomes (extreme consensus, between-group bi-polarization) than the explicit version. In the baseline condition ($N = 10$), the implicit argument-communication only generated extreme consensus (in roughly 40% of the runs) and between-group bi-polarization (~60%). Conversely, in the baseline condition the explicit argument-communication produced moderate consensus in most of the simulation runs, while the rest of the runs were roughly evenly divided between extreme consensus and between-group bi-polarization. The absence of moderate consensus as model equilibrium in the implicit version might explain the higher absolute number of runs that converged to bi-polarization (compared to the explicit version).

There might be another reason why the implicit argument-communication generates more between-group bi-polarization. In both the implicit and explicit versions of the argument-communication, an agent does not update her opinion when she receives an argument of the same valence (pro or con) as the one she forgets. However, a central difference between implicit and explicit argument-communication is that the explicit argument-communication carries some additional probability that the communication of an argument does not change the recipient's opinion: this is what we call 'argument redundancy', and happens when the communicated argument is already known to the receiving agent. By contrast, the implicit argument-communication model has no argument redundancy, as it does not track which arguments are considered. As a consequence, the implicit argument-communication generates more opinion changes than explicit argument-communication. This small difference affects the reinforcement process that is responsible for the emergence of bi-polarization. Consider, for instance, an agent who holds 3 pro and 1 con argument. According to the homophily principle, this agent will most likely be exposed to another pro argument, which according to the implicit argument-communication model will likely intensify her positive opinion. Under the explicit argument-communication regime, this is also the most likely outcome, but it is less likely than under implicit argument-communication, as there is also a positive chance that the agent will receive a pro-argument she already considers. This means that, under the explicit argument-communication regime, the self-reinforcing process of homophily and argument-communication is weaker, which makes bi-polarization a less probable and consensus a more probable outcome of influence dynamics.

Third, Fig. 1 reveals that while the model with explicit argument-communication is able to generate moderate consensus, dynamics did not lead the bigger populations into this equilibrium. With $N=100$, moderate consensus emerged only rarely, leading us to the conclusion that this difference between explicit and implicit argument communication affects the long-term outcomes of the dynamics mainly in small teams. However, it should also be noted that teams with 100 members rarely occur, if at all, in real organizations.

This effect of group size in the model with explicit argument communication can be derived from earlier work on the X–S. These studies have shown that moderate consensus is harder to reach in bigger population, because coordination on a single argument vector can take very long in big populations. Even when most agents hold moderate opinions, it is possible that the population will at some moment develop a small bias towards one of the poles of the opinion scale. Due to homophily, agents leaning towards one of the poles will most likely be exposed to further arguments that intensify their opinions. In a population with little opinion variation, agents that moved towards the pole can pull others with them, sparking a collective extremization of opinions, similar to the empirically observed opinion shifts in the experiments of the polarization paradigm from social psychology (Myers and Lamm 1976). In bigger populations, such a scenario is more likely, because it takes these populations longer to reach a consensus on moderate opinions, giving them more time to at some moment develop a small opinion bias that is subsequently intensified.

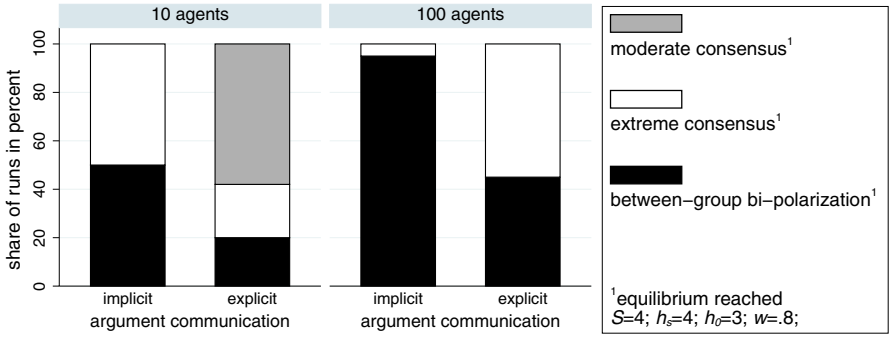

Fig. 2 Effect of implicit vs. explicit argument communication for high impact of opinion d opinion disagreement on similarity ($h_o = 3$). All other parameters are taken from the baseline condition

The effect of the team size in Fig. 1 is the fourth main finding: bigger teams are more likely to experience between-group bi-polarization than smaller teams, under both argument-communication regimes. For the explicit argument-communication, this trend could be the consequence of the previous finding: as simulations with bigger teams were less likely to converge to moderate consensus, there was a higher relative proportion of runs that converged to the other possible outcomes, extreme consensus and between-group bi-polarization. This explanation does not hold for the implicit version of the argument communication, where moderate consensus never emerges as simulation outcome, but still simulation runs were more likely to converge to between-group bi-polarization in big teams than in small teams. This result is both unexpected and puzzling. We acknowledge the need for further research to understand this effect.

To assess the robustness of the four main results of this first experiment, we conducted a ceteris-paribus replication of the experiment shown in Fig. 1 with a higher impact of opinion disagreement on similarity ($h_o = 3$). Figure 2 shows the results.

Figure 2 shows that the four main findings described for Fig. 1 could be replicated. Also under $h_o = 3$, moderate consensus occurs only with explicit argument communication, and extreme opinion outcomes are thus more likely with implicit argument-communication. Concerning the team size, we again find that high N suppresses moderate consensus, and makes between-group bi-polarization more likely. In a further robustness test, we also repeated the experiment of Figs. 1 and 2 with lower and higher initial congruency ($w = 0.5$ and $w = 0.9$). Again, the four main findings could be replicated.

