Online Appendix to

In the short term we divide, in the long term we unite. Demographic crisscrossing and the effects of faultlines on subgroup polarization

Michael Mäs, Andreas Flache, Károly Takács, and Karen A. Jehn

In the main paper we reviewed two theories that have been used to explain effects of the strength of demographic faultlines on intergroup polarization in work teams. The first theory, Lau and Murnighan's (1998) approach, expects that in teams with strong faultlines homophilious selection of interaction partners and influence during interaction leads to the development of subgroups with polarized opinions. The second, a classical sociological theory (Colson 1954; Evans-Pritchard 1939; Flap 1988; Ross 1920), argues that crisscrossing actors, who share demographic attributes with several subgroups, help to overcome conflicts in groups.

We argued that Lau and Murnighan's theory also factors the effects of demographic crisscrossing in, but they have overlooked a crucial implication. Demographic crisscrossing implies that even teams with strong faultlines will overcome intergroup polarization in the long run, although they might suffer from it in the short term. We developed and analyzed a computational model of the opinion and network dynamics in work teams. This model is based on Lau and Murnighan's main assumptions and therefore allows studying the consistency of our informal reasoning.

This online appendix provides material that could not be included in the paper for lack of space but which is useful as additional illustration. In Section one, we provide a closer examination of the two ideal-typical runs presented in the paper. We use here more elaborate outcome measures than in the paper in order to provide a more detailed description of the opinion dynamics of the two runs. For instance, we address here the development of the argument distribution in the simulated teams. Furthermore, we introduce here two short animations of the network and opinion dynamics of the two ideal-typical runs. The movies describe very intuitively the opinion and network dynamics in work teams with and without crisscrossing actors.

In the paper we argued that all work teams that contain crisscrossing actors will overcome group splits that were caused by strong faultlines. We show in Sections two and three that this happens faster when faultlines are weak and selection of interaction is based on weak homophily. This result makes the counter intuitive effect even more puzzling. Even though teams with strong faultlines have a stronger tendency to split up and even though they also remain split up longer, they are faster in arriving at consensus in the long run. In the paper, we discussed an explanation of why the model generates this effect. Here, we present additional analyses that support this explanation. In section two, we focus on how long it takes work teams to overcome group splits. In Section three, we focus on the actual consensus formation at the end of the sense making process. In particular, we provide analyses on the development of the polarization measure in simulation runs that started with maximal

polarization and show that the advantage of strong faultline teams develops at the very end of the process.

In Section four, we summarize additional analyses on the robustness of our results. In particular, we replicated the results of the paper under alternative parameter settings. First, we added a second opinion dimension (K=2). Second, we increased the cognitive capabilities of the agents, assuming that each agent considers 6 instead of 4 arguments (S=6) for opinion formation. In both cases, the results of the paper could be replicated.

Section five focuses on the dropping of arguments. The model assumes that agents base their opinion on the S most recent arguments. Thus, arguments that are not sufficiently recent are dropped. To test whether this assumption has crucial effects on the model's implications, we have implemented alternative dropping rules. In particular, we included that a random argument is dropped. We show in section five that this did not affect the core results of the main paper. Furthermore, we experimented with the assumption that agents tend to drop arguments that are not in line with their current opinion. We show in section five that this dropping rule makes subgroup polarization very likely.

Section six focuses on the selection of interaction partners. Our model is based on the assumption that team members tend to interact with those colleagues who have similar demographic characteristics and who hold similar opinions. The selection of demographically similar colleagues is a core assumption of Lau and Murnighan's theory (Lau and Murnighan 1998). However, the authors did not include the assumption of selection based on opinion similarity. We argue in the main paper that both dimensions of selection should be included in our formal model because both mechanisms have been supported by empirical research (Byrne 1971; McPherson et al. 2001). In Section six we study to which degree the inclusion of opinions in the selection of interaction partners affects the predictions of our model. For this purpose, we adjusted the model in a way such that agents select their interaction partners only based on their demographic attributes and ignore opinions. We conducted computational experiments with this new model version and compared results to the findings of the main paper. We demonstrate in section six that all effects could be replicated. Furthermore, we show results of a model version where agents select their interaction partners only based on opinions. As one would expect, we found that effects of demographic faultlines disappear if one assumes that demographic attributes have no influence on the behavior of the agents. These results demonstrate that the effects that we report in the paper are not driven by our assumption that selection of interaction partners is based on demographic attributes and opinions.

1. Closer examination of the two ideal-typical runs

In the main paper, we discussed two ideal-typical simulation runs that illustrate opinion dynamics in work teams with strong and weak faultlines. We mainly applied graphical techniques (see Figure 3 and 4), which do not inform about the distribution of arguments that agents use to form their opinions. Arguments, however, are highly relevant for the equilibrium conditions. In the following, we describe two additional outcome measures that we used to study model dynamics. We then apply these measures to the two ideal-typical runs that we discussed in the paper. Furthermore, we describe two animation movies of the opinion dynamics in the two runs.

The main dependent variable of our analyses is the level of subgroup polarization in work teams. In the main paper, we operationalized this concept with a measure called *polarization*. *Polarization* captures the degree to which the group can be separated into a small set of factions that are mutually antagonistic in the opinion space and have maximal internal agreement. To compute *polarization*, we use the variance of pairwise opinion agreement across all pairs of agents in the population, where agreement is ranging between -1 (total disagreement) and +1 (full agreement), measured as one minus the average distance of opinions.

Polarization informs about the degree to which a work team has split up into subgroups with opposing opinions. However, it does not depict if an opinion split occurred along the demographic faultline, which is a central proposition of faultline theory. Therefore we consider also the *attribute-opinion covariance*, *cov(fix;flex)*, as an indicator of the degree to which demographic differences align with opinion differences. Technically, this index is computed as the covariance between the vector of pairwise demographic dissimilarities and the pairwise opinion dissimilarities, where we computed for every pair of actors *i* and *j* the average dissimilarity across all opinions and across all demographic attributes, respectively. Technically, the corresponding dissimilarity measures $\Delta_{i,j}^{fix}$ and $\Delta_{i,j}^{flex}$, are given by equations (4a) and (4b).

$$\Delta_{i,j}^{fix} = \frac{1}{D}\sum_{d=1}^{D}\left|c_{i,d} - c_{j,d}\right| \tag{4a}$$

$$\Delta_{i,j}^{flex} = \frac{1}{K}\sum_{k=1}^{K}\left|o_{i,k} - o_{j,k}\right| \tag{4b}$$

The attribute-opinion covariance is then calculated as the covariance of the differences in opinions with demographic differences, across all pairs of agents, as given by equation (5). The higher the value of this measure, the stronger is the relationship between demographic attributes and opinions.

$$cov(fix, flex) = \frac{\sum_{j\neq i}\left(\left(\Delta_{i,j}^{fix} - \overline{\Delta^{fix}}\right)\left(\Delta_{i,j}^{flex} - \overline{\Delta^{flex}}\right)\right)}{N(N-1)} \tag{5}$$

To provide insight into the dynamics of the arguments used in the team, we use the measure *argument diversity*. To calculate it, the computer program counts the number of distinct[1] relevance matrices in the team. This number is then divided by *N*. Thus, if *argument diversity* takes the value 1 then each agent holds a different argument matrix. As long as not all agents base their opinions on exactly the same arguments, they still hold different opinions, or there is at least a chance that through the exchange of arguments their opinions may change. To express that perfect unity corresponds to no diversity whatsoever, we set *argument diversity* =0 if there is perfect consensus in arguments, meaning that all agents use exactly the same arguments and have, by implication, the same opinions. Hence, $0 \leq$ *argument diversity* ≤ 1.

