Advances in Complex Systems
Vol. 21, Nos. 6 & 7 (2018) 1850017 (32 pages)

DOI: 10.1142/S0219525918500170

ABOUT RENEGADES AND OUTGROUP HATERS: MODELING THE LINK BETWEEN SOCIAL INFLUENCE AND INTERGROUP ATTITUDES

ANDREAS FLACHE

Department of Sociology,
University of Groningen, Grote Kruisstraat 2/1,
9712 TG Groningen, The Netherlands
a.flache@rug.nl

Received 4 November 2017
Revised 12 July 2018
Accepted 19 July 2018
Published 11 September 2018

Polarization between groups is a major topic of contemporary societal debate and research. Formal models of opinion dynamics try to explain how intergroup polarization arises from simple first principles of social interaction. In existing models, intergroup attitudes affect social influence in the form of homophily or xenophobia, fixed tendencies of individuals to be more open to influence from ingroup members or distance themselves from attitudes of outgroup members. These models generate polarization between groups, but they neglect a central insight from empirical research. Intergroup attitudes are themselves subject to social influence in interactions with both in- and outgroup members. A model is proposed in which the attitude which is subject to social influence is also an intergroup attitude. It affects in turn the influence process itself. Furthermore, it is shown how this changes model predictions about process and conditions of polarization between groups. More complex patterns of intergroup relations emerge than in a model with fixed xenophobia. Especially, a renegade minority ('outgroup lovers') is found to have a key role in avoiding mutually negative intergroup relations and even elicit reversed polarization, resulting in a majority of individuals developing a negative attitude towards their ingroup and a positive one for the outgroup.

Keywords: Opinion dynamics; intergroup attitudes; polarization; social influence; homophily.

1. Introduction

It is a central question in research on opinion dynamics whether and under what conditions societies polarize, falling apart into deeply antagonistic factions. Recent developments that underscore the importance of this question are, for example, the rise of right-wing populist parties in many Western societies, the Brexit referendum in the U.K., or the trend towards deepening divisions in the political landscape of the U.S. [1, 14, 18, 22, 29, 41].

Research on intergroup relations points to negative intergroup attitudes as an important source of opinion divisions in society. Intergroup attitudes are positive or negative attitudes individuals hold about different groups in their society, including their ingroup. Negative attitudes towards outgroups can arise from feelings of threat individuals perceive in the face of immigration or ethnic diversity in their social environment [48, 50, 58, 68, 60]. Negative outgroup attitudes can also form when people generalize from experiences with individual members of outgroups toward their attitude about the outgroup as whole. Studies in the tradition of contact theory [2, 16, 55], as well as research on attitude generalization and stereotype change [37, 38, 53] demonstrated generalization from the interpersonal to the group level both for positive as well as negative interpersonal experiences [15, 63].

Attitudes about a group not only arise from interpersonal interactions and perceptions of threat, but they also shape how individuals interact with and are influenced by members of different groups. This role of intergroup attitudes has been recognized by researchers developing formal models of the dynamics of opinion formation. A number of models [3, 5, 17, 21, 23, 36, 43, 57] incorporated the possibility of repulsive (or, negative) social influence, assuming that individuals strive to be dissimilar to people they dislike, accentuating disagreement with others if these are perceived as being too discrepant (also called xenophobia).

Building on research on intergroup dynamics and social identity [7, 64], further models [59] assume that individuals may change their opinion to adopt a position prototypical for their ingroup or to distance themselves from an opinion perceived as prototypical for an outgroup (see also [34, 35] for a similar approach). It has in particular been assumed that perceived dissimilarity not only arises from disagreement in opinions between individuals, but also from demographic differences between individuals representing fixed characteristics like gender or ethnicity [20, 24, 25, 31, 47]. These models show how demographic diversity can give rise to polarization between demographic groups, in particular when these groups differ in a number of demographic dimensions at the same time, generating demographic faultlines [39] in a population.

Formal models of opinion dynamics take into account that negative intergroup attitudes can foster disagreement and polarization between groups, but to my knowledge existing models curiously neglect an important insight from earlier research. Intergroup attitudes are at the same time themselves subject to social influence and are shaped by opinion dynamics. Empirical research on effects of direct and extended intergroup contact [52, 55, 69] has established how attitudes about groups are shaped by interactions both between members of the same group and members of different groups, via direct or indirect experiences with group members, or via social influence [67]. As this paper will argue, incorporating socially influenced intergroup attitudes into formal models of opinion dynamics can crucially change model predictions, thus giving us new insights into the robustness of results obtained from previous modeling research, and it improves the alignment between formal models and evidence from substantive studies of intergroup dynamics.

If intergroup attitudes not only affect social influence between individuals but also simultaneously are shaped by social influence processes, very different theoretical implications can arise about the role that intergroup attitudes have for societal polarization. For example, if xenophobia and its counterpart of ingroup favoritism [61] are a hard-wired element in social influence, models have an inherent tendency to generate opinion polarization that divides a population along the lines of its salient demographic divisions. However, if xenophobia can be unlearned, for example when individuals in different groups discover sufficient ground for agreement and can form mutually positive relations, consensus in a society could be more robust against increasing demographic diversity than suggested by models with hardwired xenophobia. Also, as research on intergroup attitudes has pointed out, people can sometimes have negative attitudes towards their own group and prefer outgroups above their ingroup, for example if their ingroup has low social status in their society, or is perceived to violate important personal norms [6].

The presence of such 'renegades' could profoundly change polarization dynamics when intergroup attitudes are open to influence. Renegades can influence their fellow ingroup members to adopt a more critical perspective on the own group and think more positively about outgroups. This might prevent intergroup polarization where it would develop otherwise if intergroup attitudes were fixed and xenophobic. But the opposite is also possible. A small critical mass of 'outgroup haters' might convince a moderate majority within their own group to become more critical of the outgroup, fueling in the process increasing disagreement between groups that eventually gives rise to polarization between groups.

To explore these possibilities, a simple model is proposed that extends previous models assuming fixed intergroup attitudes. The opinion variable in the model is the intergroup attitude about two groups (0 and 1), where higher values represent a more positive attitude towards group 1 as compared to group 0. I analyze with computational experiments how conditions and dynamics of opinion polarization change if openness to outgroup influence depends on this intergroup attitude, rather than being static and hardwired. A description of the model is given in Sec. 2. Section 3 presents results of simulation experiments. The paper concludes with a discussion of the substantive interpretation of results and possible directions for future research in Sec. 4.

2. Model

The model presented here draws upon and extends earlier models that combine assimilative and repulsive social influence in opinion formation. A recent overview on these and related models in the agent-based modelling literature can be found in [26]. To concentrate on the effects of introducing socially influenced intergroup attitudes, other aspects of the model are chosen to be maximally simple. To begin with, the population of N individuals consists of only two distinct groups and group membership is stable over time. Every individual i is member of either group 0 or group 1,

indicated by its group membership $g_i \in \{0, 1\}$. Furthermore, individuals influence each other with regard to only one opinion dimension o.

The key difference to earlier models is that this opinion represents the intergroup attitude, which in turn affects how individuals are influenced by ingroup members and outgroup members. More precisely, the opinion value o_{it} expresses the attitude an individual i has at time point t about group 1 and $1 - o_{it}$ is the attitude towards group 0. Higher values indicate a more positive attitude towards group 1 and a less positive attitude towards group 0. Following most earlier models in the literature, I constrain the value range of opinions, assuming $0 \leq o_{it} \leq 1$. If $o_{it} < 0.5$, individual i has a more positive attitude towards group 0 than towards group 1, while $o_{it} > 0.5$ indicates the opposite relation. A value of $o_{it} = 0.5$ expresses indifference between the two groups. It should be noted that for members of group 1, the opinion o models their evaluation of the ingroup relative to the outgroup, whereas for members of group 0, it captures the evaluation of the outgroup relative to the ingroup.

Dynamics of the model unfold in a sequence of consecutive interaction events, in which in every event a pair of two different population members i and j is selected with equal probability from all possible pairs in the population for an interaction. Notice that this imposes the simplification that interactions are not constrained by network structure or spatial distances. This interaction regime draws upon the pairwise interaction mechanism introduced by Deffuant and co-authors [13].

All individuals k who are not involved in an interaction at time point t do not change their opinions, thus $o_{k,t+1} = o_{kt}$. If i and j do interact, then both can modify their current opinions to move closer towards or away from the opinion of the interaction partner as given by Eqs. (1) and (2).

$$o_{i,t+1} = o_{it} + \Delta o_{it} = o_{it} + \mu w_{ijt}\ (o_{jt} - o_{it}), \tag{1}$$

$$o_{j,t+1} = o_{jt} + \Delta o_{jt} = o_{jt} + \mu w_{jit}\ (o_{it} - o_{jt}). \tag{2}$$

Following [13], the model includes a parameter μ $(0 < \mu \leq 0.5)$ in Eqs. (1) and (2) that defines the rate of opinion change and will be kept at $\mu = 0.5$ throughout in the present paper. The influence weights w_{ijt} and w_{jit} in Eqs. (1) and (2) represent the direction and magnitude of the influence of i on j and j on i, respectively. Weights are constrained by $-1 \leq w_{ij} \leq 1$. A positive weight w_{km} entails assimilative influence (k moving her opinion closer towards m's opinion), whereas a negative weight imposes differentiation (k moving her opinion away from m's opinion). With $w_{km} = 0$, k does not change her opinion, reflecting indifference towards the source of influence, m. In this basic form, Eqs. (1) and (2) allow interactions to push the opinion outside of the opinion interval $[0, 1]$ if weights are negative. In this case, the resulting opinion is truncated to the interval boundary that was crossed by the opinion shift. In some models that combine assimilation and differentiation, opinions are constrained with more sophisticated approaches (e.g. [23, 24, 34]), but this seems to have little effect on the main model dynamics.

The link between intergroup attitudes and social influence is implemented in the way how weights are computed. I adopt a two-step process. In the first step, the discrepancy d_{ijt} is computed that individual i experiences at time point t between herself and individual j. Discrepancy is based on the current level of disagreement $|o_{jt} - o_{it}|$, whether i and j belong to the same group, $|g_j - g_i|$, and on the attitude that i currently holds towards the group to which j belongs. Discrepancy is constrained by $0 \leq d_{ijt} \leq 1$. In the second step, the actual influence weight w_{ijt} is obtained from a weight function that takes the discrepancy d_{ijt} as argument. This function is assumed to be monotonous but not necessarily linear. Broadly, the larger the discrepancy, the lower the influence weight. More precisely, discrepancy is computed as given by Eq. (3).

$$d_{ijt} = \beta_O|o_{jt} - o_{it}| + \beta_D|g_j - g_i| + \beta_A[g_j(1 - o_{it}) + (1 - g_j)o_{it}]. \tag{3}$$

The parameters $\beta_O, \beta_D, \beta_A$ in Eq. (3) scale the relative impact that respectively, *opinion* disagreement, *demographic* differences and the intergroup *attitude* towards j's group have on d_{ijt}. I impose the constraint $\beta_O + \beta_D + \beta_A = 1$. If intergroup attitudes play no role ($\beta_A = 0$), this implementation is compatible with earlier models that incorporate only opinion disagreement and fixed xenophobia in otherwise the same framework (e.g. [20, 24, 25]). The new element of intergroup attitudes is added such that a more positive attitude o_{it} of i towards group 1 means less discrepancy if the interaction partner j belongs to group 1, whereas it means more discrepancy if j belongs to group 0. Conversely, if j belongs to group 0, then a higher value of o_{it} implies more discrepancy and a lower value means less discrepancy.