4.3 Results 2: Competing notions of homophily

Both extreme consensus and bi-polarization are expected to occur less likely within a given time frame when homophily affects the effectiveness of an interaction rather than the likelihood that the interaction occurs. In addition, when these outcomes

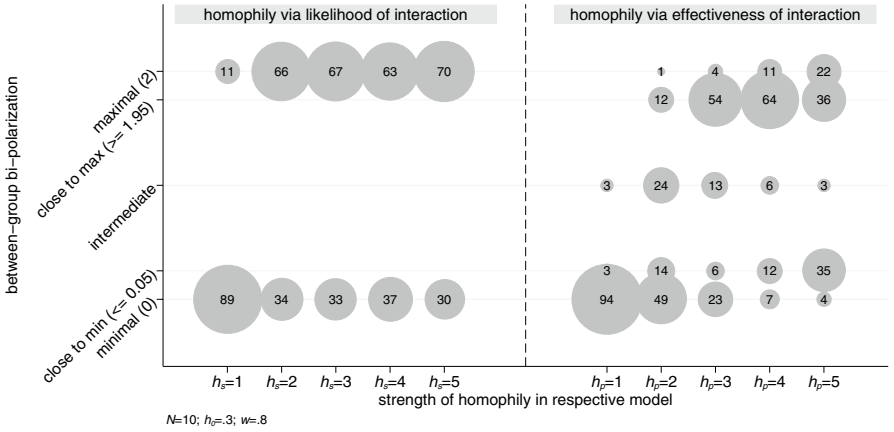

Fig. 3 Effects of the conceptualization of homophily on model predictions after 10,000 opinion updates per agent. Baseline condition, 100 independent realizations per condition

emerge, they should do so after a higher number of interaction events compared to a model in which homophily affects the likelihood of interaction.

To compare the two competing notions of homophily, we used the model with implicit argument communication, combining it with the two different versions of homophily from X–S and I–E. Figure 3 depicts results of this variation for the baseline condition. As a further test, we also varied homophily strength (h_s and h_p). Earlier work on the X–S has shown that homophily strength increases bi-polarization in a model with explicit argument communication (Mäs and Flache 2013). We wanted to know whether this result extends to both versions of the model with implicit argument communication. In the "Appendix", we provide in addition a comparison of the effect of homophily strength across all model versions.

Moderate opinion consensus is not an equilibrium candidate of the two models with implicit argument communication. Therefore, we quantify outcome differences in terms of *between-group bi-polarization*, that is, the absolute difference between the average opinions of the two demographic groups. Figure 3 reports the share of runs with an extreme consensus (between-group polarization = 0), and the share of runs characterized by a perfect split between the two subgroups after 10,000 interaction events per agent (between-group bi-polarization = 2). Since not all runs reached a state of equilibrium, Fig. 3 further shows the share of runs that had not reached equilibrium but ended instead with an opinion distribution very close to either extreme consensus or maximal between-group polarization. Finally, the figure informs about the share of runs that were not close to one of the equilibria even after 10,000 interaction events per agent. The size of the bubbles in Fig. 3 corresponds to the share of runs with the respective opinion distribution. In addition, the labels in the center of the bubble indicate the exact number of runs observed.

Figure 3 shows that both conceptualizations of homophily are able to explain the emergence of opinion consensus and a split into two opposing groups.

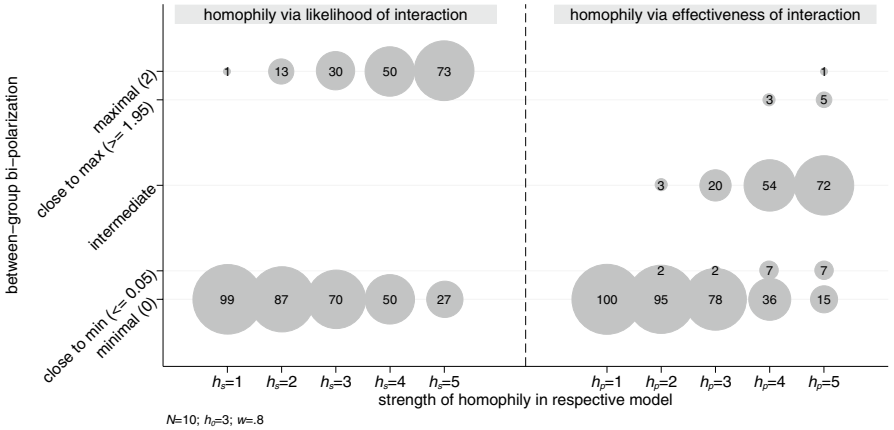

Fig. 4 Effects of the conceptualization of homophily on model predictions after 10,000 opinion updates per agent. Baseline condition except for $h_o = 3$, 100 independent realizations per condition

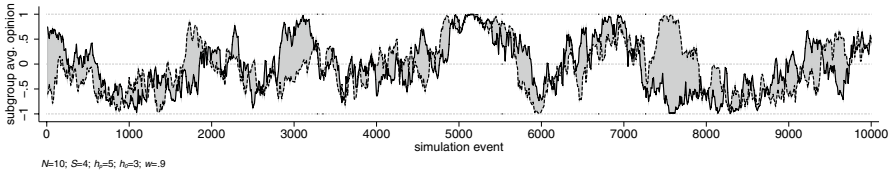

Fig. 5 Evolution of the average opinion in the two groups in an ideal–typical simulation run of the model with homophily conceptualized as interaction effectiveness

Furthermore, both models generate more group splits when homophily is stronger (see h_s and h_p). The central difference between the two models is that the dynamics of the model with homophily conceptualized as influence effectiveness tend to rest longer in intermediate states and opinion distributions that are close to one of the equilibria of the dynamics. This results in less extreme outcomes after 10,000 interaction events, consistently with our expectation.

As a further test, we conducted a ceteris-paribus replication of this experiment with a high value for the impact of opinion disagreement on similarity ($h_o = 3$). Figure 4 reports the results and shows that the qualitative effects of the conceptualization of homophily are robust against this variation the parameter h_o.