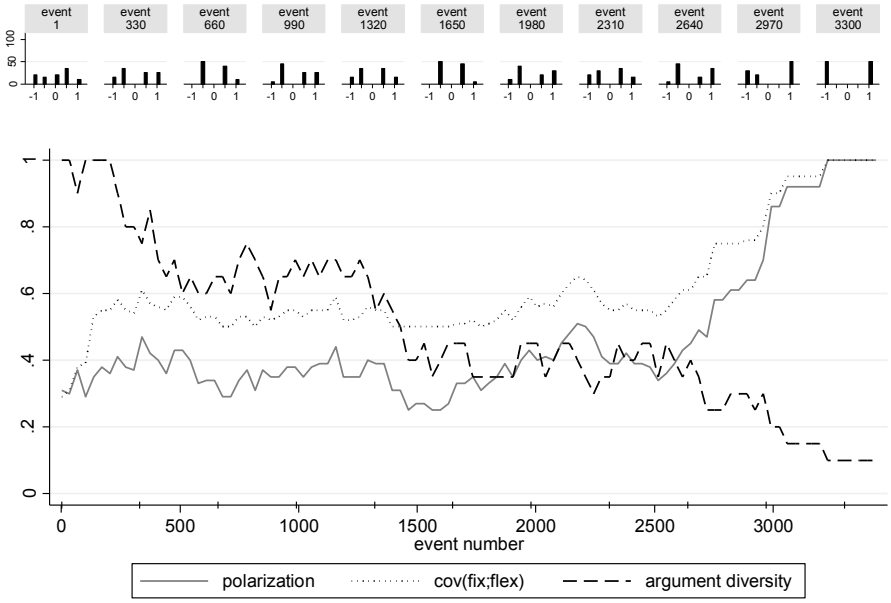
Ideal-typical run without crisscrossing agents. Figure 1 describes the ideal-typical simulation run with maximal faultline strength (*f*=1) that we discussed in the paper.

[1] Two matrices are treated as similar if the same arguments hold nonzero relevance values.

We assumed for this simulation relatively strong homophily (h=5) for all agents. Moreover, we imposed a relatively strong correlation of initial opinions with demographic attributes (w=.8). This lead in this run to an initial Pearson correlation between the opinion and the three demographic attributes of .77. Under these conditions Propositions 1, 1a and 1b predict that the strong faultline in the group should breed intergroup polarization. Furthermore, once the team split up into two demographically homogenous groups with maximally opposing opinion, it is impossible that the split is overcome. The reason is that there are no crisscrossing actors in the team that could build up indirect communication between unconnected team members.

FIGURE 1

Ideal-typical run with maximal faultline strength (f=1, h=5, w=.8)



The upper graphs of Figure 1 report the opinion distribution at different time points. The lower graph shows the development of the three outcome measures *polarization*, *attribute-opinion covariance* and *argument diversity*, measured every 34th event. The ticks that cross the x-axis of the bottom graph indicate to which simulation event the respective bar graph above corresponds. The bar graph for event 1 shows that initially the opinion was almost uniformly distributed. Accordingly, *polarization* took a moderate value (.310). *Argument diversity* adopted its maximal value (1), indicating that each of the 20 agents based the opinion on a different set of arguments. Interaction between members of different subgroups was very rare, but not impossible due to some similarities in opinions. The initial correlation between opinion and demographic attributes (w=.8) resulted in an attribute-opinion covariance of .29 at the outset. This shows that the opinions in the two subgroups already initially tended to be somewhat different from each other. Due to strong homophilious selection, agents were exposed to either mainly pro or mainly con arguments. As expected, this resulted in the polarizing dynamic of argument communication that Lau and Murnighan (1998) predicted. As Figure 1 shows, *polarization* and *attribute-opinion covariance* increased in the process of interaction and eventually reached their maximal values. This shows that the simulation run ended in a clear group split into

two equally large subgroups, which held maximally antagonistic opinions. The final value of *argument diversity* (.1) shows that the two subgroups coordinated independently from each other on two subgroup-specific vectors of either only pro or only con arguments. It was, thus, impossible that any agent's opinion would ever change again, because only agents that hold identical opinions (based on the same arguments) had a nonzero probability to interact.

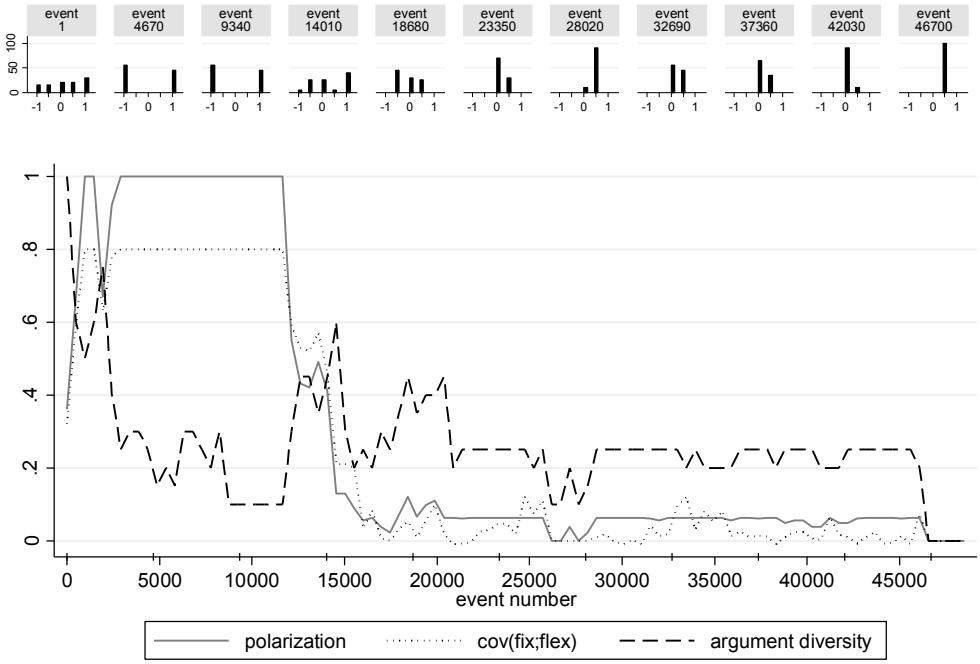
This online appendix also provides a short animation film (McFarland and Bender-deMoll 2007) that further illustrates network and opinion dynamics of this ideal-typical simulation run (see the file "max _faultline.mov"). In the film, each agent is represented by a circle. Circle color indicates to which demographic subgroup the agents belongs. A pair of agents is connected by a line if the agents' overall similarity (sim_{ij}) takes a non-zero value, indicating that there is a positive probability that the two agents engage in interaction. Furthermore, agents are arranged in such a way that pairs of agents with a high opinion similarity are placed closer to each other (Kamada and Kawai 1989). The film shows that initially most pairs of agents were connected and interacted. However, the colors of the circles show that agents that belonged to the same demographic subgroup already initially tended to hold similar opinions. The opinion differences between the demographic subgroups intensified during the simulation run and, eventually, became maximal. At the end of the simulation run, the overall similarity between members of the two subgroups was zero (there are no lines connecting the subgroups) and convergence of opinions was impossible.