The actual influence weight w_{ijt} is computed from the discrepancy d_{ijt} as $w_{ijt} = f(d_{ijt})$, with a nonlinear weight function f adapted from [46] as given by Eq. (4). This weight function implements a nonlinear relation between discrepancy and the impact that a source of influence j has on i's opinion. Following earlier implementations (e.g. [36]) this function has been chosen to model that if the level of discrepancy is such that the opposing psychological motivations towards assimilation and differentiation are roughly in balance, changes in discrepancy have relatively little effect on the weight. More technically, the function f yields a weight that is positive if discrepancy falls below a threshold level of 0.5, the midpoint of the range of possible discrepancy levels. The weight turns negative once discrepancy exceeds this threshold. Changes in discrepancy close to the threshold level have only a relatively small impact on the weight such that the influence weight is close to zero in this region. The weight changes more steeply in discrepancy as d approaches the boundaries of its theoretically possible range of $[0, 1]$.

$$f(d) = \begin{cases} (1 - 2d)^2, & d < \dfrac{1}{2}, \\ -(2d - 1)^2, & d \geq \dfrac{1}{2}. \end{cases} \tag{4}$$

3. Results

3.1. *Design and measures*

I present here results from two computational experiments. The first experiment explores whether and if so how conditions and dynamics of polarization change, if we model intergroup attitudes as socially influenced rather than given by 'fixed xenophobia'. The main comparison is between the impact of an intergroup attitude that is fixed (gradually increasing β_D, letting $\beta_A = 0$) as opposed to a socially influenced intergroup attitude (gradually increasing β_A, letting $\beta_D = 0$). To preview, Experiment 1 will reveal that even a relatively small proportion of initial renegades can profoundly alter opinion dynamics if intergroup attitudes are socially influenced. In the second experiment, it is therefore tested how fixed versus socially influenced intergroup attitudes affect outcomes if there is a considerably higher proportion of initial renegades than in Experiment 1.

The results presented in this paper are the first step of a fuller exploration of the behavior of the model proposed here. Thus, I keep a number of conditions constant across all the experiments reported below. To begin with, population size is assumed to be constant throughout and fixed at $N = 100$, where groups 0 and 1 have equal size ($N_0 = N_1 = 50$). The rate of opinion change is fixed at $\mu = 0.5$. Further constant conditions were discussed above and include that there is only one opinion dimension o, and that interaction is equally likely for every pair of population members.

The key outcome of interest in the simulation experiments is the degree of polarization both within and between the two groups. To assess between-group polarization, I measure the absolute value of the difference between the mean opinions in both groups, $|\overline{o_{g=1}} - \overline{o_{g=0}}|$. If this difference is close to one, this is a clear sign of strong between-group polarization. A low difference between the mean opinions of the groups, however, does not necessarily show that there is no polarization in the population as a whole. The population can also fall apart into two opposed factions containing members of both groups. To distinguish this form of 'population polarization' from between-group polarization, I measure population polarization P_t at time point t as the variance of all pairwise opinion distances in the population (adapted from [23]), as given by Eq. (5).

$$P_t = \frac{4}{N^2} \sum_{ij}^{i=N,j=N} (|o_{jt} - o_{it}| - \overline{|o_{kt} - o_{mt}|})^2. \tag{5}$$

In Eq. (5), $\overline{|o_{kt} - o_{mt}|}$ denotes the average opinion distance across all pairs (km) in the population. The minimum level of polarization ($P = 0$) obtains when all pairwise distances are zero, corresponding to full consensus in the population.[a] P obtains its maximal value of 1 if the population is split into two equally large factions with

[a] Other than in [23], the set of pairs includes here also the self-distances in the pairs (i, i). While these are known to be zero by definition, including self-distances simplifies definition and interpretation of the measure.

maximal mutual disagreement and full agreement within each of the factions. It is also useful to assess to which extent polarization occurs within each of the two groups separately. For this, the measure P_t is computed separately for each of the two groups ($P_{t,g=0}, P_{t,g=1}$).

3.2. *Simulation Experiment* 1: *The relative weight of intergroup attitudes in social influence*

Previous modeling work has shown how fixed xenophobia can increase opinion polarization between groups, particularly when the demographic differences are clear and salient (e.g. [20, 24, 25]). This suggests that the higher the relative weight β_D of belonging to different subgroups is in computing the discrepancy between individuals, the more likely attitudes in the population polarize between the two subgroups. In Experiment 1, it will be assessed whether this result changes if intergroup attitudes are themselves subject to social influence. I start from an initial opinion distribution that imposes a mild ingroup bias in both groups. This reflects the notion of ingroup favoritism. Technically, the initial opinions in both groups are randomly drawn from two Beta distributions that are symmetric around the midpoint of the opinion interval, with $Beta(7.5, 10)$ for group 0 and $Beta(10, 7.5)$ for group 1. The mean opinion in this initial distribution is about 0.43 for group 0 and about 0.57 for group 1. Initial opinions in both groups have the same expected standard deviation of approximately 0.12. Figure 1 visualizes the corresponding probability density functions together with the aggregate distribution for the population as a whole.

In Experiment 1, the relative impact of the intergroup attitude on discrepancy is systematically varied across the entire range of possible values. Two conditions are distinguished. First, in the 'fixed xenophobia' condition, the parameter that is manipulated is β_D, keeping $\beta_A = 0$ and $\beta_D + \beta_O = 1$. In the 'socially influenced intergroup attitude' condition, the parameter manipulated is β_A, keeping $\beta_D = 0$, under the constraint $\beta_A + \beta_O = 1$. In both conditions, the manipulated parameter is

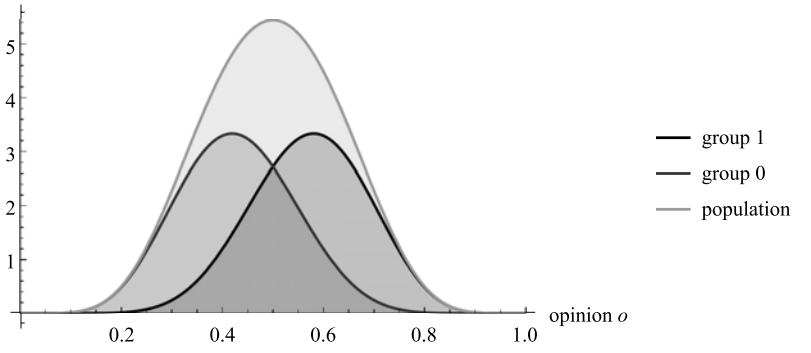


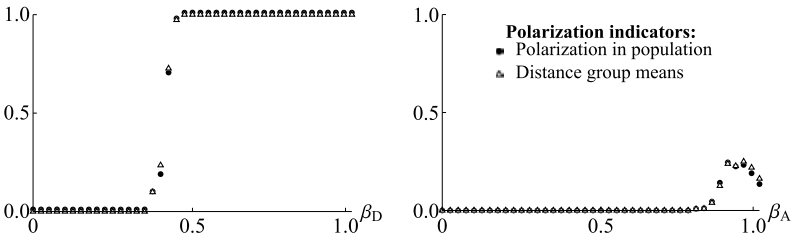
Fig. 1. Initial opinion distribution in Experiment 1 (group 0: Beta(7.5,10), group 1: Beta(10,7.5)).

varied from 0 to 1 in steps of 0.025, yielding 41 different levels of impact of intergroup attitudes. At one end of the spectrum, discrepancy between two individuals i and j depends exclusively on the current level of disagreement in the attitude towards group 1, $|o_{jt} - o_{it}|$. In the 'fixed xenophobia condition' this occurs for $\beta_D = 0$, in the 'socially influenced intergroup attitude' condition, this happens for $\beta_A = 0$. At the other end of the spectrum, with $\beta_D = 1$ or $\beta_A = 1$, disagreement is irrelevant for the discrepancy between i and j. Here, only demographic difference (fixed xenophobia) or i's attitude towards j's group (socially influenced intergroup attitude) matter for the direction and magnitude of influence of j on i.

3.2.1. *Experiment* 1: *Fixed xenophobia condition* ($\boldsymbol{\beta}_A = 0$)

The left part of Fig. 2 shows the effect of β_D on two indicators of between-group polarization in the fixed xenophobia condition ($\beta_A = 0$). The two indicators are the absolute distance of group means ($|\overline{o_{g=1}} - \overline{o_{g=0}}|$) and the indicator for polarization in the population (P_t). The results shown are averages of the outcomes of 100 independent realizations of the simulation per level of β_D, measured after 30.000 simulation events. A small offset has been added to the indicator of population polarization to avoid overlap of points with identical values.

Figure 2(a) shows how the population shifts from a consensus regime into a regime with maximal between-group polarization when fixed xenophobia exceeds a critical threshold level of approximately $\beta_D = 0.4$. The curve exhibits a sharp increase in the average measures of the distance between group mean opinions ($|\overline{o_{g=1}} - \overline{o_{g=0}}|$) and of the level of polarization in the population, P_t, when β_D passes the critical level. Below that point, both outcome indicators yield zero, indicating that after 30.000 iterations all population members have adopted the same intergroup attitudes and there are no discernible differences between the two groups. Despite an initial distance between the group means of about 0.145, the initial disagreement within and between groups is not sufficiently large in this region of the



(a) Fixed xenophobia condition ($\beta_A = 0$)

(b) Condition with socially influenced intergroup attitude ($\beta_D = 0$)

Fig. 2. Results of Experiment 1. Effect of weight of intergroup attitude β_D or β_A, respectively, on polarization indicators after 30.000 simulation events, averaged across 100 realizations per level of β_D (left) or β_A (right).

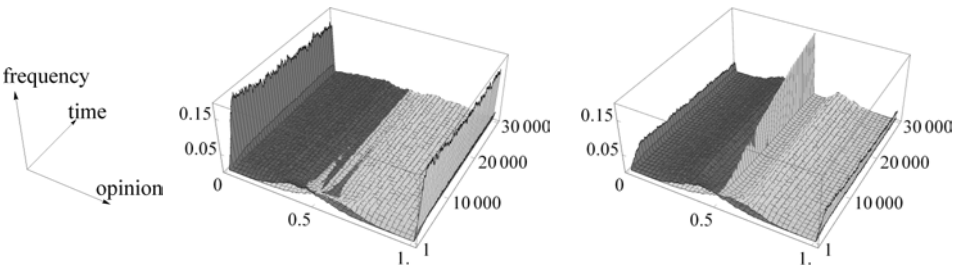
parameter space to trigger repulsive influence dynamics that could counter the pressures towards assimilation. Eventually all individuals converge to an opinion close to the initial population mean value of 0.5. However, when β_D exceeds the critical level, only little disagreement is needed to give individuals a negative influence weight w towards each other if they do not belong to the same group. For example, for $\beta_D = 0.4$, repulsive influence is triggered between individuals from different groups if their disagreement exceeds a level of approximately 0.15. By contrast, only a disagreement of 0.8 or more can elicit repulsive influence between members of the same group. In this region of the parameter space, the initial between-group differences are large enough to trigger repulsive influence dynamics between members of different groups, pushing their attitudes increasingly away from those of the outgroup towards the extreme poles of $o = 0$ and $o = 1$ for groups 0 and 1, respectively. Assimilation within the groups assures that all group members eventually agree with these extremely negative outgroup attitudes. The close match of the two polarization indicators in Fig. 2(a) further confirms that all polarization is between the groups and not within groups. Both $P_{t,g=0}$ and $P_{t,g=1}$ are approximately zero after 30.000 simulation events.