Figures 3 and 4 and additional results provided in the "Appendix", demonstrate that under the model that combines implicit argument communication and homophily via influence effectiveness, and for most values of homophily strength (except for $h_o = 3$ and $h_s = h_p = 1$, see Fig. 4), a considerably smaller share of runs converges to perfect consensus or maximal bi-polarization than in the other models. This is shown by the fact that all runs display either minimal or maximal between-group bi-polarization in the left-hand panels of Figs. 3 and 4, whereas in most cases the right-hand panels show runs with non-extreme values. The reason

is that unlike the other two models, this model (with homophily via influence effectiveness) generates dynamics that can last very long before an equilibrium is reached. For illustration, Fig. 5 shows the trajectories of the average opinions in the two subgroups in an ideal–typical simulation run. The size of the gray area at a given simulation event, thus, shows the degree of between-group bi-polarization. The dynamics are very rich. For instance, after every agent's opinion had been updated 5140 times, in one of the two groups, all members had adopted an opinion of 1. The average opinion in the other group was close to the same pole (avg = .99155) but there was still one agent who was not maximally extreme ($o_{i,t}$ = .95775). While the population was very close to reaching an extreme consensus, Eq. 3 implies that it is very unlikely that the opinion of this agent shifted onto the pole. This shift required that the agent dropped a con argument, an event that is very unlikely when an agent holds an opinion so close to the opinion pole of + 1. Instead, dynamics moved the system away from the consensus equilibrium when the agent at some moment communicated a con argument to one of the other agents.

Only the combination of homophily via influence effectiveness and implicit argument communication makes it possible that outcomes can occur in which the opinion distribution moves so close to one of the extreme outcomes that it is almost impossible to reach eventually. The reason is that the weighing of the impact of an argument with similarity can result in extremely small steps of opinion change, especially with high homophily strength. This property of the model also makes the opposite equilibrium (bi-polarization) very difficult to reach, as is also illustrated by the ideal–typical run in Fig. 5. After 7560 opinion updates per agent, the opinion averages in the two groups were − 98348 and .98001 respectively. In other words, the population was highly bi-polarized. Also in this setting, however, it was relatively unlikely that the population reached the state of maximal bi-polarization. To eventually reach the pole towards which a subgroup tended, all of its members would have needed to drop their last remaining opposite argument and replace it with one leading them all the way towards the extreme. However, as Eq. 3 implies, this was very unlikely to happen because in this situation agents held on average only about 1% arguments in favor of the opposite end of the opinion spectrum.

Reaching a perfectly bi-polarized opinion distribution is particularly time consuming when homophily (h_p) is strong. The problem is that an agent who has adopted a maximally extreme opinion will likely be exposed to an argument challenging her opinion when interacting with an agent holding a very different opinion. The resulting opinion shift away from the opinion pole, however, can be extremely small when h_p adopts high values. The updated opinion value will therefore be very close to the opinion pole, making it extremely unlikely that the agent drops the counter argument and returns to a maximally extreme view.

Figure 6 shows results that test the expectation that simulations take on average more interaction events before a convergence state is reached, when homophily was implemented via the effectiveness of influence rather than via likelihood of interaction. For Fig. 6, we conducted simulation experiments for the baseline condition with a time limit of 5×10^5 interaction events per agent. We found that all runs

Fig. 6 Boxplots showing the effect of the conceptualization of homophily on the logarithm of the duration of the dynamics measured in simulation events (100 simulation runs per treatment, $S=4$, $S_0=3$, $w=.8$, $h_o=.3$; $N=10$, only models with implicit argument-communication)

but one reached a stable rest point. Runs of the model with homophily modeled as influence effectiveness lasted particularly long when homophily was strong. The opposite effect, however, was found when homophily was implemented as increased likelihood of interaction. Here, homophily strength decreased the duration of the dynamics, an effect that is strong but visually diluted by the logarithmic scale of the y-axis in Fig. 6.

To further assess the robustness of results for the effects of the implementation of homophily on between-group bi-polarization, additional tests were conducted, varying initial congruency (w), and the number of arguments in agents' memory (S) for a specific scenario in which a group split was possible, but not easy to obtain within 10,000 interaction events per agent. The reported finding turned out to be robust. In the models with implicit argument communication, extreme outcomes were less likely to emerge within a given time frame for the model with effectiveness-homophily. Further results of these tests are reported in the "Appendix".

5 Conclusion and discussion

Strong demographic faultlines have been identified as a possible reason why diversity can hamper a team's cohesion and performance. While empirical research has pointed to a number of moderating conditions for the effects of demographic faultlines, computational modellers have recently begun to address the task of understanding these effects with models of the complex and interdependent dynamics of social relations and social influence in teams. We focused here on one agent-based modelling approach, the model of persuasive argument communication proposed by Mäs et al. (2013), Mäs and Bischofberger (2015) and Mäs and Flache (2013), which closely builds on the fundamental processes of reinforcing influence and homophily central in Lau and Murnighan's (1998) original theory of faultlines.

The model of persuasive argument communication points to intriguing theoretical hypotheses about the dynamics of group splits in teams. First, it highlights a number of conditions that are required before a demographic faultline can really induce a

group-split, including sufficiently strong initial congruency of opinions and demographics, and sufficiently strong homophily. Second, it allows explaining group-split dynamics without making the empirically debated assumption of repulsive forces in social influence, used by earlier formal accounts of group-split dynamics. In addition, Mäs et al. (2013) showed how the addition of "criss-crossing" actors connecting demographically separated subgroups could prevent group-split dynamics despite a strong faultline in a team. As such, the model of persuasive argument communication highlights theoretical directions research could take to test possible strategies organizations could employ to preclude between-group polarization in teams with a strong faultline. This potential practical use of the model of persuasive argument communication makes it highly important to carefully assess the robustness of its main theoretical predictions against alternative theoretically plausible specifications of the micro-processes of reinforcing influence and homophily that are at the heart of the model.