Ideal-typical run with 3 crisscrossing agents. Figure 2 reports the development of the output measures that we observed in the ideal-typical run with 3 crisscrossing actors (see also Figure 4 of paper).

Figure 2 shows that initially (event 1) *polarization* was relatively small (.36). This was due to the random assignment of arguments and opinions. *Polarization*, however, increased in the first phase of the simulation run, showing that agents adopted more extreme opinions in this phase of the dynamic. The second bar graph shows that the team split up into two subgroups with maximally opposing opinions. What is more, *opinion-attribute covariance* increased to value of .8, the theoretical maximum in a team with faultline strength of 0.8. This indicates that the group split occurred along the demographic attributes. To be more precise, the nine agents that held the value 1 at all three demographic attributes adopted an opinion value of 1. Eight agents, holding the value -1 at all demographic attributes, coordinated at an opinion of -1. The three crisscrossing agents of the team held an opinion of -1 because they shared two demographic characteristics with agents that held the value -1 at all three demographic attributes. Similarity (sim_{ij}) between members of the two demographic subgroups (1,1,1 and -1,-1,-1) took the value zero. As a consequence, there was no interaction between the subgroups but very frequent interaction within the subgroups.

FIGURE 2

Ideal-typical runs with 3 crisscrossing agents (f=.8, h=5, w=.8)



Besides the opinion differences, crisscrossing agents interacted with all team members, what lead at event 1773 to a temporary decrease of *polarization* and *attribute-opinion covariance*. One of the crisscrossing agents adopted an argument of a team member that held the opposing opinion. In one of the subsequent events, the crisscrossing agent communicated the new argument to a team member from the demographic subgroup for which this argument had not been relevant before. However these two agents did not communicate the new argument to further members of that subgroup. Accordingly, the new argument gradually lost its relevance and subgroup differences increased again. At event 8730, also the arguments eventually perfectly aligned with demographic differences. That is, the members of the two subgroups based their opinions on two different sets of either only pro or only con arguments. This is also indicated by an *argument diversity* of 0.1 (N*0.1 = 2 different argument vectors). Note that the remaining 12 arguments were not relevant for the team members and were not considered in upcoming interactions.

After event 11826 the group split was overcome. An agent holding the value -1 at all demographic attributes interacted with a crisscrossing agent. This agent adopted an argument that changed his opinion and communicated it in the subsequent events to other members of his demographic subgroup. As the bar graph indicates, those agents that held an opinion of -1 before, adopted more moderate opinions after event 14010. One of the nine agents, who held an opinion of +1 in the split phase, likewise changed his opinion after this event.

The opinion changes increased similarity (sim_{ij}) between many members of the different subgroups. Interaction between them became more frequent and opinions further converged. This is evident from the *almost* perfect opinion consensus at which the group arrived at event 28020 (see bar graph). At this point, most agents based

their opinion on the same number of pro and con arguments. But there was still a lot of variation with regard to the underlying arguments, as indicated by *argument diversity* (=.15), which shows that there were three different combinations of arguments present in the team. Finally, by event 46700 the team arrived at an overall consensus.

We provide a short animation film (McFarland and Bender-deMoll 2007) of this simulation run (see the file "strong_faultline.mov"). In the film, the three red squares represent the crisscrossing actors. The film shows that the team first split up into two subgroups with maximally different opinion. However, there was still some interaction between all team members and the crisscrossing actors (lines between red squares and the subgroup with the opposing opinion). It becomes obvious, that the crisscrossing actors brought the group together again and, thus, made opinion consensus possible.

Note that the film of this ideal-typical run is shorter than the film of the run with maximal faultline strength. This does not mean that this run ended faster. Actually, it took the run with maximal faultlines fewer events to arrive at equilibrium (see Figure 1 and 2). However, in the run with maximal faultline strength the main dynamics occurred at the end of the process (see also Figure 1) what made it necessary to include information on a larger number of consecutive events. This increased the length of the film.

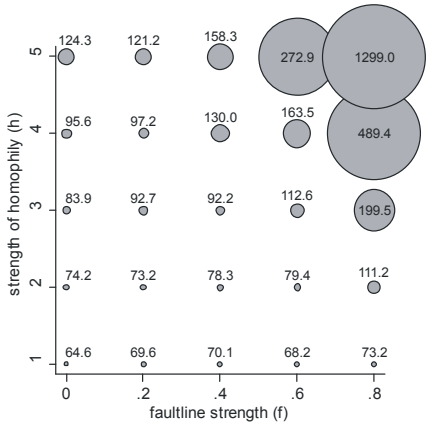
2. How long does it take teams to find perfect consensus?

In the paper we show that teams with strong faultlines can split up into subgroups with polarized opinions. However, as long as there are crisscrossing actors in the team, the split will be overcome and the team arrives at a stable opinion consensus. Furthermore, it turned out that teams with strong faultlines arrive at consensus faster than teams with weak faultlines. In the paper we argued that this counter intuitive effect develops at the end of the sense making process, a phase which Lau and Murnighan refer to as the "challenging phase". In this phase, the communication structures of strong faultline teams make teams faster in arriving at consensus. In the following we present analyses that confirm this interpretation. In particular, we show here that it takes teams with strong faultlines and strong homophily longer to overcome group splits. In other words, teams with strong faultlines tend to split up more than teams with weak faultlines and they also tend to overcome splits at a later stage of the challenging phase. Nevertheless, in strong faultline teams the overall number of events needed to arrive at consensus is smaller than in teams with weak faultlines.

In the paper we report analyses on the maximal value of *polarization* that was reached in a simulation run. Now, we turn to the number of the simulation event, when this value was obtained the last time in the respective run. This serves as an indicator for the length of the split phase. Figure 3 reports how this measure is associated with faultline strength and homophily strength. The sizes of the bubbles in Figure 3 indicate the average length of the split phase. The position of the bubbles expresses the related value of faultline strength and homophily strength. Figure 3 shows that it took simulated teams longer to overcome the group split when the demographic faultline was strong and when homophily was strong.