3.2.2. *Experiment* 1: *Socially influenced intergroup attitude condition* ($\boldsymbol{\beta}_D = 0$)

Figure 2(b) shows results for the condition in which intergroup attitudes are socially influenced ($\beta_D = 0$). It visualizes the effect of β_A on the same average indicators of between-group polarization and their distribution across 100 independent realizations, measured after 30.000 simulation events. A comparison of Figs. 2(a) and 2(b) reveals clear differences. One commonality between both conditions of Experiment 1 is that opinion dynamics result in consensus if the impact of intergroup attitudes falls below a critical level. However, for β_A a level of at least 0.8 must be reached before the population leaves the consensus regime, as opposed to about 0.4 for β_D. Moreover, the degree of intergroup polarization is much lower than for fixed xenophobia, even when β_A approaches its maximum level of 1.0. Further inspection of the results reveals that in this region none of the runs with socially influenced intergroup attitudes has generated perfect between-group polarization. Finally, we observe that the effect of β_A on polarization is not monotonous. Figure 2(b) shows how both polarization indicators peak at approximately $\beta_A = 0.9$, at a level around 0.25, and then decline down to about 0.2 when β_A increases further.

3.2.3. *Experiment* 1: *Explanation of differences between conditions*

Why is the effect of β_A in the condition with socially influenced intergroup attitudes so different from the effect of β_D in the fixed xenophobia condition? Below the critical level of approximately $\beta_A = 0.8$, opinion dynamics are dominated by assimilation. As we see, much higher levels of β_A than of β_D are required to leave the consensus

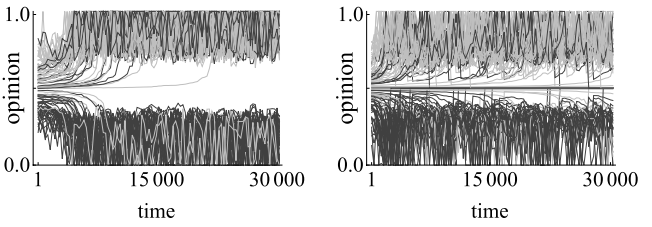


(a) $\beta_A = 0.9$ (b) $\beta_A = 1$

Fig. 3. Experiment 1: Change of aggregated opinion distributions over 100 independent realizations for two levels of β_A in the condition with socially influenced intergroup attitudes ($\beta_D = 0$). Dark: group 0, Light: group 1.

regime. This follows directly from the discrepancy function in Eq. (3). If $\beta_A = \beta_D$, then the 'xenophobia term' $\beta_D|g_j - g_i|$ always adds more to discrepancy than the 'attitude term' $\beta_A[g_j(1 - o_{it}) + (1 - g_j)o_{it}]$ does, unless i's attitude towards the outgroup is maximally negative. Given that intergroup attitudes are only mildly biased at the outset, this explains the narrower range of levels of β_A in which the population escapes from the pull of consensus. But this does not yet explain the nonmonotonous effect that we observe for β_A. To shed more light on the different dynamics of intergroup polarization with different values of β_A (keeping $\beta_D = 0$), Fig. 3 describes the overall tendencies in these dynamics. The figure presents the average frequencies of the opinion distributions that evolved across 100 independent realizations of the experiment in the two conditions where $\beta_A = 0.9$ and $\beta_A = 1$, respectively (holding $\beta_D = 0$).

Figure 3 shows that there is some extent of polarization between groups both for $\beta_A = 0.9$ and for $\beta_A = 1$. In both conditions the emergent aggregated opinion distributions show peaks at both extreme poles. While these extremal peaks are much more pronounced for $\beta_A = 0.9$, the peak around the neutral attitude 0.5 is clearly more pronounced for $\beta_A = 1$. This explains why the average level of polarization drops between these two parameter values in Fig. 2(b). At the same time, in both conditions members of both groups adopt positions across the entire spectrum, including extreme ingroup lovers, ingroup haters and moderates. Especially for $\beta_A = 1$ smaller peaks arise within both groups at attitudes that are moderately in favor of the ingroup. Inspection of the dynamics suggests that for individuals on these positions, different opposing pressures to modify their attitudes tend to cancel each other out, such that they can retain their moderate pro-ingroup position at least temporarily. The pressures balancing each other are the opposing attractions towards maximally extreme ingroup lovers and moderates within their own group, and the push towards differentiation from moderate and extreme outgroup members. Overall, in line with the moderate ingroup bias that was imposed in the initial distribution, members of both groups remain on average slightly more positive about



(a) $\beta_A = 0.9$ (b) $\beta_A = 1$

Fig. 4. Experiment 1: Single runs with socially influenced intergroup attitudes. Dark: group 0, Light: group 1.

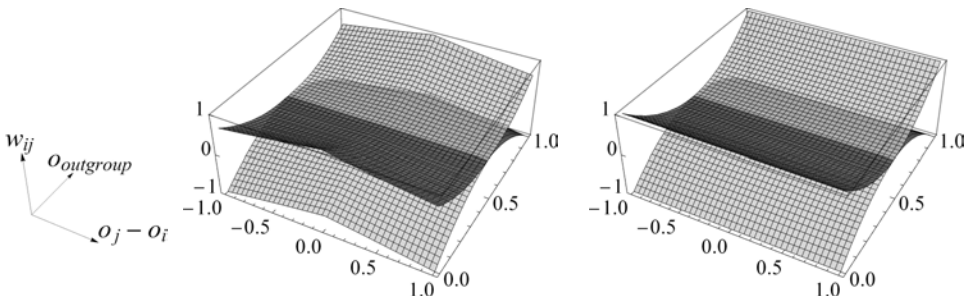
their ingroup than about the outgroup and occur more frequently at the pole of extreme ingroup love than at the pole of maximal ingroup hate.

Inspection of the trajectories of single individuals in single simulation runs gives more insight into why some individuals turn into ingroup lovers and others into ingroup haters. Figure 4 shows the evolution of the opinions of all individuals in the population in a specific run. The left part of the figure shows how in both groups individuals who initially have a more positive ingroup attitude ($o > 0.5$ for group 1, $o < 0.5$ for group 0) are likely to move towards the pole of a maximally positive ingroup attitude, while those who initially are more critical ($o < 0.5$ for group 1, $o > 0.5$ for group 0) tend to move in the opposite direction. With β_A close to or equal to 1.0 and $\beta_D = 0$, the opinion movement of a particular individual is primarily dominated by her intergroup attitude. Individuals who have a more favorable attitude of their ingroup ($o > 0.5$ for group 1, $o < 0.5$ for group 0) are moving towards the opinion of an ingroup member if they interact with one, and move away from the opinion of outgroup members they encounter. Given that in the initial distribution the average attitude of group 1 is above 0.5 while the average attitude of group 0 is below 0.5, this explains why most group 1 members move upwards and most group 0 members move downwards in the dynamics shown by Fig. 4(a). As these individuals feel even more positively about their ingroup in the process, their initial tendency becomes self-reinforcing, pushing them towards the boundaries of maximal ingroup love combined with maximal outgroup hate. Conversely, the minority of initial "renegades" in both groups ($o < 0.5$ for group 1, $o > 0.5$ for group 0) is on average attracted towards the mean opinion of the outgroup and shifts away from the attitudes of their ingroup members. These individuals become increasingly critical of their ingroup and tend to move towards the opposite extreme.

But why do the dynamics not settle into an equilibrium state within the time frame we observe? Individuals who strongly like their ingroup and thus dislike the outgroup, are not happy with having the same attitude than 'renegade' outgroup members have who agree with them. At the same time, these renegade outgroup members 'chase' them in the attitude space because they like to agree with members of the outgroup they love. Once an ingroup lover of one group interacts with an

ingroup hater of the other group, the ingroup lover has a negative influence weight towards the member of the other group and will therefore — paradoxically — be pushed to shift away from a prior positive attitude towards his own group. However, given that ingroup lovers occur more frequently on both sides than ingroup haters do, it is likely that the ingroup lover will be pulled back sooner or later after an interaction with another ingroup member. Figure 4 shows that these destabilizing dynamics occur for both $\beta_A = 0.9$ and $\beta_A = 1$, given $\beta_D = 0$. However, on average many more individuals adopt a moderate attitude at any point in time for $\beta_A = 1$ and many more extremist attitudes occur for $\beta_A = 0.9$. How this difference comes about can be further seen from inspecting the influence weight functions towards in- and outgroup members in both conditions. Figure 5 depicts the influence weight as a function of the intergroup attitude and of the opinion disagreement between two individuals who interact, separately for $\beta_A = 0.9$ and $\beta_A = 1$, with $\beta_D = 0$ in both cases.

Figure 5 explains why there is such a strong tendency for individuals to adopt a moderate intergroup attitude if $\beta_A = 1$. If $\beta_A = 1$ and actor i has a neutral attitude ($o = 0.5$) then the influence weight w_{ij} is zero towards both ingroup and outgroup members, regardless of the level of disagreement between i and j. Thus, once random movements have led an individual into the position $o = 0.5$, this individual will no longer change her attitude. This is different if $\beta_A = 0.9$. Here, disagreement still affects the influence weight. Although with a moderate intergroup attitude of $o = 0.5$, individuals have the same influence weight towards members of both groups, a neutral influence weight of $w = 0$ can only obtain if the level of disagreement between i and j is exactly 0.5 for both ingroup and outgroup members j. It is impossible to obtain a configuration in which more than one individual in the population can at the same time disagree by 0.5 with both all ingroup and all outgroup members, so that for $\beta_A = 0.9$ we do not observe the emergence of a stable attitude of $o = 0.5$. By contrast, $o = 0.5$ is always an equilibrium attitude for i if $\beta_A = 1$. An analytical proof for this claim can be found in the appendix.



(a) $\beta_A = 0.9$, $\beta_D = 0$ (b) $\beta_A = 1$, $\beta_D = 0$

Fig. 5. Weight function of group 0 member with socially influenced intergroup attitudes. Dark: weight towards ingroup, light: weight towards outgroup.

To sum up, Experiment 1 has shown distinctive differences between fixed xenophobia ($\beta_A = 0$) and socially influenced intergroup attitudes ($\beta_D = 0$). Overall, the model with socially influenced intergroup attitudes generates between-group polarization in a much narrower range of conditions. For those conditions where some intergroup polarization emerges, it does so to a much smaller extent than under the model with fixed xenophobia. Moreover, the dynamics and opinion distributions that emerge if there is no consensus (approximately for $\beta_A > 0.8$), are more complex and richer than those we find for higher levels of fixed xenophobia. With fixed xenophobia, the outcome is an equilibrium state with clear-cut between-group polarization above a level of approximately $\beta_D = 0.4$. Under socially influenced intergroup attitudes, we observe instead what could be called 'mixed instable attitude polarization', with co-existence of extreme and moderate opinions within both groups. Especially, both groups can contain a minority of 'renegades' with negative opinions about their ingroup in these conditions and, moreover, dynamics never settled down to an equilibrium state[b] in the time period simulated in Experiment 1. A further robustness test showed that the qualitative differences between the two conditions persist but become much smaller, if from the outset there are less renegades in both groups.