A comparison with alternative modelling approaches in formal models of social influence in general (cf. Flache et al. 2017) and more recent implementations of persuasive argument communication in particular (Feliciani et al. 2017) reveals two important distinctions in both processes. In modelling reinforcing influence, arguments can be explicitly represented in a model or be implicit, inferred from the opinions agents adopt. In modelling homophily, similarity can be assumed to affect the likelihood or the effectiveness of an interaction in which arguments are communicated. We developed a modelling framework that allowed us to separately compare the effects of both distinctions on model dynamics across a new 'hybrid' model and two earlier implementations proposed in the literature.

We found that all three model versions could generate the outcome of group-split in a team with a strong faultline, but there are also differences in the conditions and the dynamics of between-group bi-polarization across the models. We observed that implicit argument communication generated more between-groups bi-polarization and extreme consensus than the explicit argument-communication. This difference is explained by the fact that between-group bi-polarization and extreme consensus are the only outcome equilibria in the implicit argument-communication regime, whereas the explicit version produces a third possible outcome, moderate consensus. We found that this difference was limited to teams of small size, however: in large teams, the emergence of moderate consensus is highly unlikely with the explicit version, too. Additionally, we found that the team's size has an interesting and unanticipated effect on the implicit argument-communication: we observed that, under most parameter configurations, between-group bi-polarization is more likely to emerge in big teams than in small ones.

Also the conceptualization of homophily affected team dynamics in our theoretical studies. When homophily affected the likelihood of interaction rather than its effectiveness, this resulted in more extreme opinion outcomes. Simulations more likely converged on bi-polarization or extreme consensus within a given time frame, and group-splits were more likely to occur.

In a nutshell, the present research generally supports the robustness of the persuasive argument model as a tool to theoretically disentangle the dynamics and conditions of group-split in teams with a strong demographic faultline. At the same

time, our study has pointed to a number of potential limitations that highlight directions for future research and underline important lessons for computational models of team dynamics in a more general sense. Our study showed in particular how plausible and seemingly small modifications in the formalization of micro-processes of social influence can have profound consequences for model dynamics and theoretical predictions that require careful inspection. For example, the model with implicit argument communication was found to not generate the outcome of moderate consensus that is possible with explicit argument communication, also not in the small-sized populations that are relevant for modelling teams in organizations. A further example of a profound and intricate effect of a seemingly small change were the erratic opinion dynamics that was induced by the combination of homophily via effectiveness of interaction, and implicit argument communication. In this model, opinion differences between stable and unstable outcomes could become so small over time, that careful inspection of the robustness of its implications against the problem of floating point inaccuracy (Galán et al. 2009; Polhill et al. 2005) is needed in future studies employing similar combinations of assumptions[10].

Notwithstanding its limitations, our research has highlighted potentially highly relevant insights, not only for research on organizations facing diversity in teams with strong faultlines. The important role of the conceptualization of homophily that we identified resonates with a wider debate on the forms of homophily. In real life, opinion formation processes by argument communication can happen in situations where homophily can play out in either one of the two ways, or both at the same time. In an online social network, for example, homophily can typically take both forms. To begin with, homophily can influence who a user might decide to engage in a discussion with, whom to send a 'friendship request', or whom to 'follow'. In this sense, homophily affects the likelihood of interaction. At the same time, the opinion of close friends and relatives might count more than the opinion of complete strangers. Here, homophily is affecting the effectiveness of the interaction. In some other settings, there are structural constraints that exclude the individual from the decision of whom to interact with. This is the case, for example, when all discussants are asked to present their point of view in turns, so that everybody is exposed to everybody else in the same manner. Such interaction structure rules out the possibility for homophily to affect the chances of interaction, while it can still affect the effectiveness of the interaction: one can still choose whether to care, or to pay attention to whomever is speaking, based on the opinion that is being expressed, or the group identity of who is voicing it.

Our theoretical results also suggest an intriguing interpretation of empirical findings of Strandberg et al. (2017) about conditions for polarization in a group discussion where the discussion is either moderated or not. In a moderated discussion, a facilitator guides the discussion by ensuring inclusion of discussants

[10] To assess this possibility, we replicated the model independently with Mathematica (Wolfram Research 2018), a programming tool that allows for much higher numerical precision than we could obtain with NetLogo. For those conditions we replicated, no qualitative differences could be found in the long-term implications of the models.

and equality of the discussion. In terms of our implementations of homophily in persuasive argument communication, the moderation of a discussion limits homophily via likelihood of interaction, as participants cannot choose whom to interact with based on their preferences. Therefore, we can say that a moderated discussion favors homophily via effectiveness of interaction rather than as likelihood of interaction. People may put more weight on arguments obtained from others they agree with, but they cannot escape exposure to argument also from actors with dissimilar views. The main finding of this empirical study was that discussion between like-minded individuals is significantly less likely to produce group polarization when the discussion was moderated, compared to a free discussion. This finding resonates with the predictions of the model of persuasive argument communication: with homophily via effectiveness of interaction, the model predicts that extreme consensus and bi-polarization are less likely to emerge.

These considerations highlight interesting directions for future theoretical and empirical work. On the one hand, theoretical work could explore whether other processes of opinion formation are sensitive/to the way how homophily is conceptualized and implemented. On the other hand, empirical work could test these effects in real world settings with experiments by manipulating—for example—the link between similarity of opinions and likelihood of interaction. First studies (see Mäs and Flache 2013) along this line have been conducted. We believe our theoretical work here has provided new insights that could inspire future empirical tests.