FIGURE 3

Average number of event when polarization took the run's maximal value the last time



Why do we find this effect? If a team experiences subgroup polarization, the split will be overcome when a crisscrossing actor interacts with a subgroup member who holds a different opinion. One of them learns a new argument and subsequently spreads it in the respective subgroup. Thus, the convergence process starts with an interaction between a crisscrossing member and an agent that holds an opposing opinion. Such events, however, are relatively rare because such pairs of agents differ on two of three demographic attributes and hold opposing opinions. Nevertheless, these events will occur and they are more likely when the team comprises many crisscrossing actors and when homophily is weak.

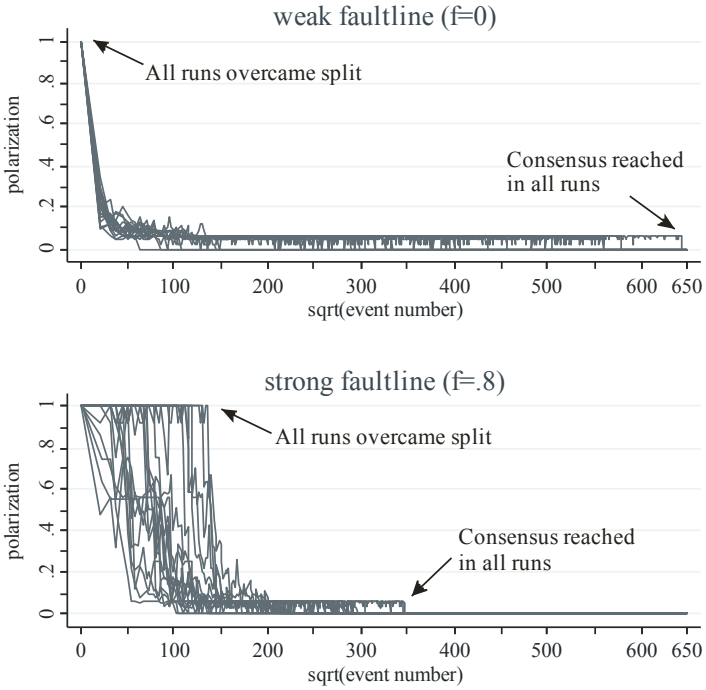
Our indicator of the length of the split phase might be problematic. There could be teams that never reach a serious level of *polarization* but happen to have a small peak in *polarization* at the end of the sense making process. Such cases might be falsely interpreted as an indication for very long lasting splits. However, we showed in the paper that the maximal value of *polarization* that the runs reached is positively associated to faultline and homophily strength (see Figures 6 and 8). This implies that the misinterpretation is most likely when faultlines and homophily are weak because in the remaining conditions short term polarization is usually high. Figure 3, however, shows that these runs where the misinterpretation is most likely, the splits are overcome most quickly. We find this even though our measure for the split length might be problematic.

3. Length of the convergence process

The simulation experiments have revealed that teams with strong faultlines and strong homophily might arrive faster at perfect consensus even though they likely suffer from subgroup polarization in the short term. We have conducted further simulations to confirm this effect, studying 20 simulations runs with strong (f=.8) and weak faultlines (f=0). Under both conditions, we imposed strong homophily (h=5). Furthermore, all runs started with perfect opinion polarization (w=1). As in the paper, we ran the simulations until all agents adopted the same opinion and based it on the same set of arguments.

FIGURE 4

Development of the *polarization* in 20 runs with weak and strong faultlines each



The line graphs of Figure 4 show the development of *polarization* for 20 randomly chosen runs with weak (*f*=0) and strong (*f*=.8) faultlines. We rescaled the horizontal axis by taking the square root of the number of events, because most of the change in the outcome measures occured in the early phase of the dynamics. As imposed, *polarization* was maximal at the beginning of all runs of Figure 4. In runs with weak faultlines, *polarization* declined quickly. This was due to the larger number of crisscrossing actors in teams with weak faultlines, which, in turn, allowed for a faster diffusion of arguments and opinions across demographic subgroups. In most runs with strong faultlines, however, the opinion split remained stable. But while opinions converged quicker with a weak faultline, it also took longer until the teams finally arrived at perfect consensus (on opinions and arguments). By contrast, in the simulation runs with the strong faultline it took more events until arguments were exchanged between subgroups and *polarization* started to decline. However, once a team's subgroups started to influence each other, opinions converged faster than in the teams with weak faultlines.

4. Replication of results under alternative parameter combinations

Replication of results with two opinions (K=2)

In the main paper, we focus on subgroup polarization on a single opinion dimension (K=1). In the following, we show that our results can be replicated when a second opinion dimension (K=2) is included.

To illustrate, Figure 5 shows ideal-typical opinion dynamics under K=2. For this simulation run, we assumed the same conditions as for the typical run with strong faultlines (f=.8) from the paper (see Figure 4 in the paper):

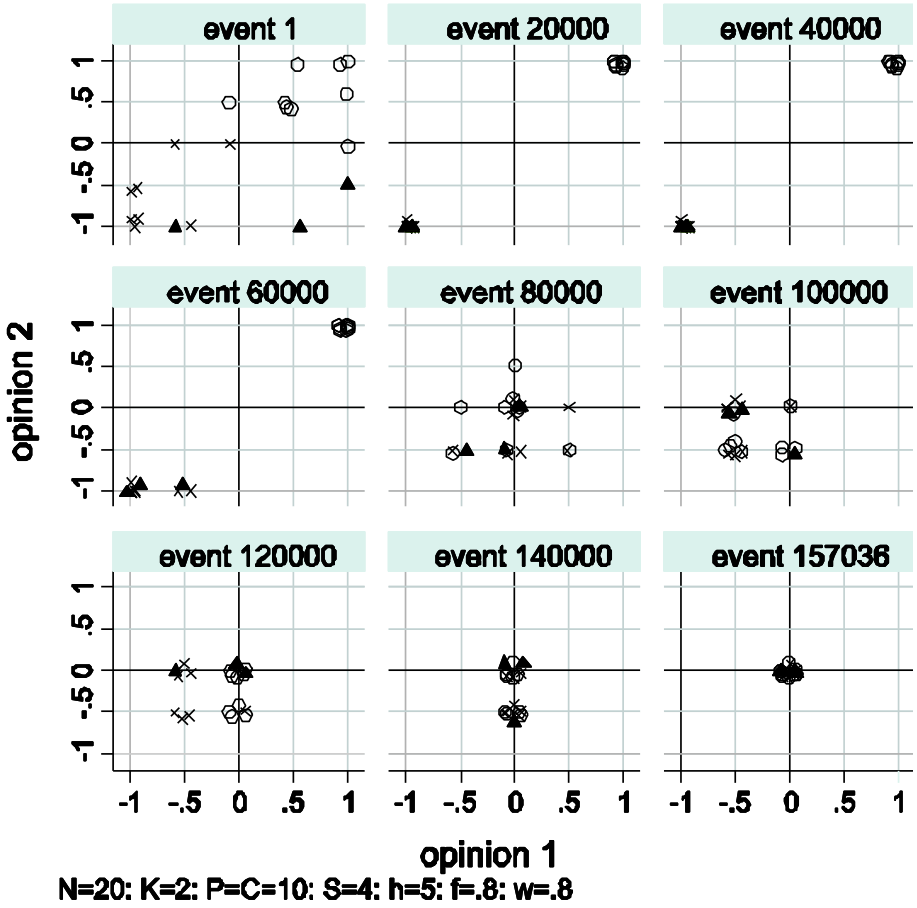
- 20 agents (N=20),
- opinions are based on 4 arguments (S=4)
- for both opinions, there are 10 pro and 10 con arguments (P=C=10)
- strong homophily (h=5)
- strong initial congruency on both opinion dimensions (w=.8 for both opinion dimensions).