3.3. *Experiment* 2: *Renegades and reversed intergroup polarization*

Experiment 1 showed that the presence of renegades in the initial distribution can considerably change opinion dynamics when intergroup attitudes are socially influenced rather than fixed. Therefore, Experiment 2 focuses on the role of renegades, testing what happens if the initial opinion distributions show a much larger variance than in Experiment 1, so that the proportion of renegades is considerably higher in both groups.

Figure 6 shows the Beta distributions used in Experiment 2. With about 0.067, the initial distance in group means is approximately half as large as in Experiment 1, but the standard deviation within both groups is about twice as large (0.229 in Experiment 2 versus 0.115 in Experiment 1).

In Experiment 2, I want to assess whether a larger share of renegades can prevent between-group polarization under conditions for which it is likely to occur otherwise. As a conservative test, socially influenced intergroup attitudes will in Experiment 2 be gradually replaced by fixed xenophobia, rendering it increasingly difficult to prevent between-group polarization. That is, starting from a baseline in which social influence is exclusively moderated by intergroup attitudes ($\beta_A = 1$), the relative weight of fixed xenophobia will be increased from $\beta_D = 0$ to $\beta_D = 0.2$, and finally $\beta_D = 0.4$. Furthermore, it is assumed throughout Experiment 2 that the (low) disagreement in attitudes between members from different groups has no impact on the

[b] In fact, for the conditions of Experiment 1 shown in Figs. 3(b) and 4(b), I found in further simulation runs that dynamics failed to converge within 1 billion (10^9) simulation events, while equilibrium was in some cases reached within about 100 million simulation events for $\beta_A = 0.92$ in the condition with socially influenced intergroup attitudes.

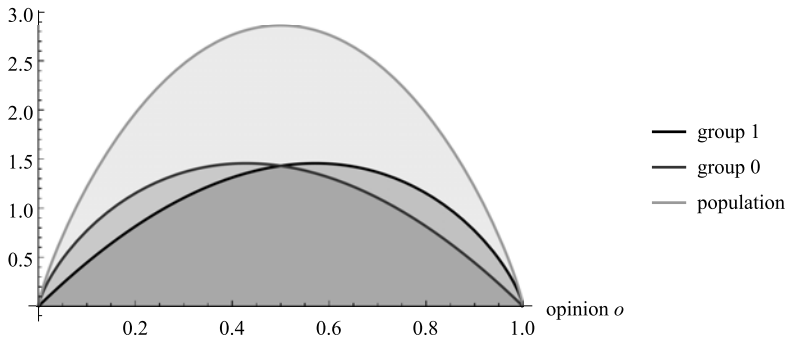


Fig. 6. Initial opinion distribution in Experiment 1 (group 1: Beta(2,1.75), group 0: Beta(1.75,2)).

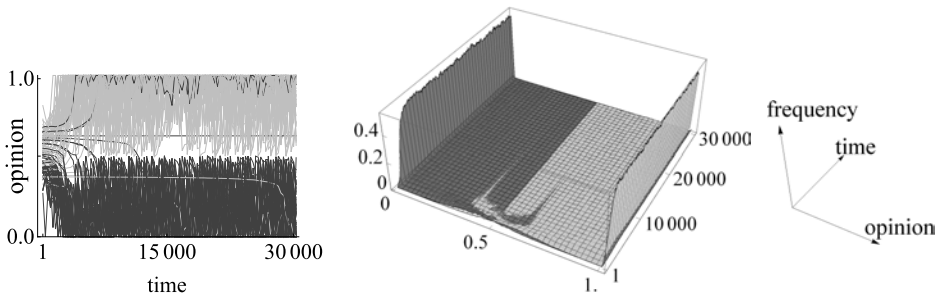
discrepancy individuals perceive, in order to prevent that the relatively small initial attitude difference between the groups makes consensus too easy to reach ($\beta_O = 0$, $\beta_D + \beta_A = 1$).

Notice that with this approach it is assumed that individuals can simultaneously experience a fixed degree of outgroup aversion and at the same time be socially influenced to develop a positive attitude towards the outgroup. This captures the idea that although negative views of outgroups can be 'unlearned', there are also simultaneous psychological processes and social-structural reasons (like speaking similar languages or sharing familiar social norms) that let us be more attracted to more similar people, all other things being equal (e.g. [9]). The relative values of β_D and β_A can be seen as modeling how deeply entrenched in a given social setting are conditions that sustain xenophobia relative to social influence towards more positive intergroup attitudes.

3.3.1. *Experiment* 2: *Baseline with small proportion initial renegades*

First, I consider as baseline a condition with a low share of renegades, to then show how the presence of renegades changes dynamics. The baseline is given by the same initial distribution than for Experiment 1. The dynamics for $\beta_D = 0$, $\beta_A = 1$ in this situation are those shown in Fig. 3(b), where we observe a state in which moderate attitudes dominate. Figure 7 shows how these dynamics change considerably if the impact of fixed xenophobia is slightly increased to $\beta_D = 0.2$, $\beta_A = 0.8$.

This small increase from $\beta_D = 0$ to $\beta_D = 0.2$ suffices to push the population into a state with a high level of between-group polarization. In both groups a share of about 80% of the group members move on average to the attitude of maximal ingroup love and outgroup hate, with an average distance between the group means of about 0.69. As the individual trajectories of a single run shown in Fig. 7(a) illustrate, between-group polarization is not perfect because a small fraction of renegades who moved to the other side cause perpetual movement of ingroup lovers in and out of the state of maximal outgroup hate. While this outcome could still be seen as moderate



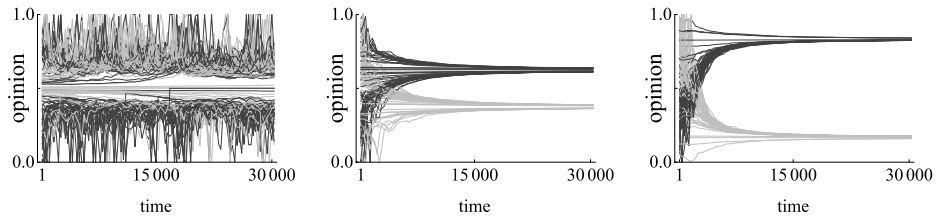
(a) Single run (b) Distribution opinions

Fig. 7. Baseline condition Experiment 2, $\beta_A = 0.8$. Dark: group 0, light: group 1.

between-group polarization, a further increase of the weight of fixed xenophobia to $\beta_D = 0.4$ pushes the population into a state of nearly maximal between-group polarization (average $|o_{g=1} - o_{g=0}| \approx 1$).

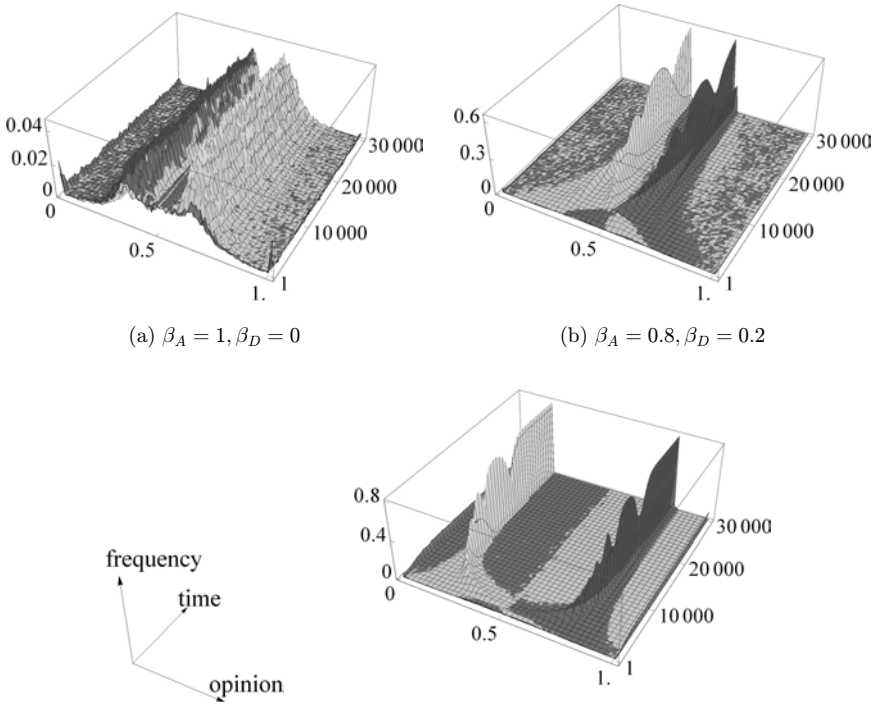
3.3.2. *Experiment* 2: *Large proportion initial renegades*

Next, the baseline condition is compared to the scenario with a larger initial share of renegades (see Fig. 6). Figure 8 shows individual trajectories of typical single runs for $\beta_D = 0$, $\beta_D = 0.2$ and $\beta_D = 0.4$, respectively ($\beta_O = 0, \beta_D + \beta_A = 1$). Figure 9 depicts the corresponding change of the opinion distribution averaged over 100 runs per condition. The results reveal that the larger share of initial renegades fundamentally changes how fixed xenophobia affects the dynamics of between-group polarization. The pattern we observe for $\beta_D = 0.2$ and $\beta_D = 0.4$ can be characterized as 'reversed intergroup polarization'. Members of the two groups end up disagreeing like they do with 'regular' between-group polarization, but this time both like their outgroup and dislike their ingroup. This can also be seen from comparison of the absolute difference between-group means $|o_{g=1} - o_{g=0}|$, with the signed difference $o_{g=1} - o_{g=0}$. While the latter is negative for $\beta_D = 0.2$ and $\beta_D = 0.4$, the former is positive and increasing as β_D increases from 0 upwards.



(a) $\beta_D = 0, \beta_A = 1$ (b) $\beta_D = 0.2, \beta_A = 0.8$ (c) $\beta_D = 0.4, \beta_A = 0.6$

Fig. 8. Experiment 2: Individual trajectories with large share of initial renegades. Dark: group 0, light: group 1.



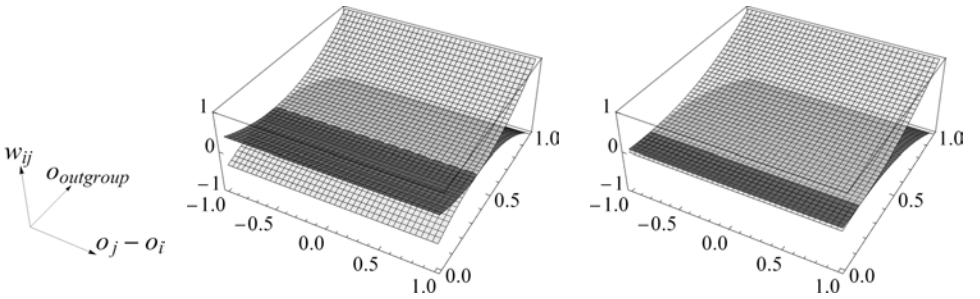
(a) $\beta_A = 1, \beta_D = 0$

(b) $\beta_A = 0.8, \beta_D = 0.2$

(c) $\beta_A = 0.6, \beta_D = 0.4$

Fig. 9. Experiment 2: Change of opinion distributions averaged over 100 runs under large share of initial renegades. Dark: group 0, light: group 1.