Acknowledgements The authors' list is ordered by contribution. The authors are grateful to the members of the research group Norms and Networks at the University of Groningen for their helpful feedback, and would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Appendix

Effects of congruency, *w*, on the argument-communication (explicit versus implicit)

Figures 7 and 8 show the robustness of the findings concerning the effects of the explicit versus implicit argument-communication to changes in the congruency parameter (see Fig. 1). Congruency affects the correlation between agent's initial opinion and group identity—since higher congruency implies stronger between-group bi-polarization at the outset of the simulation run, we expect that congruency

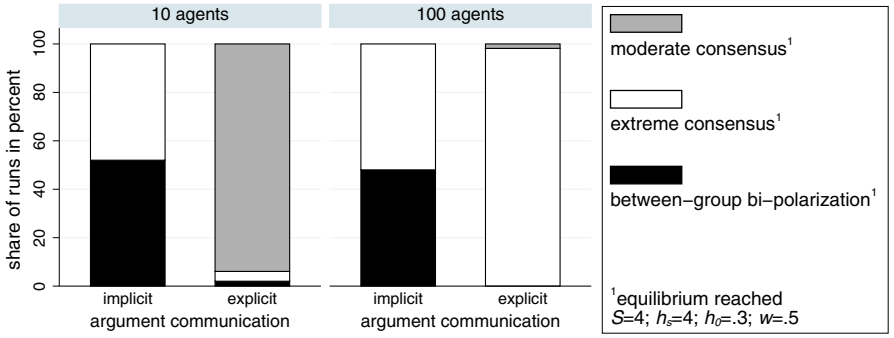

Fig. 7 Replication of the experiment in Fig. 1 with $w = 0.5$

increases the chances that the team converges to between-group bi-polarization to emerge.

Figures 7 (no congruency: $w = 0.5$) and 8 (strong congruency: $w = 0.9$) show that the results are consistent with this expectation. Furthermore, the four main findings reported in our results are confirmed (see Fig. 1). Specifically, we observe that (1) the implicit version of the argument-communication does not converge to moderate consensus; (2) the implicit version produces more extreme opinion outcomes (that is, more extreme consensus and more between-group bi-polarization); and (3) in big teams, moderate consensus is rare even for the explicit argument-communication. Lastly, we observed that (4) between-group bi-polarization is more likely in big teams than in small teams. The latter finding does not appear for the explicit argument-communication with no congruency ($w = 0.5$), where between-group bi-polarization hardly ever emerged regardless of the size of team.

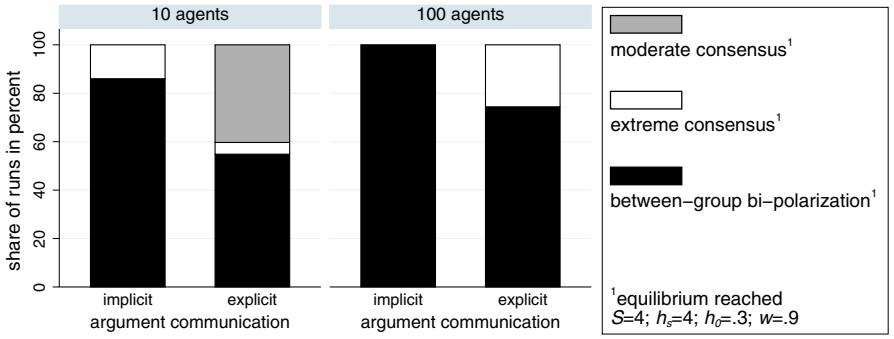

Fig. 8 Replication of the experiment in Fig. 1 with $w = 0.9$

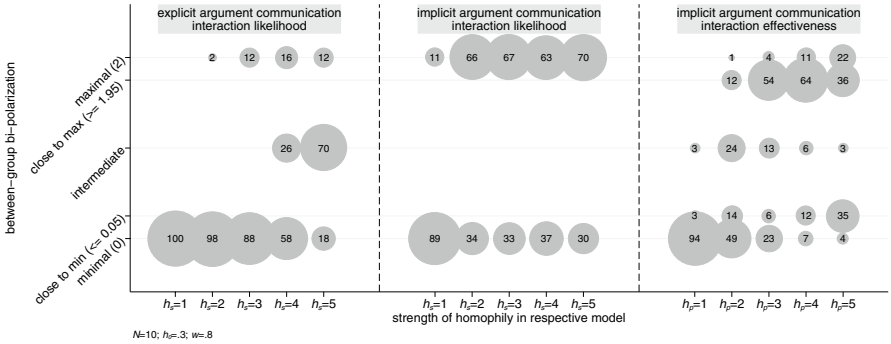

Fig. 9 Experiment of Fig. 3 main paper extended with model combining explicit argument communication with homophily via likelihood of interaction. From left to right, the panels show the setup of X–S, the hybrid version, and I–E

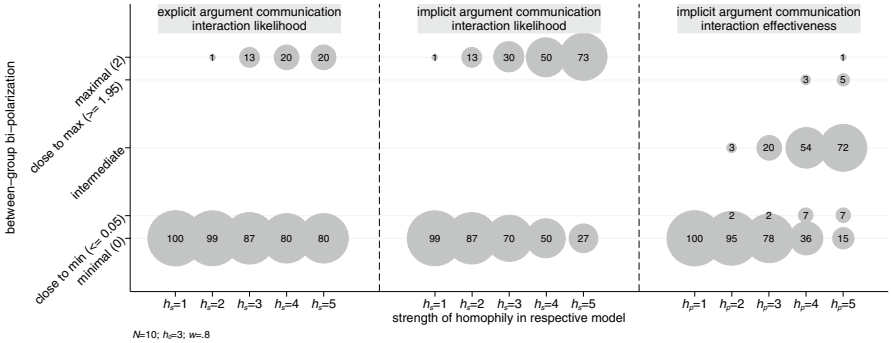

Fig. 10 Experiment of Fig. 4 main paper extended with model combining explicit argument communication with homophily via likelihood of interaction. From left to right, the panels show the setup of X–S, the hybrid version, and I–E

Effects of homophily-strength on between-group disagreement, compared across all three models

In Figs. 3 and 4 we reported results comparing different implementations of homophily, always assuming an implicit argument-communication. With Figs. 9 and 10 we integrate these results, by comparing them with a model version where the argument-communication is modelled explicitly (as in the X–S).