Figure 5 shows two-dimensional opinion distributions at nine different points in time. The circles and the crosses indicate to which demographic subgroup the respective agent belonged. The opinions of the three crisscrossing actors are shown with triangles. To prevent that the symbols overlap when agents hold identical opinions, we have added a small random value to the opinions of the agents (only in the Figure, not in the simulations).

The graph for event 1 shows that both opinions were uniformly distributed at the outset. However, high congruency (w=.8) created on both opinion dimensions clear differences between the demographic subgroups. In line with our expectations and findings from the paper, the Figure shows that the simulated work team experiences subgroup polarization on both opinion dimensions. However, due to the crisscrossing actors there was still some exchange of arguments between the two demographic subgroups. After 60,000 simulation events, the perfect opinion split was overcome and the convergence process started. After 157,036 events, perfect consensus was reached.

FIGURE 5

Ideal-typical run with 3 crisscrossing agents and two issues (*f*=.8, *h*=5, *w*=.8)



We have conducted three simulation experiments to test whether the result of the main paper can be replicated under K=2. In the first experiment we focused on the effects of faultline strength and homophily strength on opinion polarization. Therefore, we studied conditions with weak (*f*=0) and strong faultlines (*f*=.8) and with weak (*h*=1) and strong homophily (*h*=5). For each of the four parameter combinations, we conducted 100 independent replications. The initial congruency was fixed at *w*=.9 for both opinion dimensions. To compute *polarization*, we used the identical measure as in the paper but adjusted it to two opinion dimensions. Thus, polarization was measured as the variance of pairwise opinion agreement across all pairs of agents in the population, where agreement is ranging between -1 (total disagreement) and +1 (full agreement), measured as one minus the *average* distance of opinions over the *K* dimensions.

Similar to the Figures in the paper, Figure 6 informs about the average increase in *polarization*. The Figure shows that we could replicate the central findings of Figures 6 and 8 from the paper. First, stronger demographic faultlines lead to a stronger increase in *polarization* (see Figure 6 in the paper). Second, the faultline effect was moderated by homophily strength. That is, homophily needs to be strong to find polarizing effects of strong faultlines (see Figure 8 in the paper).

FIGURE 6

Average maximal opinion *polarization* over *f*, by *h* (*K*=2, *w*=.9; 100 runs per bar)

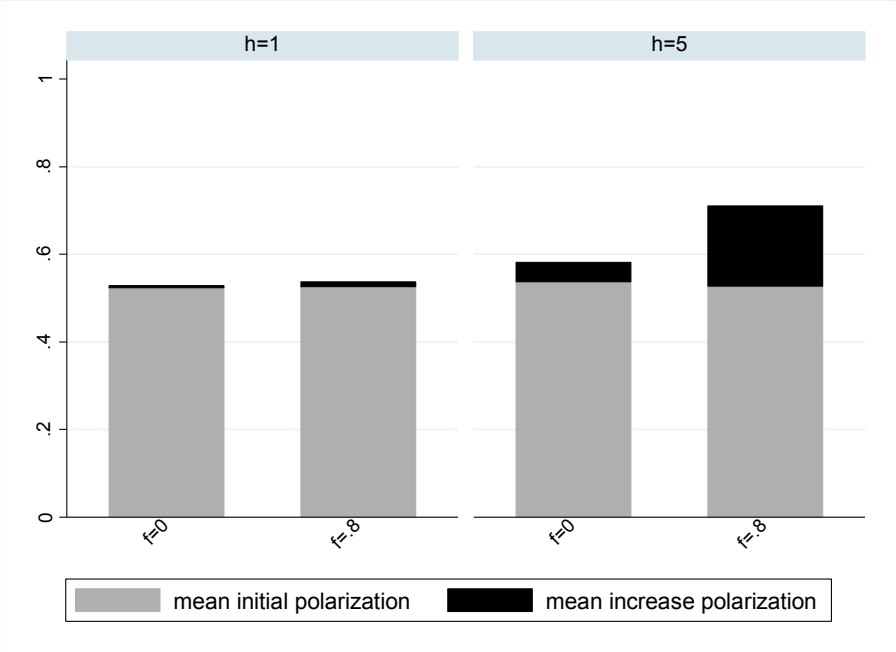
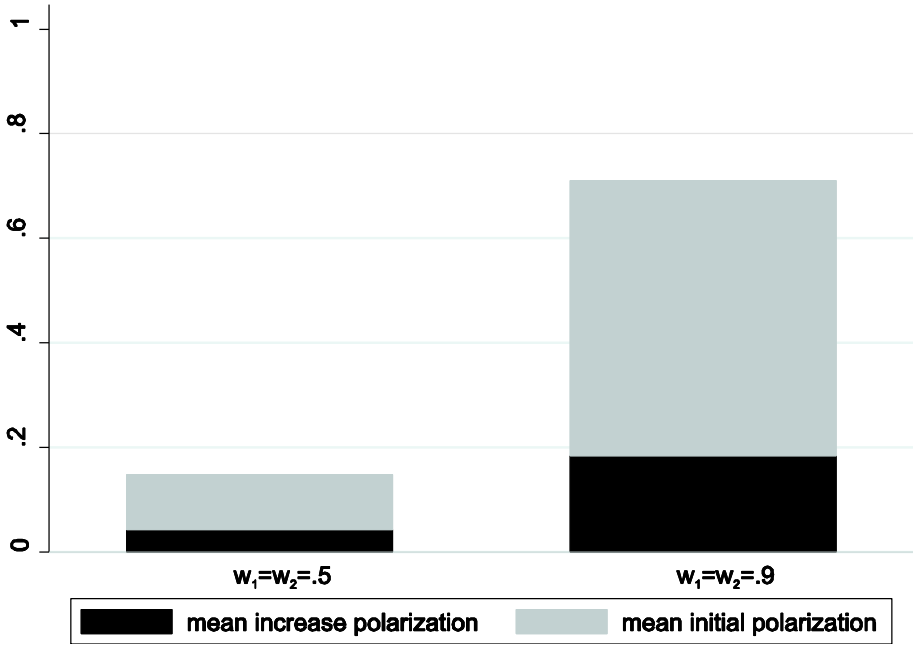


FIGURE 7

Average maximal opinion polarization over *w*, (*K*=2, *f*=.8, *h*=5; 100 runs per bar)



The remaining simulation experiments with *K*=2 focused on the effects of initial congruency (*w*). In experiment 2, we imposed a strong faultline (*f*=.8) and a strong homophily (*h*=5) and varied the initial congruency between *w*=.5 (no congruency) and *w*=.9 (very high initial congruency). We conducted 100 independent replications per condition. Also on this experiment, we imposed the same level of congruency to both opinions. In other words, the congruency between the demographic attributes and the

first opinion (w_1) was the same as the congruency between the demographic attributes and the *second* opinion (w_2). Figure 7 summarizes the results.