Higher levels of fixed xenophobia increased the extent of ingroup love and outgroup aversion in both groups when the initial share of renegades was small (Experiment 1). Figures 8 and 9 show how in Experiment 2 this effect is reversed by a high share of initial renegades. Now, from $\beta_D = 0.2$ on, in both groups a consensus develops on 'universal outgroup love' ($o < 0.5$ for group 1, $o > 0.5$ for group 0). Paradoxically, the higher the weight of fixed xenophobia, the more positive the attitude towards the outgroup becomes (and the more negative the attitude towards the ingroup), on which all members of the same group eventually agree. The individual trajectories depicted in Fig. 8 illustrate how this happens. For $\beta_D = 0.2$ and $\beta_D = 0.4$, initial renegades within both groups change their attitudes much less than initial ingroup lovers do. The reason is that renegades have a negative attitude towards their own group and are therefore not open for assimilative influence from fellow ingroup members. At the same time, ingroup lovers wish to assimilate towards the attitudes of their fellow ingroup members, including the renegades. Thus, the influence relation between ingroup lovers and renegades is asymmetrical. While renegades stay their course, ingroup lovers adopt the negative view of the renegades about their own ingroup, because they wish to agree with their fellow ingroup members.



(a) $\beta_D = 0.2, \beta_A = 0.8$ (b) $\beta_D = 0.4, \beta_A = 0.6$

Fig. 10. Weight function of group 0 member with socially influenced intergroup attitudes. Dark: weight towards ingroup, light: weight towards outgroup.

This counter-intuitive dynamic is further strengthened by renegades' attraction towards the attitudes of outgroup members, and ingroup lovers' desire to differentiate from outgroup members. This explains the initial shift of extreme renegades towards slightly less negative views of their ingroup that we can observe for $\beta_D = 0.2$ and $\beta_D = 0.4$ in Fig. 8. The reason is not assimilation towards the ingroup, but it is assimilation towards the majority of the outgroup members who simultaneously develop an increasingly critical attitude towards their own group.

The position of the equilibrium attitudes to which individuals settle in Experiment 2 can be best understood by an analysis of the weight functions f imposed by the values for $\beta_O, \beta_D, \beta_A$ that we use in the experiment. Figure 10 charts the weight functions for the three conditions of Experiment 2.

Figure 10 makes clear that in the conditions shown there is a value for the intergroup attitude o at which the influence weights of both ingroup members and outgroup members are zero. This is the level of o at which the two surfaces in the figures intersect. If all individuals within a group adopt this value of o, they are no longer susceptible to outside influences and have reached an equilibrium state. Figure 10 shows how increasing β_D pushes this equilibrium intergroup attitude towards an increasingly negative opinion about the ingroup. Substantively, this can be interpreted as the interplay of two counterbalancing forces. Fixed xenophobia leads an individual to differentiate from outgroup members. However, if the same individuals hold a favorable socially influenced attitude towards the outgroup, this can neutralize the force towards differentiation and thus result in an equilibrium situation. The stronger fixed xenophobia, the more positive the attitude towards the outgroup needs to be to compensate for the effect of fixed xenophobia. Conversely, fixed xenophobia makes an individual more open to assimilate towards the opinion of an ingroup member (by helping to reduce discrepancy to below the critical level needed for a positive influence weight) but an unfavorable attitude towards the ingroup can neutralize this assimilative force with an equally strong repulsive

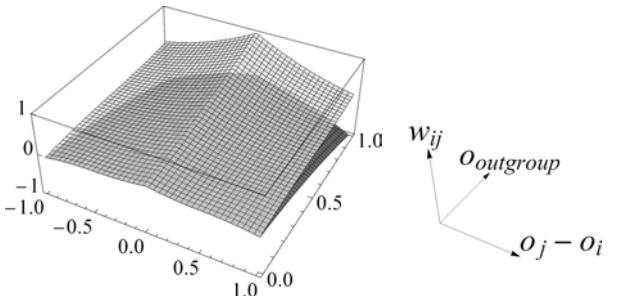
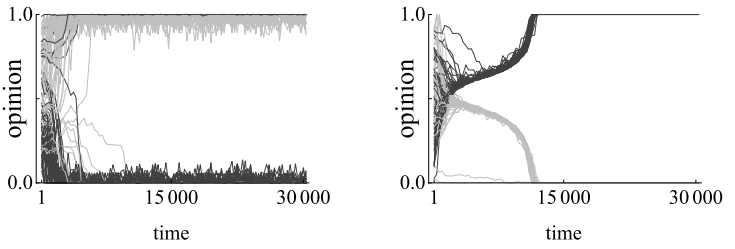


Fig. 11. Weight function of group 0 member with socially influenced intergroup attitudes. Dark: weight towards ingroup, light: weight towards outgroup $\beta_O = \beta_D = \beta_A = \frac{1}{3}$.

pressure. The appendix shows analytically how higher β_D shifts the equilibrium intergroup attitude downward in this condition.

The stronger the force of fixed xenophobia, the more favorable the attitude towards the outgroup needs to be to obtain an intergroup attitude in equilibrium. The limiting case is $\beta_A = \beta_O$. When fixed xenophobia and socially influenced intergroup attitude have an equally strong impact on perceived discrepancy, the only stable attitude is a maximally negative view of the ingroup. This holds also when there is an additional impact of disagreement ($\beta_O > 0$) as demonstrated by Fig. 11 for the case where all three sources of discrepancy have the same impact on the influence weight w_{ij}. The figure shows that the influence weight of an ingroup member always exceeds the weight of an outgroup member except for the limiting case of $o = 0$ in which both weights are equal for the same level of disagreement. Notice that not both weights are always zero in this case, this only happens if in addition the disagreement between i and j in the attitude o is exactly 0.5. The situation depicted by Fig. 11 suggests that for this parameter vector there can be an equilibrium state with extreme 'reversed intergroup polarization'. This is the state in which both groups adopt a maximally negative view of their own group. In this case, members of both groups are not pulled towards the opposite view of the outgroup they love, because their disagreement with this outgroup is too large. This state is likely to emerge with a large fraction of initial renegades, because these renegades are immune to influence from ingroup members with a favorable view of their own group, while the latter are attracted towards the opinion of the renegades. As a test, I conducted a simulation of this scenario, using the initial distribution of Experiment 2.

Figure 12 shows two typical runs of this condition. In one of these runs (12(b)), a quick reversal of the initial tendencies of both groups occurs such that almost perfect 'reversed intergroup polarization' arises and stabilizes. In the other run (12(a)), the population moves into 'regular' between-group polarization. However, some renegades join the other side. As the left part of Fig. 12 shows, this prevents the dynamics from settling into an equilibrium because ingroup lovers in both camps are pushed to



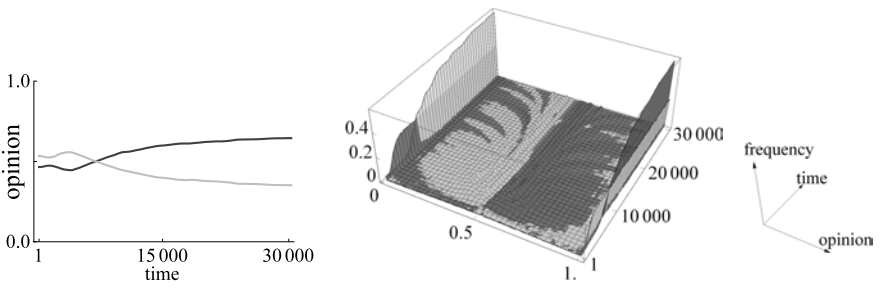
(a) Mixed between-group polarization (b) Reversed between-group polarization

Fig. 12. Individual trajectories for two typical simulation runs with $\beta_O = \beta_D = \beta_A = \frac{1}{3}$, large share initial renegades. Dark: group 0, Light: group 1.

differentiate from the attitudes of the renegades who simultaneously belong to the outgroup and happen to agree with them.

Figure 13 shows the corresponding dynamics of the average trajectories of the mean opinions within both groups (13(a)) and of the distribution of opinions in both groups (13(b)), aggregated over 100 independent realizations.

Figure 13 confirms how a sufficient critical mass of renegades at the outset can change the dynamics of intergroup polarization. On average, more individuals develop into ingroup haters and outgroup lovers in this condition than the other way around. Still, on average both groups contain a variety of different attitudes, caused by the instances in which the 'mixed between-group polarization' occurs that is shown in Fig. 12(a). The non-monotonous change of average attitudes shown in Fig. 13(a) is caused by the aggregation of two different dynamics with different speeds, as illustrated by Fig. 12. The 'regular' intergroup polarization builds up faster than the 'reversed' intergroup polarization, resulting in the aggregate in a shift of the average attitude in both groups first towards the initially predominant direction, followed by a subsequent reversal when the slower dynamic of reversed intergroup polarization pulls the aggregate mean attitudes in the opposite direction.



(a) Mean opinions (b) Opinion distribution

Fig. 13. Average dynamics of opinions aggregated over 100 realizations for $\beta_O = \beta_D = \beta_A = \frac{1}{3}$. Dark: group 0, light: group 1.

4. Discussion and Conclusion

Formal models of opinion dynamics hitherto consider intergroup attitudes mainly in terms of fixed xenophobia. This implies that individuals differentiate more likely from outgroup members than from ingroup members, and assimilate more likely towards ingroup members than towards outgroup members. These models are prone to generate the outcome of between-group polarization, bi-polarization of opinions aligned with group boundaries. However, models of this type also neglect an important insight from research on intergroup relations. Intergroup attitudes are themselves subject to social influence from ingroup members as well as from outgroup members.

In this paper, a model was proposed that extends previous models of opinion dynamics in demographically diverse groups. The model integrates the feedback loop between social influence on intergroup attitudes and the effect that those intergroup attitudes have on influence from ingroup members and from outgroup members. Simulation experiments show that the effects of fixed xenophobia on intergroup polarization can profoundly differ from those of socially influenced intergroup attitudes. Broadly, I find that not only are the conditions for intergroup polarization more restrictive when intergroup attitudes are open to social influence. Also, when intergroup attitudes have a stronger impact on social influence this does not necessarily increase between-group polarization, in contrast with models assuming fixed xenophobia. Instead, a model with socially influenced intergroup attitudes can generate richer dynamics, including outcomes in which both groups contain ingroup lovers, individuals with a moderate intergroup attitude, and ingroup haters or, renegades. Especially, when there is a sufficient initial critical mass of renegades who hold a more favorable attitude of the outgroup than of the ingroup, this critical mass can trigger a dynamic of reversed intergroup polarization in which eventually all members of a group side with the renegades in being critical of their ingroup.

The simulation experiments show that in modeling intergroup opinion dynamics, we can not safely neglect the possibility that intergroup attitudes not only affect social influence but also are subject to social influence. At the same time, these simulation results should not readily be seen as guidance for empirical research on intergroup relations. The model presented here constitutes a first step in integrating social influence with intergroup attitude dynamics and it relies on a number of simplifications and abstractions that need to be carefully explored in future work. In what follows, the most important simplifications are discussed.