The two main findings reported previously extend to the third model version: we observe that all model versions can generate between-group bi-polarization, and that higher homophily strength (*hs* and *hp*) has a positive effect on the share of runs that converged to perfect between-group bi-polarization.

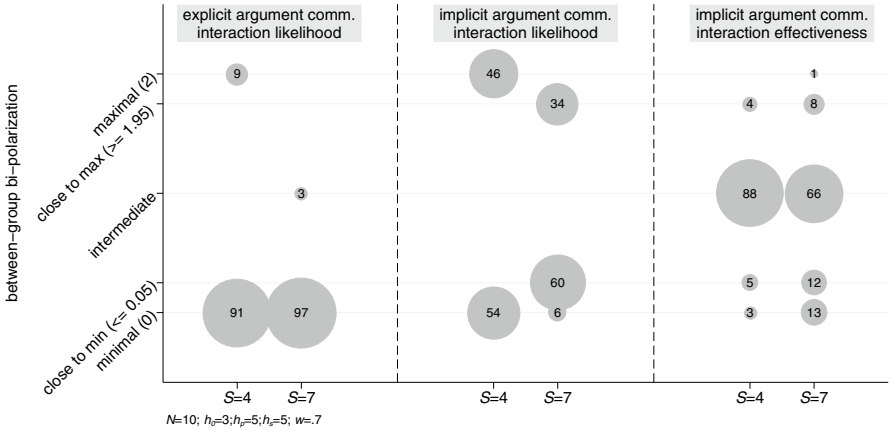

Fig. 11 Effect of the number S of arguments that agents consider in the three models (100 replications per condition). From left to right, the panels show the setup of X–S, the hybrid version, and I–E

Effects of agents' memory size, *S*, compared across all three models

Figure 11 informs about how between-group polarization is affected by the number S of arguments that agents consider, comparing $S=4$ with $S=7$. We selected a specific scenario in which between-group bi-polarization within 10,000 interaction events is possible, but not easy to obtain, and compared the effects of the number of arguments across all three models. In line with our expectations, we observe that among the models with implicit argument communication convergence on extreme outcomes is less likely in the model with effectiveness-homophily, than in the model with interaction-homophily. This result appears to be robust across $S=4$ and $S=7$. For the two models with homophily conceptualized as an increased chance of interaction, Fig. 11 suggests furthermore that higher numbers of arguments make opinion polarization less likely. This effect obtains because the parameter S affects the size of the opinion shifts (see Eq. 4). A higher number of arguments implies that opinion shifts are smaller when agents adopt a new argument. As a consequence, it takes longer until the two subgroups have adopted very different opinion, which in turn increases the chances that arguments are communicated across subgroup boundaries and create consensus. For the model with homophily conceptualized as influence effectiveness, the precise effect of S is relatively hard to interpret because the majority of the simulations had not reached equilibrium or a state very close to one of the equilibria.

Effects of congruency, *w*, compared across all three models

Figure 12 shows how the initial degree w of congruency affects model outcomes. Again, we chose a specific scenario in which between-group bi-polarization within 10,000 interaction events is possible, but not easy to obtain, and compared the effects

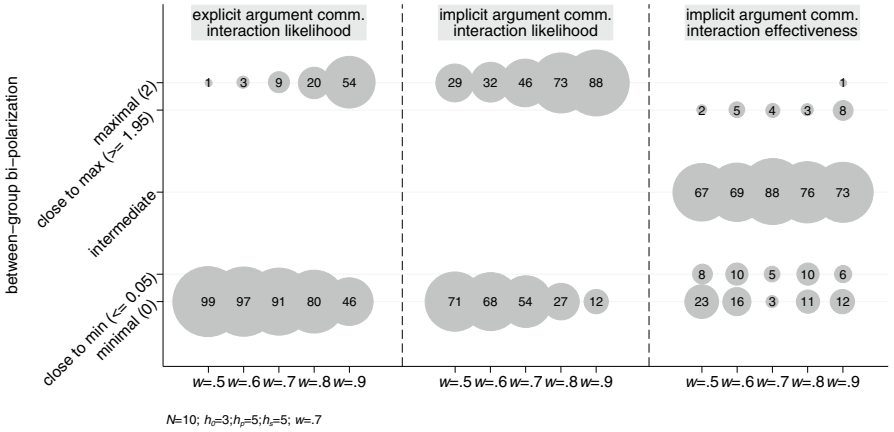

Fig. 12 Effect of the initial degree w of congruency (100 replications per condition). From left to right, the panels show the setup of X–S, the hybrid version, and I–E

of congruency w on between-group polarization across all three models. In line with our expectations, we observe across all levels of congruency that among the models with implicit argument communication convergence on extreme outcomes is less likely in the model with effectiveness-homophily, than in the model with interaction-homophily. For all three models, we found that opinion polarization is more likely for higher levels of w, where dynamics depart from a more bi-polarized state. This effect, however, is relatively weak for the model that combines implicit argument communication with homophily conceptualized as influence effectiveness. The reason is again the property of the model to generate unstable dynamics that lead the system very close to an equilibrium but also very likely lead them away from this rest point. The initial opinion distribution, therefore, affected the long-term model outcomes only in those runs that happen to reach an equilibrium very early.