Figure 7 shows that also the effect of the initial congruency could be replicated (compare with Figure 7 in the main paper). Our results show that the effect of initial congruency on short term polarization did not differ significantly between K=1 and K=2. Even with a strong faultline, there was only a slight increase in *polarization* when the initial congruency was low. However, the increase in *polarization* was stronger when the initial congruency was high (t=8.39). Statistical analyses showed that the size of the effect of initial congruency (w for K=1 and w_1=w_2 for K=2) on the increase in polarization in the conditions with one opinion dimension did not differ significantly from the effect under K=2 (t=1.5). Moreover, there was no discernible difference in the polarization level between K=1 and K=2 if congruency is high (t=-1.77). This is consistent with our reasoning that the number of opinion dimensions only affects results when initial congruency is weak (see discussion section of main paper).

Comparing the size of the gray bars in Figure 7 and the corresponding bars in Figure 7 of the main paper, one can see that the initial degree of subgroup polarization was clearly weaker under K=2 than under K=1. Note that this is a logical consequence of our method of creating initial congruency. According to this method an initial congruency of e.g. w_1=.9 creates a strong correlation between the first demographic dimension and the first opinion dimension. In other words, most agents hold an opinion that corresponds to the first demographic attribute and there will be very few agents that do not fit into this pattern. Applying the same method to generate the opinion distribution of the second dimension results in a number of agents that do not fit into the imposed pattern, too. However, these agents will most likely not be the same agents as those who do not fit into the relationship between the demographic attribute and the first opinion dimension. Because our measure of subgroup polarization is based on the overall dyadic overlap with regard to all opinion dimensions, polarization must be smaller under K=2 than under K=1.

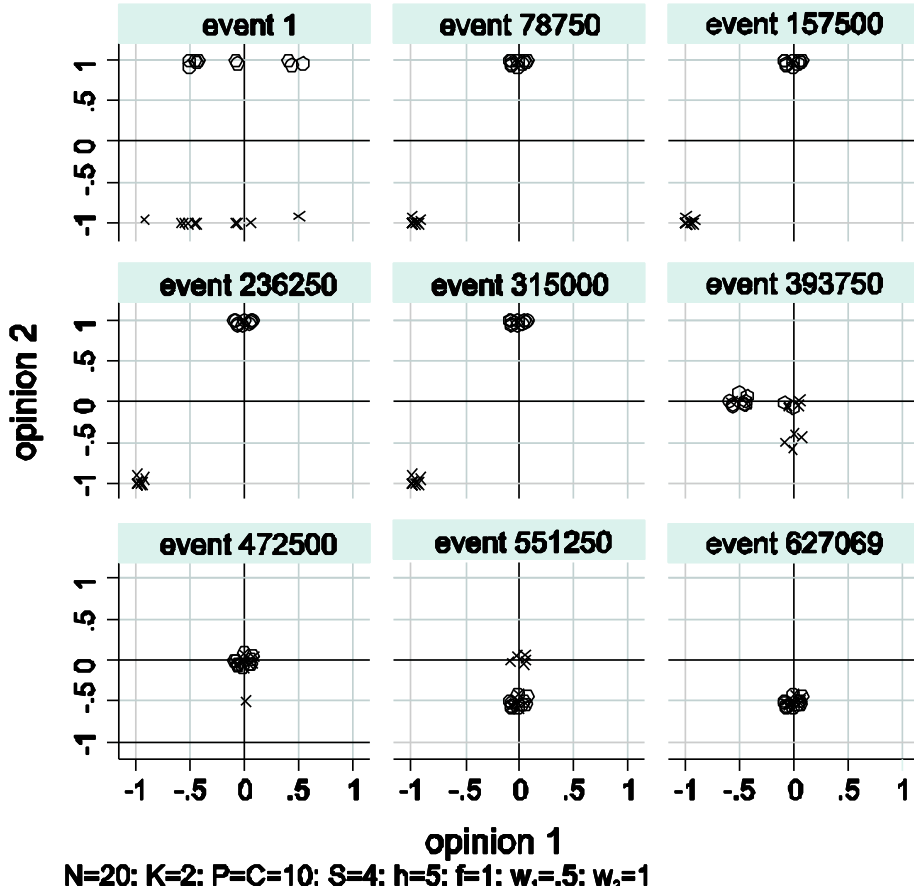
Do we also find subgroup polarization when only one of the opinions correlated with the demographic attributes? Figure 8 depicts dynamics found in a typical simulation run where the first opinion is unrelated to the demographic attributes (w_1=.5) and the second opinion correlates perfectly (w_2=1). In addition, we imposed a perfect faultline (f=1). Thus, the simulated group did not contain demographically crisscrossing actors. For graphical reasons, we have added a small random value to the opinions of the agents (only in the Figure, not in the simulations).

Figure 8 shows that initially the two demographic subgroups (indicated by circles and crosses) were perfectly polarized on the second opinion dimension. However, the first opinion was unrelated to the demographic attributes. Because of the strong homophily, agents interacted mainly with their ingroup members and, thus, also developed a subgroup consensus on the first opinion dimension. Most likely, the subgroups will find consensus on a moderate opinion (see the circle group), because the first opinion was randomly distributed at the outset (w_1=.5). However, it can happen, that a subgroup develops an extreme opinion also on the first dimension (see the cross group in Figure 8), but it is unlikely that both subgroups happen to develop extreme and opposing opinions. If it does happen, then the subgroups differ on all demographic and opinion dimensions and dynamics settle. However, Figure 8 shows the most likely outcome. Independently from each other, both subgroups developed a

consensus but the opinion differences between the subgroups were not maximal. It follows that there was still interaction and exchange of arguments between the two subgroups. Eventually, the subgroups have overcome the split and developed consensus.

FIGURE 8

Ideal-typical run with two issues with different initial congruency (f=1; h=5; w_1=.5; w_2=1)



The third simulation experiment studied the effects of initial congruency on subgroup polarization under the assumption that the two opinions differ in the congruency with the demographic attributes (see Figure 9). Therefore, we simulated work teams with a strong faultline (f=.8) and strong homophily (h=5). We imposed a strong initial congruency on the first opinion dimensions (w_1=.9) and varied the initial congruency on the second dimension between no (w_2=.5) and a very strong congruency (w_2=.9).