First, it is assumed in the current model that there is only one attitude dimension that is subject to social influence, and this attitude dimension simultaneously is the intergroup attitude. This simplification was chosen to focus on what happens to opinion dynamics if intergroup attitudes are socially influenced. At the same time, both earlier modeling work (e.g. [5, 24, 25, 34, 35, 42]) as well as empirical evidence (e.g. [29]) suggest that opinion differences on issues which are not necessarily linked to group identities can align over time with 'fixed' differences between groups. But

opinions on such issues may also create bridges between groups. Future work should explore whether socially influenced intergroup attitudes mitigate the degree to which demographic differences can elicit between-group polarization also on other issues than intergroup attitudes, and under which conditions.

Second, the model also abstracts from homophily [40, 49], the clustering of social relations within groups. In the current model all pairs of individuals can meet with equal probability, but this is an unrealistic description of most empirical settings in which social influence occurs. Homophily may be caused by structural patterns of social interaction that systematically sort similar people into similar foci [19] where they meet and interact, like schools, neighborhoods, or workplaces. But people may also prefer to interact with similar rather than dissimilar others, as has been found in studies of social network formation in various contexts [54, 62]. Previous computational modelling work on social influence dynamics has shown how both structurally imposed network homophily [20, 24] as well group-based homophily in selection of network neighbors in dynamic networks [28, 30] can profoundly affect social influence dynamics. Both forms of homophily could also lead to different conclusions about the impact of socially influenced intergroup attitudes. For example, fixed xenophobia would get much less of a chance to drive between-group polarization if members of different groups are unlikely to interact due to structural homophily [20, 24]. At the same time, reversed intergroup polarization may become less likely if renegades could change their group membership [28]. In this case, they might 'defect' to the other group and thus lose their influence on their former peers because they are now considered to be outgroup members. These possibilities can be explored in future work by integrating both structural segregation as well as dynamic networks into the framework proposed here.

A third simplification is the assumption that there is an inverse relation between ingroup attitudes and outgroup attitudes. In the model presented here, a more positive attitude towards the ingroup implies a less positive attitude towards the outgroup and vice versa. Drawing on social identity theory [65], empirical research on intergroup attitudes found support for such a negative association (e.g. [51]) under certain conditions. However, this assumption should be seen as simplification of an empirically more complex relationship between ingroup attitudes and outgroup attitudes (see e.g. [8, 56]) that can be incorporated in the modeling framework in future research. Changing this assumption may, for example, yield different insights about the role of renegades. If renegades can at the same time have a favorable view of their own group as well as of an outgroup, a critical mass of renegades may less readily lead to the reversed between-group polarization that has been shown for the present model. The reason is that renegades would then be open to influence from both ingroup members and outgroup members.

A further avenue for future research is exploration of alternative assumptions about the influence process as such. The model presented here belongs to the broader class of models that combine assimilative and repulsive influence (e.g. [5, 23, 24, 34, 43, 44]. These models are prone to generate bi-polarization under a large range of

conditions. While empirical studies point to some evidence for repulsive influence in experimental and field research [33, 41], recent experimental research testing repulsive influence more systematically, found no support [11, 66]. This suggests that repulsive influence may be less easily triggered in social interactions than most models assume. Alternative assumptions about influence processes studied in the literature include 'bounded confidence' [13, 32], the assumption that influence is constrained by lack of similarity [4], or modified by 'biased assimilation' [12]; exchange of persuasive arguments [47], a persistent pull towards fixed initial tendencies [27] or a 'strive for uniqueness' [45]. Hitherto the literature is plagued by an insufficient understanding of the differences and similarities between these alternative models of influence [26]. However, broadly we can say that most of these alternative specifications tend to generate opinion diversity more in the form of clustering around a broader range of alternatives in the opinion space, rather than bi-polarization. Future work should thus analyze whether and how using alternative specifications of the influence process affects the fundamental differences between fixed xenophobia and socially influenced intergroup attitudes that were identified for the present model.

Finally, the model presented here specifies group identities as fixed individual properties. This portrays a world in which the group identity of an individual is always observable and can never change. Moreover, it includes the assumption that all individuals agree about who belongs to which group. However, social identity theory and research on the formation of social and ethnic boundaries [70] shows how individuals themselves can change their identification with social categories and, moreover, outsiders can change their view about the group to which someone belongs and of the salience of social distinctions in society. Future extensions should consider the possibility that individuals modify their identification with and perception of social categories. This would, for example, allow to include the possibility that renegades do no longer see themselves as members of their ingroup or are not perceived as 'proper' ingroup members any more by others. This may diminish the influence renegades can have on other members of their social category.

As a last point, it should be noted that more general analytical treatments than those offered here would be desirable. While most of the results shown here hold up when somewhat different parameter values are chosen, exactly how robust results are and under which conditions qualitatively different outcomes obtain has only been established for some of the parameter values that were manipulated in the simulation experiments or robustness tests. Other parameter values were chosen to reflect empirical or theoretical considerations, or a simple baseline scenario (such as the assumption of equally sized subgroups). A promising basis for arriving at more general results might be work by authors in control theory and mathematical social sciences [3, 17, 57] in recent years. Trying to explain persistent disagreement despite assimilative influence, they adopted the same approach that was used in this paper and that was developed earlier in the ABM-literature [24, 36, 43, 44], combining both assimilative and repulsive influence. The recent studies [3, 17, 57] could prove general

theorems about conditions under which permanent disagreement can obtain under such opinion dynamics.

Reflecting earlier analytical insights from the formalization of social balance theory proposed by Cartwright and Harary in the 1950s [10], Altafini [3] showed for a time-continuous model that bi-polarization in the form of 'bipartite consensus' can be reached when a static network of positive and negative influence relations is structurally balanced, falling apart into two subgroups with positive ties within and negative ties between the subgroups. A very similar result was shown with a time-discrete myopic best-response approach [17]. The outcome of 'bipartite consensus' closely resembles some of the equilibria found for the models in this paper, namely those where only two opinions survive that are symmetrical to the midpoint of the opinion spectrum (see also Appendix). Yet, the conditions they found are not directly applicable for our purposes, because they assume static networks, whereas socially influenced intergroup attitudes imply that the signs in a graph of social influence relations can change over time. The recent work of Proskurnikov *et al.* [57] introduced also time-varying signed graphs of positive and negative influence relations and could prove that consensus or bi-polarization arises if under a dynamic of relational change the network remains strongly connected beyond some time point. Possibly, future work could establish analytically if the dynamics of changing pairwise discrepancies generated by the model proposed here satisfy this condition for a specific set of parameters, providing more general explanations of why the models generate consensus, bi-polarization or more complex mixed patterns in specific cases.

The model presented in this paper employs a number of simplifications that suggest directions for future research. It also lacks a comprehensive analytical treatment so far. This notwithstanding, it points to a potentially important insight hitherto overlooked by models which construe intergroup attitudes as fixed xenophobia. Group identities that differentiate between ingroup members and outgroup members do not necessarily foster opinion polarization between groups. When intergroup attitudes are themselves subject to social influence, this may prevent polarization between groups if the extent of initial ingroup bias is moderate. Or, it may lead to a complex landscape of emergent attitudes about ingroup and outgroup, including extreme ingroup love, moderate intergroup attitudes and ingroup hate within all groups at the same time. This suggests that the model proposed here opens up a wide range of possibilities for future theoretical investigation of the role of socially influenced intergroup attitudes in opinion dynamics.

Acknowledgments

The author wishes to acknowledge that this paper has benefited from insightful discussions with members of the research group NNC (Norms and Networks) at the Department of Sociology, University of Groningen, and from the instructive comments of two anonymous reviewers. A preliminary version of this paper was presented at the Interdisciplinary Workshop on Opinion Dynamics and Collective

Decision 2017, held at Jacobs University Bremen, Germany, July 5–7, 2017 (organizer: Jan Lorenz). Discussions with the workshop participants importantly helped to improve the paper.

Appendix A

In this appendix, some analytical results are derived for the conditions under which a distribution of intergroup attitudes o can be stable. These results will be related to outcomes of the simulation experiments presented in the main paper.

To begin with, conditions are derived for the situation where only fixed xenophobia and socially influenced intergroup attitudes have an effect on perceived discrepancy, i.e. $\beta_O = 0$.

The case of β_O=0

Under which conditions does an intergroup attitude o_{it} constitute an equilibrium attitude in the sense that the influence of both ingroup members and outgroup members on individual i is zero?

Given the monotonicity of the weight function in Eq. (4), for a given intergroup attitude o_{it}, the weight w_{ijt} towards an individual j is zero if and only if i experiences a discrepancy of $d_{ijt} = \frac{1}{2}$ with j. Thus the condition under which o_{it} is an equilibrium attitude is equivalent to the condition that the discrepancy i experiences is exactly $\frac{1}{2}$ both for an ingroup member j as well as for an outgroup member j. Notice that with $\beta_O = 0$, the attitude held by j is irrelevant for the discrepancy. Equations (A.1) and (A.2) give the discrepancies that follow from Eq. (3) for an ingroup member and an outgroup member, respectively, given $\beta_O = 0$. The time index t is omitted, because this concerns equilibrium conditions. Furthermore, it is assumed without loss of generality that o_i is the attitude about the outgroup of i.

$$d_{\text{in}} = \beta_A o_i, \tag{A.1}$$

$$d_{\text{out}} = \beta_D + \beta_A(1 - o_i). \tag{A.2}$$

From these equations, we can derive the intergroup attitudes for which the influence weights are zero. These are the attitudes at which discrepancy is exactly 0.5. Without loss of generality, consider the outgroup attitude of an agent. From (A.1), we obtain the attitude o_i^{in} towards the outgroup for which a focal agent i experiences $d_{\text{in}} = \frac{1}{2}$ when interacting with an ingroup member. From (A.2), we can find the attitude o_i^{out} towards the outgroup for which $d_{\text{out}} = \frac{1}{2}$ when i interacts with an outgroup member. Equations (A.3) and (A.4) show the corresponding attitudes.

$$o_i^{\text{in}} = \frac{1}{2\beta_A}, \tag{A.3}$$

$$o_i^{\text{out}} = 1 + \frac{2\beta_D - 1}{2\beta_A}. \tag{A.4}$$

Equations (A.3) and (A.4) provide a proof that explains one of the results from Experiment 1 reported in the main paper. If $\beta_A = 1$, then the equations yield $o_i^{\text{in}} = o_i^{\text{out}} = \frac{1}{2}$, which shows that $o_i = \frac{1}{2}$ is always an equilibrium attitude for i if $\beta_A = 1$, because if an agent hold this intergroup attitude, the influence weight is zero. This explains why in Experiment 1 we see over time an increasing number of individuals adopt $o_i = \frac{1}{2}$ if $\beta_A = 1$.

Equations (A.3) and (A.4) also explain the reversed intergroup polarization found in Experiment 2. More specifically, the equations explain why, given $\beta_O = 0$, the outgroup attitudes towards which both groups move constitute increasing ingroup hate and outgroup love as β_D increases. For illustration, consider the corresponding equilibrium values for the outgroup attitudes that ensue from Eqs. (A.3) and (A.4) for $\beta_O = 0$ and the values of β_D used in Experiment 2, which were 0, 0.2 and 0.4, respectively. The corresponding outgroup attitudes are 0.5, 0.625 and 0.833, respectively. More generally, Eqs. (A.3) and (A.4) imply that the larger β_D, the farther the equilibrium outgroup attitude shifts away from a neutral outgroup attitude towards outgroup love and ingroup hate. For values of $\beta_D > \frac{1}{2}$, no feasible solutions exist for (A.3) and (A.4), because both equations yield outgroup attitudes larger than one.