References

Anzola D, Barbrook-Johnson P, Cano JI (2017) Self-organization and social science. Comput Math Org Theory 23(2):221–257. https://doi.org/10.1007/s10588-016-9224-2

Axelrod R (1997) The dissemination of culture—a model with local convergence and global polarization. J Confl Resolut 41(2):203–226

Baldassarri D, Bearman P (2007) Dynamics of political polarization. Am Sociol Rev 72(5):784–811

Banisch S (2010) Unfreezing social dynamics: synchronous update and dissimilation. In: A Ernst, S Kuhn (eds) Proceedings of the 3rd world congress on social simulation (WCSS 2010), Kassel

Banisch S, Olbrich E (2017) Opinion Polarization by Learning from Social Feedback. ArXiv Preprint (arXiv:1704.02890)

Bowers CA, Pharmer JA, Salas E (2000) When member homogeneity is needed in work teams: a meta-analysis. Small Group Res 31(3):305–327

Byrne D (1971) The attraction paradigm. Academic Press, New York

Carter AB, Phillips KW (2017) The double-edged sword of diversity: toward a dual pathway model. Soc Pers Psychol Compass 11(5):e12313. https://doi.org/10.1111/spc3.12313

Chen L, Gable GG, Hu H (2013) Communication and organizational social networks: a simulation model. Comput Math Org Theory 19(4):460–479. https://doi.org/10.1007/s10588-012-9131-0
Conte R, Giardini F (2016) Towards computational and behavioral social science. Eur Psychol 21(2):131
Cowan N (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci 24(1):87–114
Dandekar P, Goel A, Lee DT (2013) Biased assimilation, homophily, and the dynamics of polarization. Proc Natl Acad Sci USA 110(15):5791–5796
Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. Adv Complex Syst 03(01n04):87–98. https://doi.org/10.1142/S0219525900000078
Duclos JY, Esteban J-M, Ray D (2004) Polarization: concepts, measurement, estimation. Econometrica 72(6):1737–1772
Duggins P (2017) A psychologically-motivated model of opinion change with applications to american politics. J Artif Soc Soc Simul. https://doi.org/10.18564/jasss.3316
Ellemers N, Rink F (2016) Diversity in work groups. Curr Opin Psychol 11:49–53
Feld SL (1982) Social structural determinants of similarity among associates. Am Sociol Rev 47(6):797–801
Feliciani T, Flache A, Tolsma J (2017) How, when and where can spatial segregation induce opinion polarization? Two Competing Models. *JASSS* 20(2):6. https://doi.org/10.18564/jasss.3419
Flache A, Macy MW (2011) Small worlds and cultural polarization. J Math Soc 35(1–3):146–176. https://doi.org/10.1080/0022250X.2010.532261
Flache A, Mäs M (2008a) 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 Org Theory 14(1):23–51. https://doi.org/10.1007/s10588-008-9019-1
Flache A, Mäs M (2008b) Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion. Simul Model Pract Theory 16(2):175–191
Flache A et al (2017) Models of social influence: towards the next frontiers. J Artif Soc Soc Simul 20(4):2
Fu G, Zhang W (2016) Opinion formation and bi-polarization with biased assimilation and homophily. Phys A 444:700–712
Galán JM et al (2009) Errors and artefacts in agent-based modelling. JASSS 12(1):1
Grow A, Flache A (2011) How attitude certainty tempers the effects of faultlines in demographically diverse teams. Comput Math Org Theory 17(2):196–224. https://doi.org/10.1007/s10588-011-9087-5
Harrison JR, Carroll GR (2002) The dynamics of cultural influence networks. Comput Math Org Theory 8(1):5–30. https://doi.org/10.1023/A:1015142219808
Hegselmann R, Krause U (2002) 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
Isenberg DJ (1986) Group polarization: a critical review and meta-analysis. J Pers Soc Psychol 50(6):1141
Kitts JA (2006) Social influence and the emergence of norms amid ties of amity and enmity. Simul Model Pract Theory 14(4):407–422
Kurahashi-Nakamura T, Mäs M, Lorenz J (2016) Robust clustering in generalized bounded confidence models. J Artif Soc Soc Simul. https://doi.org/10.18564/jasss.3220
La Rocca CE, Braunstein LA, Vazquez F (2014) The influence of persuasion in opinion formation and polarization. EPL 106(4):40004
Lau DC, Murnighan JK (1998) Demographic diversity and faultlines: the compositional dynamics of organizational groups. Acad Manag Rev 23(2):325–340. https://doi.org/10.5465/AMR.1998.533229
Lau DC, Murnighan JK (2005) Interactions within groups and subgroups: the effects of demographic faultlines. Acad Manag J 48(4):645–659
Lazarsfeld PF, Merton RK (1954) Friendship and social process: a substantive and methodological analysis. In: Berger M, Abel T, Page CH (eds) Freedom and control in modern society. Van Nostrand, New York, pp 18–66
Leslie LM (2017) A status-based multilevel model of ethnic diversity and work unit performance. J Manag 43(2):426–454. https://doi.org/10.1177/0149206314535436
Lindenberg S (1992) The method of decreasing abstraction. In: Coleman JS, Fararo TJ (eds) Rational choice theory. Advocacy and critique. Sage, Newbury Park, pp 3–20
Liu Q, Wang X, Zhao J (2015) Multi-agent model of group polarisation with biased assimilation of arguments. IET Control Theory Appl 9(3):485–492
Mark NP (1998) Beyond individual differences: social differentiation from first principles. Am Sociol Rev 63(3):309–330