FIGURE 9

Average maximal opinion polarization over w_2, (w_1=.9, K=2, f=.8, h=5; 100 runs per bar)

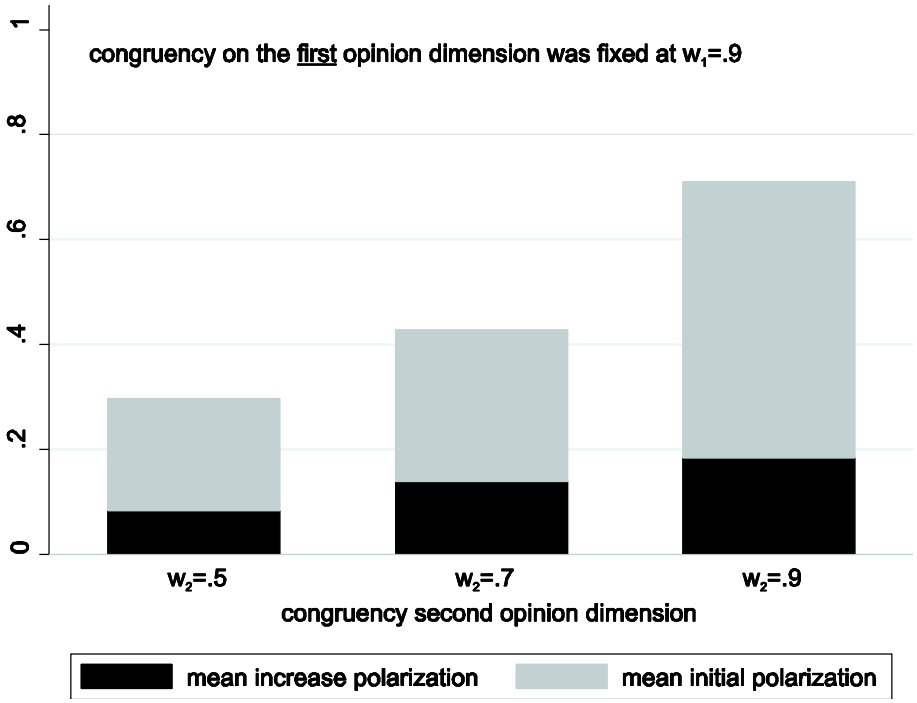
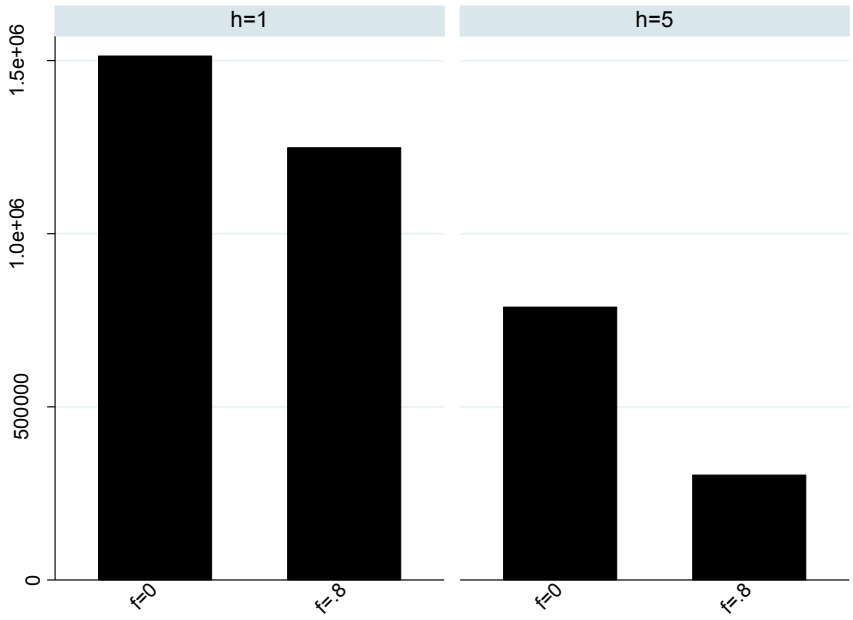


Figure 9 shows that also the results of the third simulation experiment confirm the results of the main paper. In teams with strong faultlines and strong homophily, subgroup polarization is the stronger, the stronger the initial congruency is. In addition, Figure 9 shows that the initial congruency needs to be strong on both opinion dimensions.

Finally, we used the data of the first simulation experiment to test whether the counter intuitive finding of the paper could be replicated with K=2. Figure 9 in the main paper shows that strong demographic faultlines and stronger homophily lead to faster consensus formation. Below, Figure 10 shows the corresponding results from experiment 1. Again, the new experiment confirms the results of the paper.

FIGURE 10

Average number of events until the teams arrived at an overall consensus over *f*, by *h* (*K*=2, w_1=w_2=.9; 100 runs per bar)



Replication of results with higher S

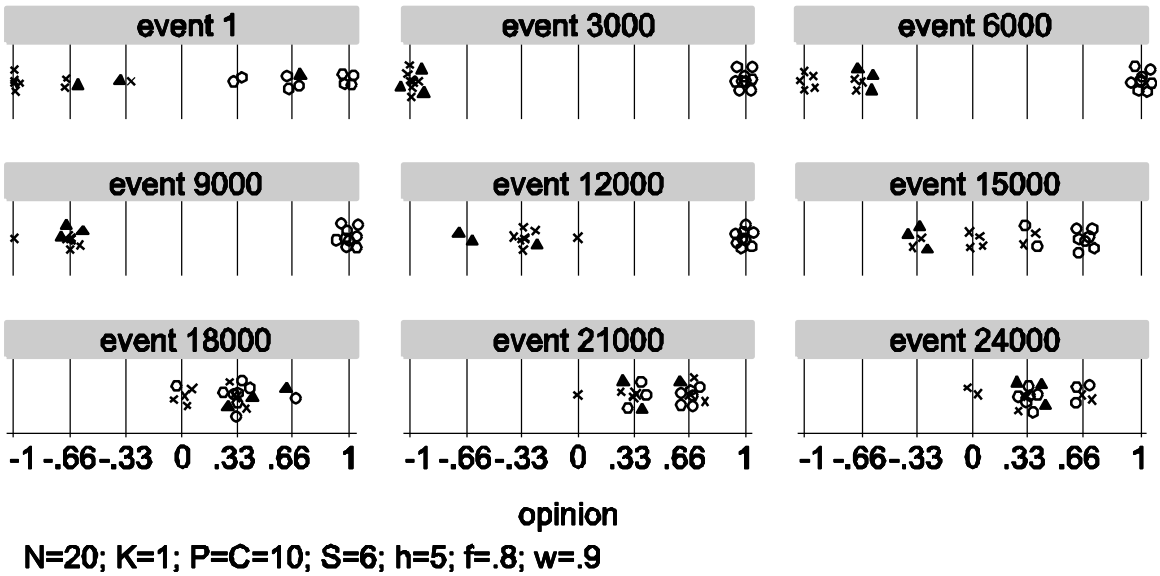
Our model takes into account that humans have limited cognitive capabilities. In particular, we assume that each agent considers a limited number of *S* arguments for opinion formation. In the simulation experiments, we fixed the value of *S* to four (Cowan 2001). To demonstrate that the results of these experiments do not depend on this assumption, we replicated the experiments with *S*=6.