The case of $\beta_O > 0$

In what follows, conditions are analyzed for equilibria with perfect between-group polarization in which there is perfect consensus within both of the two groups, but not necessarily between them. Notice that many of the outcomes found in the simulation experiments with socially influenced intergroup attitudes, do not fall into this category, because they comprised ingroup lovers as well as renegades in both groups at the same time. As will be shown below, equilibrium outcomes with perfect between-group polarization can be characterized analytically.

For this, it is useful to distinguish between the attitudes in which members of group 0 and group 1, respectively, experience discrepancy $d = \frac{1}{2}$ towards both ingroup members and outgroup members. From now on, it is assumed that the attitude o expresses the attitude towards group 1. Let d_{in}^0 denote the discrepancy that members of group 0 experience when interacting with an ingroup member. Assuming that ingroup members fully agree with the focal individual i, Eq. (3) implies the value of d_{in}^0 as given by Eq. (A.5).

$$d_{\text{in}}^0 = \beta_A o_{\text{in}}^0. \tag{A.5}$$

Solving (A.5) for the condition of zero weight or, $d_{\text{in}}^0 = \frac{1}{2}$, yields the condition for the equilibrium attitude o_{in}^0 of group 0 given by (A.6).

$$o_{\text{in}}^0 = \frac{1}{2\beta_A}. \tag{A.6}$$

Notice that this condition is the same that was obtained for $\beta_O = 0$. This follows from the premise that all group 0 members have the same attitude towards group 1 in perfect between-group polarization, so that their mutual disagreement is zero.

The discrepancy that members of group 1 experience when interacting with an ingroup member can be similarly obtained from Eq. (3), as given by (A.7), which yields (A.8) upon solving for $d^1_{\text{in}} = \frac{1}{2}$.

$$d^1_{\text{in}} = \beta_A(1 - o^1_{\text{in}}), \tag{A.7}$$

$$o^1_{\text{in}} = 1 - \frac{1}{2\beta_A}. \tag{A.8}$$

The conditions given by (A.6) and (A.8) are necessary conditions for an equilibrium consensus within both groups, but they may not be sufficient. Possible solutions are further constrained by the requirement that in both groups also the discrepancy towards outgroup members must satisfy $d = \frac{1}{2}$. Equations (A.9) and (A.10) express the discrepancies towards outgroup members as function of the intergroup attitude o_i and the disagreement Δo towards the attitude of the outgroup members for group 0 and group 1, respectively.

$$d^0_{\text{out}} = \beta_O \Delta o + \beta_D + \beta_A(1 - o_i), \tag{A.9}$$

$$d^1_{\text{out}} = \beta_O \Delta o + \beta_D + \beta_A o_i. \tag{A.10}$$

Equilibrium requires that $o_i = o^0_{\text{in}}$ in (A.9) and $o_i = o^1_{\text{in}}$ in (A.10). Substituting the term from (A.6) for o_i in (A.9) and from (A.8) for o_i in (A.10), and then solving both (A.9) and (A.10) for the disagreement with the outgroup Δo^* that satisfies the equilibrium conditions, we obtain for both equations the same solution:

$$\Delta o^* = \frac{1 - \beta_A - \beta_D}{\beta_O}. \tag{A.11}$$

By virtue of the premise $\beta_O + \beta_D + \beta_A = 1$ we obtain $\Delta o^* = 1$ as the only feasible level of disagreement between the groups that satisfies these equilibrium conditions. From (A.6) and (A.8) it follows that this level can in equilibrium only be obtained if $\beta_A = 0.5$, which yields $o^0_{\text{in}} = 1$ and $o^1_{\text{in}} = 0$. Thus with $\beta_A = 0.5$, maximal 'reversed intergroup polarization' constitutes an equilibrium.

To be sure, there are also other conditions possible under which no change occurs in the population. In particular, the basic influence Eqs. (1) and (2) imply that an individual will never change her opinion after interaction with another individual who holds the same attitude.

This follows, because $\Delta o^* = 0$ if $o_{jt} - o_{it} = 0$, regardless of the influence weight w. Thus, every state in which all members of the population have the same attitude is an equilibrium. Furthermore, every state in which there is consensus within each of the two groups can be an equilibrium if the condition is satisfied that for both groups the discrepancy experienced in interaction with outgroup members meets $d = \frac{1}{2}$. Thus,

equilibrium also is reached if there is consensus within each of the two groups and in addition (A.12) and (A.13) hold.

$$d^0_{\text{out}} = \beta_O \Delta o + \beta_D + \beta_A(1 - o^0), \tag{A.12}$$

$$d^1_{\text{out}} = \beta_O \Delta o + \beta_D + \beta_A o^1. \tag{A.13}$$

From the requirement $d^0_{\text{out}} = d^1_{\text{out}}$ it follows that $\beta_A(1 - o^0) = \beta_A o^1$, which is equivalent with (A.14).

$$(1 - o^0) = o^1. \tag{A.14}$$

This condition is not sufficient to guarantee equilibrium, but it shows that if there is an equilibrium with perfect consensus within each of the two groups, the attitudes which the two groups adopt are symmetrical to the midpoint of the opinion scale. Figure 8 in the main paper shows some examples of cases where $\beta_O = 0$, further simulations confirmed that such equilibria also can obtain when $0 < \beta_O, \beta_D, \beta_A < 1$.

It also possible to give a more precise characterization of the conditions under which equilibria can obtain in which there is perfect consensus within each of the two groups. To see this, rewrite the conditions given by (A.12) and (A.13) by expanding the term Δo. There are two possibilities. In 'reversed' intergroup polarization, the equilibrium attitude of group 0 exceeds the one of group 1, i.e. $o^0 > o^1$, in 'regular' intergroup polarization it is the other way around, i.e. $o^1 > o^0$. First, rewrite (A.12) and (A.13) for reversed intergroup polarization:

$$d^0_{\text{out}} = \beta_O(o^0 - o^1) + \beta_D + \beta_A(1 - o^0) = \frac{1}{2}, \tag{A.15}$$

$$d^1_{\text{out}} = \beta_O(o^0 - o^1) + \beta_D + \beta_A o^1 = \frac{1}{2}. \tag{A.16}$$

Solving these two conditions simultaneously for o^0 and o^1 yield

$$o^0 = \frac{2\beta_A + 2\beta_D - 2\beta_O - 1}{2(\beta_A - 2\beta_O)}, \tag{A.17}$$

$$o^1 = \frac{1 - 2\beta_D - 2\beta_O}{2(\beta_A - 2\beta_O)}. \tag{A.18}$$

However, there is no guarantee that this solution meets the criterion $o^0 > o^1$. To assure this, we can substitute in $o^0 > o^1$ the r.h.s. terms from (A.17) and (A.18), which yields after some further rearrangement while utilizing $\beta_A = 1 - \beta_O - \beta_D$, the condition given in (A.19).

$$\left(\left(\beta_O < \frac{1}{3}(1 - \beta_D\right)\right) \wedge (\beta_O \leq \beta_D)) \vee \left(\left(\beta_O > \frac{1}{3}(1 - \beta_D\right)\right) \wedge (\beta_O \geq \beta_D)). \tag{A.19}$$

If condition (A.19) is not met, it is still possible that there is an equilibrium with regular intergroup polarization. This imposes the precondition that $o^1 > o^0$, such that finding the equilibrium conditions requires solving the system given by (A.20) and (A.21) for o^1 and o^0.

$$d^0_{\text{out}} = \beta_O(o^1 - o^0) + \beta_D + \beta_A(1 - o^0) = \frac{1}{2}, \tag{A.20}$$

$$d^1_{\text{out}} = \beta_O(o^1 - o^0) + \beta_D + \beta_A o^1 = \frac{1}{2}. \tag{A.21}$$

This yields the following solutions:

$$o^0 = \frac{2\beta_A + 2\beta_D + 2\beta_O - 1}{2(\beta_A - 2\beta_O)}, \tag{A.22}$$

$$o^1 = \frac{1 - 2\beta_D + 2\beta_O}{2(\beta_A - 2\beta_O)}. \tag{A.23}$$

The condition under which (A.22) and (A.23) satisfy $o^1 \geq o^0$ reduces to $\beta_O \geq \beta_D$. This shows that for some parameter configurations only reversed intergroup polarization is possible, for others only regular intergroup polarization, and for some parameter configurations both regular and reversed intergroup polarization are possible equilibriums. Here follow some examples. In all cases, simulations confirmed that the analytically derived equilibrium outcomes indeed constitute stable states of the dynamics.

- *Regular intergroup polarization only*: $\beta_O = 0.2, \beta_D = 0.1, \beta_A = 0.7$. This parameter-vector meets the condition for $o^1 \geq o^0$ and does not meet the condition for $o^0 \geq o^1$. The equilibrium solution obtained is $o^0 = 0.455$ and $o^1 = 0.545$ (rounded to 3 decimals).
- *Reversed intergroup polarization only*: $\beta_O = 0.1, \beta_D = 0.3, \beta_A = 0.6$. In this scenario, the equilibrium solution is $o^0 = 0.75, o^1 = 0.25$. Indeed $\beta_O \geq \beta_D$ is not met, thus the condition for $o^1 \geq o^0$ is not satisfied likewise. Also $(\beta_O < \frac{1}{3}(1 - \beta_D)) \wedge (\beta_O \leq \beta_D)$, thus the condition for reversed intergroup polarization is met.
- *Both regular and reversed intergroup polarization*: $\beta_O = 1/2, \beta_D = 1/3, \beta_A = 1/6$. In this case, reversed intergroup polarization is sustained for $o^0 = 3/5$ and $o^1 = 2/5$. Regular intergroup polarization is sustained for $o^0 = 3/7$ and $o^1 = 4/7$. However, this is no guarantee that any of these two equilibrium outcomes is reached from a random start. Simulation experiments showed examples of other feasible equilibriums. Consensus on $o = 0.5$ was obtained, as well as 'mixed between-group polarization' where all agents adopt maximally extreme positions, but both extreme positions are taken by members from both groups.

References

[1] Abramowitz, A. I. and Saunders, K. L., Is polarization a myth? *J. Polit.* **70** (2008) 542–555.

[2] Allport, G. W., *The Nature of Prejudice* (Addison-Wesley, Cambridge, MA, 1954).

[3] Altafini, C., Consensus problems on networks with antagonistic interactions, *IEEE Trans. Automat. Contr.* **58** (2013) 935–946.