Mark NP (2003) Culture and competition: homophily and distancing explanations for cultural niches. Am Sociol Rev 68(3):319–345
Mäs M, Flache A (2013) Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. PLoS ONE 8(11):e74516
Mäs M, Flache A, Helbing D (2010) Individualization as driving force of clustering phenomena in humans. PLoS Comput Biol 6(10):e1000959
Mäs M, Flache A, Takács K, Jehn KA (2013) In the short term we divide, in the long term we unite: demographic crisscrossing and the effects of faultlines on subgroup polarization. Org Sci 24(3):716–736. https://doi.org/10.1287/orsc.1120.0767
Mäs M, Bischofberger L (2015) Will the personalization of online social networks foster opinion polarization? SSRN Electron J. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2553436%5Cnhttp://papers.ssrn.com/abstract=2553436
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27(1):415–444. https://doi.org/10.1146/annurev.soc.27.1.415
Meyer B, Glenz A, Antino M, Rico R, González-Romá V (2014) Faultlines and subgroups. Small Group Res 45(6):633–670. https://doi.org/10.1177/1046496414552195
Miller GA (1956) The magical number 7, plus or minus 2—some limits on our capacity for processing information. Psychol Rev 63(2):81
Milliken FJ, Martins LL (1996) Searching for common threads: understanding the multiple effects of diversity in organizational groups. Acad Manag Rev 21(2):402–433
Myers DG (1978) Polarizing effects of social-comparison. J Exp Soc Psychol 14(6):554–563
Myers DG (1982) Polarizing effects of social interaction. In: Brandstätter H, Davis JH, Stocker-Kreichgauer G (eds) Group decision making. Academic Press, London, pp 125–161
Myers DG, Lamm H (1976) The group polarization phenomenon. Psychol Bull 83(4):602
Pelled LH (1996) Demographic diversity, conflict, and work group outcomes: an intervening process theory. Org Sci 7(6):615–631
Phillips KW (2003) The effects of categorically based expectations on minority influence: the importance of congruence. Pers Soc Psychol Bull 29(1):3–13
Phillips KW, Mannix EA, Neale MA, Gruenfeld DH (2004) Diverse groups and information sharing: the effects of congruent ties. J Exp Soc Psychol 40(4):497–510
Pinasco JP, Semeshenko V, Balenzuela P (2017) Modeling opinion dynamics: theoretical analysis and continuous approximation. Chaos, Solitons Fractals 98:210–215
Polhill JG, Izquierdo LR, Gotts NM (2005) The ghost in the model (and other effects of floating point arithmetic). JASSS
Reagans R (2011) Close encounters: analyzing how social similarity and propinquity contribute to strong network connections. Org Sci 22(4):835–849
Rouchier J, Tubaro P, Emery C (2014) Opinion transmission in organizations: an agent-based modeling approach. Comput Math Org Theory 20(3):252–277. https://doi.org/10.1007/s10588-013-9161-2
Secchi D, Gullekson NL (2016) Individual and organizational conditions for the emergence and evolution of bandwagons. Computat Math Org Theory 22(1):88–133
Shemla M, Meyer B, Greer L, Jehn KA (2016) A review of perceived diversity in teams: does how members perceive their team's composition affect team processes and outcomes? J Org Behav 37:S89–S106
Stewart GL (2006) A meta-analytic review of relationships between team design features and team performance. J Manag 32(1):29–55
Strandberg K, Himmelroos S, Grönlund K (2017) Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization. Int Polit Sci Rev. https://doi.org/10.1177/0192512117692136
Takács K, Flache A, Mäs M (2016) Discrepancy and disliking do not induce negative opinion shifts. PLoS ONE 11(6):e0157948. https://doi.org/10.1371/journal.pone.0157948
van Dijk H, Meyer B, van Engen M, Loyd DL (2017) Microdynamics in diverse teams: a review and integration of the diversity and stereotyping literatures. Acad Manag Ann 11(1):517–557
Van Knippenberg D, Schippers M (2007) Work group diversity. Annu Rev Psychol 58(1):515–541
Vinokur A, Burnstein E (1978) Depolarization of attitudes in groups. J Pers Soc Psychol 36(8):872–885
von Hohenberg BC, Maes M, Pradelski BSR (2017) Micro influence and macro dynamics of opinions. SSRN Electron J. https://www.ssrn.com/abstract=2974413

Wang D, Pi X, Pan Y (2017) The interpersonal diffusion mechanism of unethical behavior in groups: a social network perspective. Comput Math Org Theory 23(2):271–292. https://doi.org/10.1007/s10588-016-9226-0

Webber SS, Donahue LM (2001) Impact of highly and less job-related diversity on work group cohesion and performance: a meta-analysis. J Manag 27(2):141–162

Wilensky U (1999) NetLogo. http://Ccl.Northwestern.Edu/Netlogo/. Center for connected learning and computer-based modeling. Northwestern University, Evanston IL 2009. http://ccl.northwestern.edu/netlogo/. Accessed 26 Feb 2009

Williams KY, O'Reilly CA (1998) Demography and diversity in organizations: a review of 40 years off research. In: Research in organizational behavior

Wimmer A, Lewis K (2010) Beyond and below racial homophily: ERG models of a friendship network documented on facebook. Am J Sociol 116(2):583–642

Wolfram Research, Inc (2018) Mathematica

Thomas Feliciani is a PhD student at the ICS/Department of Sociology at the University of Groningen and a research assistant at the School of Sociology, at the University College Dublin. His research is focused on processes of social influence and peer review, which he studies with the methods of computational social science.

Andreas Flache took his master in computer science from the University of Koblenz-Landau and his Ph.D. in social sciences from the University of Groningen (1996). He is professor of sociology at the Department of Sociology/ICS of the University of Groningen. His main research interests concern social integration and cooperation, in particular in relation to social networks. He uses agent-based and game-theoretical modeling, social network research, laboratory experiments, survey- and interview research. Recent work has been published in sociological and social-psychological journals (e.g. Journal of Mathematical Sociology, Personality and Social Psychology Bulletin) as well as in other disciplines and interdisciplinary outlets (e.g. PLoS One, PLoS Computational Biology).

Michael Mäs is assistant professor at the Department of Sociology and the Interuniversity Center for Social Science Theory and Methodology (ICS) at the University of Groningen. He works on collective action and social integration in social networks using computational modeling techniques and laboratory experiments. He had a postdoctoral stay at the Chair of Computational Social Sciences at ETH Zurich.