[4] Axelrod, R., The dissemination of culture: A model with local convergence and global polarization, *J. Conflict Resolut.* **41** (1997) 203–226.
[5] Baldassarri, D. and Bearman, P., Dynamics of political polarization, *Am. Sociol. Rev.* **72** (2007) 784–811.
[6] Becker, J. C. and Tausch, N., When group memberships are negative: The concept, measurement, and behavioral implications of psychological disidentification, *Self Identity* **13** (2014) 294–321.
[7] Brewer, M. B., The social self: On being the same and different at the same time, *Personal. Soc. Psychol. Bull.* **17** (1991) 475–482.
[8] Brewer, M. B., The psychology of prejudice: Ingroup love or outgroup hate? *J. Soc. Issues* **55** (1999) 429–444.
[9] Byrne, D., *The Attraction Paradigm* (Academic Press, New York, London, 1971).
[10] Cartwright, D. and Harary, F., Structural balance: A generalization of Heider's theory, *Psychol. Rev.* **63** (1956) 277–293.
[11] Clemm von Hohenberg, B., Maes, M. and Pradelski, B., Micro influence and macro dynamics of opinion formation (2017), https://ssrn.com/abstract=2974413.
[12] Dandekar, P., Goel, A. and Lee, D. T., Biased assimilation, homophily, and the dynamics of polarization, *Proc. Natl. Acad. Sci. U.S.A.* **110** (2013) 5791–5796.
[13] Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G., Mixing beliefs among interacting agents, *Adv. Complex Syst. Complex Syst.* **3** (2000) 87–98.
[14] DiMaggio, P., Evans, J. and Bryson, B., Have American's social attitudes become more polarized?, *Am. J. Sociol.* **102** (1996) 690–755.
[15] Dolderer, M., Mummendey, A. and Rothermund, K., And yet they move: The impact of direction of deviance on stereotype change, *Personal. Soc. Psychol. Bull.* **35** (2009) 1368–1381.
[16] Dovidio, J. F., Love, A., Schellhaas, F. M. H. and Hewstone, M., Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions, *Group Processes & Intergroup Relations* **20** (2017) 606–620.
[17] Eger, S., Opinion dynamics and wisdom under out-group discrimination, *Math. Soc. Sci.* **80** (2016) 97–107.
[18] Evans, J. H., Have Americans' attitudes become more polarized? An update, *Soc. Sci. Q.* **84** (2003) 71–90.
[19] Feld, S. L., Social structural determinants of similarity among associates, *Am. Sociol. Rev.* **47** (1982) 797–801.
[20] Feliciani, T., Flache, A. and Tolsma, J., How, when and where can spatial segregation induce opinion polarization? Two competing models, *J. Artif. Soc. Soc. Simul.* **20** (2017), http://jasss.soc.surrey.ac.uk/20/2/6.html.
[21] Fent, T., Groeber, P. and Schweitzer, F., Coexistence of social norms based on in- and out-group interactions, *Adv. Complex Syst.* **10** (2007) 271–286.
[22] Fiorina, M. P. and Abrams, S. J., Political polarization in the American public, *Annu. Rev. Polit. Sci.* **11** (2008) 563–588.
[23] Flache, A. and Macy, M. W., Small worlds and cultural polarization, *J. Math. Sociol.* **35** (2011) 146–176.
[24] Flache, A. and Mäs, M., How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams, *Comput. Math. Organ. Theory* **14** (2008) 23–51.
[25] Flache, A. and Mäs, M., Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion, *Simul. Model. Pract. Theory* **16** (2008) 175–191.

[26] Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S. and Lorenz, J., Models of social influence: Towards the next frontiers, *J. Artif. Soc. Soc. Simul.* **20** (2017), http://jasss.soc.surrey.ac.uk/20/4/2.html.

[27] Friedkin, N. E., Norm formation in social influence networks, *Soc. Networks* **23** (2001) 167–189.

[28] Gargiulo, F. and Huet, S., Opinion dynamics in a group-based society, *EPL* **91** (2010).

[29] Gentzkow, M., Polarization in 2016, *Toulouse Netw. Inf. Technol. White Pap.* (2016), http://web.stanford.edu/∼gentzkow/research/PolarizationIn2016.pdf.

[30] Groeber, P., Schweitzer, F. and Press, K., How groups can foster consensus: The case of local cultures, *J. Artif. Soc. Soc. Simul.* **12** (2009), http://jasss.soc.surrey.ac.uk/12/2/4.html. See also: http://jasss.soc.surrey.ac.uk/12/2/4/citation.html

[31] Grow, A. and Flache, A., How attitude certainty tempers the effects of faultlines in demographically diverse teams, *Comput. Math. Organ. Theory* **17** (2011) 196–224.

[32] Hegselmann, R. and Krause, U., Opinion dynamics and bounded confidence models, analysis, and simulation, *J. Artif. Soc. Soc. Simul.* **5** (2002), http://jasss.soc.surrey.ac.uk/5/3/2.html.

[33] Hovland, C. I., Harvey, O. J. and Sherif, M., Assimilation and contrast effects in reactions to communication and attitude change, *J. Abnorm. Soc. Psychol.* **55** (1957) 244–252.

[34] Huet, S. and Deffuant, G., Openness leads to opinion stability and narrowness to volatility, *Adv. Complex Syst.* **13** (2010) 405–423.

[35] Huet, S., Deffuant, G. and Jager, W., A rejection mechanism in 2D bounded confidence provides more conformity, *Adv. Complex Syst.* **11** (2008) 529–549.

[36] Jager, W. and Amblard, F., Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change, *Comput. Math. Organ. Theory* **10** (2005) 295–303.

[37] Johnston, L. and Hewstone, M., Cognitive models of stereotype change: III. Subtyping and the perceived typicality of disconfirming group members, *J. Exp. Soc. Psychol.* **28** (1992) 360–386.

[38] Kunda, Z. and Oleson, K. C., When exceptions prove the rule: How extremity of deviance determines the impact of deviant examples on stereotypes, *J. Pers. Soc. Psychol.* **72** (1997) 965–979.

[39] Lau, D. C. and Murnighan, J. K., Demographic diversity and faultlines: The compositional dynamics of organisational groups, *Acad. Manag. Rev.* **23** (1998) 325–341.

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

[41] Liu, C. C. and Srivastava, S. B., Pulling closer and moving apart: Interaction, identity, and influence in the U.S. senate, 1973 to 2009, *Am. J. Sociol.* **39** (2015) 192–217.

[42] Macy, M. W. and Flache, A., Social dynamics from the bottom up: Agent-based models of social interaction, in *Oxford Handbook Analytical of Sociology*, Bearman, P. and Hedström, P. (eds.) (Oxford University Press, 2009), pp. 245–268.

[43] Macy, M. W., Kitts, J., Flache, A. and Benard, S., Polarization and dynamic networks. A hopfield model of emergent structure, in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, Breiger, R., Carley, K. and Pattison, P. (eds.) (The National Academies Press, Washington, DC, 2003), pp. 162–173.

[44] Mark, N. P., Culture and competition: Homophily and distancing explanations for cultural niches, *Am. Sociol. Rev.* **68** (2003) 319–345.

[45] Mäs, M., Flache, A. and Helbing, D., Individualization as driving force of clustering phenomena in humans, *PLoS Comput. Biol.* **6** (2010) e1000959.

[46] Mäs, M., Flache, A. and Kitts, J. A., Cultural integration and differentiation in groups and organizations, in *Perspect. Cult. Agent-Based Simulations Integr. Cult.*, Dignum, V. and Dignum, F. (eds.) (Springer International Publishing, Cham, 2014), pp. 71–90.

[47] Mäs, M., Flache, A., Takács, K. and Jehn, K. A., In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization, *Organ. Sci.* **24** (2013) 716–736.

[48] McLaren, L. M., Anti-immigrant prejudice in Europe: Contact threat perception, and preferences for the exclusion of migrants, *Soc. Forces* **81** (2003) 909–936.

[49] McPherson, M., Smith-Lovin, L. and Cook, J. M., Birds of a feather: Homophily in social networks, *Annu. Rev. Sociol.* **27** (2001) 415–444.

[50] Moody, J., Race, school integration, and friendship segregation in America, *Am. J. Sociol.* **107** (2001) 679–716.

[51] Mummendey, A., Klink, A. and Brown, R., Nationalism and patriotism: National identification and out-group rejection, *Br. J. Soc. Psychol.* **40** (2001) 159–172.

[52] Munniksma, A., Stark, T. H., Verkuyten, M., Flache, A. and Veenstra, R., Extended intergroup friendships within social settings: The moderating role of initial outgroup attitudes, *Gr. Process. Intergr. Relations* **16** (2013) 752–770.

[53] Paolini, S., Crisp, R. J. and McIntyre, K., Accountability moderates member-to-group generalization: Testing a dual process model of stereotype change, *J. Exp. Soc. Psychol.* **45** (2009) 676–685.

[54] Pearson, M., Steglich, C. E. G. and Snijders, T. A. B., Homophily and assimilation among sportactive adolescent substance users, *Connections* **27** (2006) 47–63.

[55] Pettigrew, T. F. and Tropp, L. R., A meta-analytic test of intergroup contact theory, *J. Pers. Soc. Psychol.* **90** (2006) 751–783.

[56] Phinney, J. S., Jacoby, B. and Silva, C., Positive intergroup attitudes: The role of ethnic identity, *Int. J. Behav. Dev.* **31** (2007) 478–490.

[57] Proskurnikov, A. V., Matveev, A. S. and Cao, M., Opinion dynamics in social networks with hostile camps: Consensus vs. polarization, *IEEE Trans. Automat. Contr.* **61** (2016) 1524–1536.

[58] Quillian, L., Prejudice as a response to perceived group threat population composition and anti-immigrant and racial prejudice in Europe, *Am. Sociol. Rev.* **60** (1995) 586–611.

[59] Salzarulo, L., A continuous opinion dynamics model based on the principle of meta-contrast, *J. Artif. Soc. Soc. Simul.* **9** (2006) 13, http://jasss.soc.surrey.ac.uk/9/1/13.html.

[60] Semyonov, M., Raijman, R. and Gorodzeisky, A., The rise of anti-foreigner sentiment in European societies, 1988–2000, *Am. Sociol. Rev.* **71** (2006) 426–449.

[61] Sherif, M., *Group Conflict and Cooperation* (Routledge and Kegan Paul, London, 1966).

[62] Stark, T. H. and Flache, A., The double edge of common interest: Ethnic segregation as an unintended byproduct of opinion homophily, *Sociol. Educ.* **85** (2012) 179–199.

[63] Stark, T. H., Flache, A. and Veenstra, R., Generalization of positive and negative attitudes toward individuals to outgroup attitudes, *Personal. Soc. Psychol. Bull.* **39** (2013) 608–622.

[64] Tajfel, H., Social categorization, social identity and social comparison, in *Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations*, Tajfel, H. (ed.) (Academic Press, London, 1978), pp. 61–76.

[65] Tajfel, H. and Turner, J. C., An integrative theory of intergroup conflict, in *The Social Psychology of Intergroup Relations*, Austin, W. G. and Worchel, S. (eds.) (Brooks/Cole, Monterey, CA, 1979), pp. 33–47.

[66] Takács, K., Flache, A. and Mäs, M., Discrepancy and disliking do not induce negative opinion shifts, *PLoS One* **11** (2016) e0157948.

[67] Van Zalk, M. H. W., Kerr, M., Van Zalk, N. and Stattin, H., Xenophobia and tolerance toward immigrants in adolescence: Cross-influence processes within friendships, *J. Abnorm. Child Psychol.* **41** (2013) 627–639.
[68] Vervoort, M. H. M., Scholte, R. H. J. and Scheepers, P. L. H., Ethnic composition of school classes, majority-minority friendships, and adolescents' intergroup attitudes in the Netherlands., *J. Adolesc.* **34** (2011) 257–267.
[69] Vezzali, L., Hewstone, M., Capozza, D., Giovannini, D. and Wölfer, R., Improving intergroup relations with extended and vicarious forms of indirect contact, *Eur. Rev. Soc. Psychol.* **25** (2014) 314–389.
[70] Wimmer, A., The making and unmaking of ethnic boundaries: A multilevel process theory, *Am. J. Sociol.* **113** (2008) 970–1022.