To illustrate that there are no qualitative differences in dynamics when *S* is increased, Figure 11 shows an ideal-typical simulation run under *S*=6. For this run, we imposed a strong faultline (*f*=.8), strong homophily (*h*=5), and a strong initial congruency. Furthermore, we assumed that there is only one opinion dimension (*K*=1) and that there exist 10 pro and 10 con arguments. In Figure 11, the circles and the crosses indicate to which demographic subgroup the respective agent belonged. The opinions of the three crisscrossing actors are shown with triangles. To prevent that the symbols overlap when agents hold identical opinions, we added a small random value to the opinions of the agents (only in the Figure, not in the simulations).

The graph for event 1 informs about the initial opinion distribution. The opinion was distributed over the complete range of the opinion scale. However, the strong congruency (*w*=.9) created clear differences between the demographic subgroups. In line with Propositions 1 and 2 and the findings from the paper, the Figure shows that the simulated work team experienced opinion polarization (see event 3,000) in the short run. However, due to the crisscrossing actors there was still some exchange of arguments between the two demographic subgroups,leading to opinion convergence in the long run. The simulation run of Figure 11 ended after 1,166,506 simulation events (not shown for graphical reasons) with perfect opinion consensus (o_i=.33 for all *i*).

FIGURE 11

Ideal-typical run with *S*=6 (*f*=.8; *h*=5; *K*=1; *w*=.9)

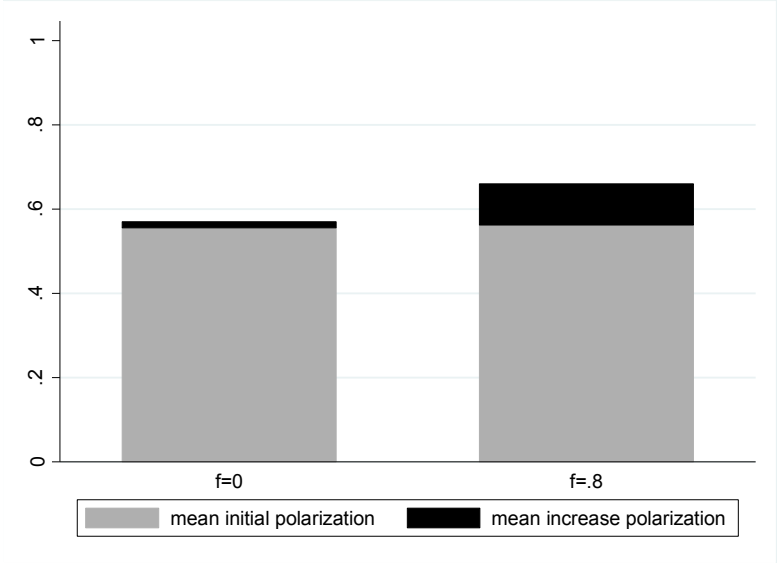


In addition, we have conducted simulation experiments to test whether the main results of the paper can be replicated also with *S*=6. The first experiment focused on the effects of faultline strength on subgroup polarization. We studied two conditions, one with a weak faultline (*f*=0) and one with a strong faultline (*f*=.8). In both conditions, we imposed a strong homophily (*h*=5) and a high initial congruency (*w*=.9).

The results of this experiment confirm the results of the paper (see Figure 6 in the paper). First, all runs ended with a perfect consensus on the opinion and the arguments (Proposition 2). Second, Figure 12 shows that there was higher short-term *polarization* in the condition with the strong faultline, compared to the weak faultline condition. This confirms Proposition 1 from the paper.

FIGURE 12

Average maximal opinion polarization over *f*, (*w*=.9, *K*=1, *h*=5; *S*=6; 100 runs per bar)



In a second experiment, we tested the results from the paper which concern the effects of homophily strength and congruency. In a 2-by-2 design, we imposed a homophily strength of *h*=1 and *h*=5 and an initial congruency of *w*=.5 and *w*=.9. Under all four conditions, we assumed a strong faultline (*f*=.8) and that the agents base their opinion on 6 arguments (*S*=6).

FIGURE 13

Average maximal opinion *polarization* over *w*, by *h* (*K*=1, *S*=6; 100 runs per bar)

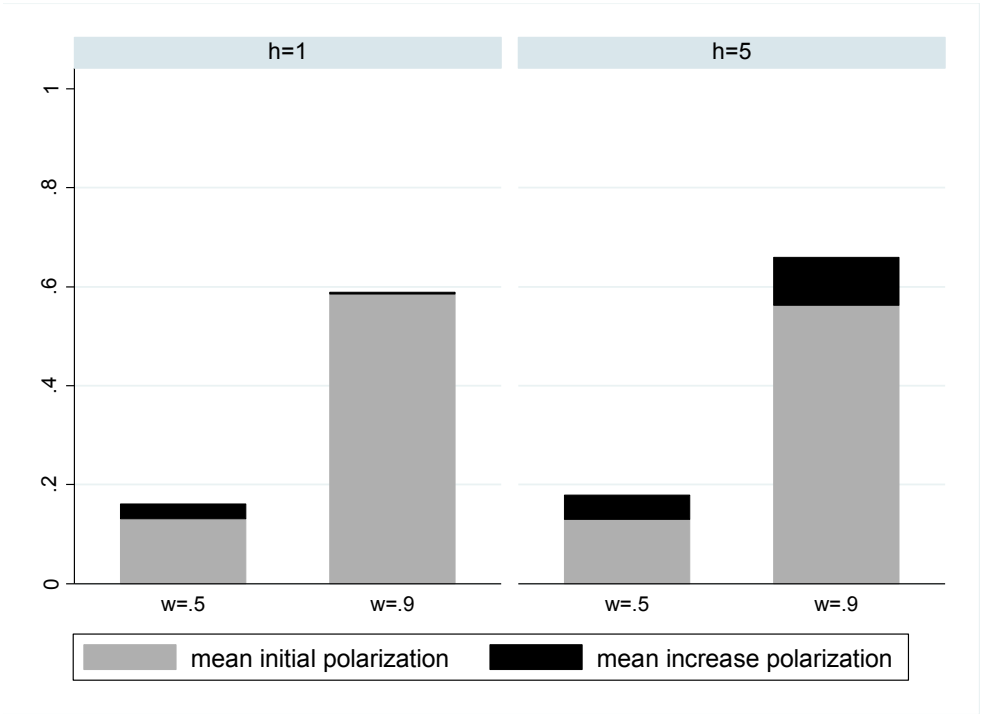
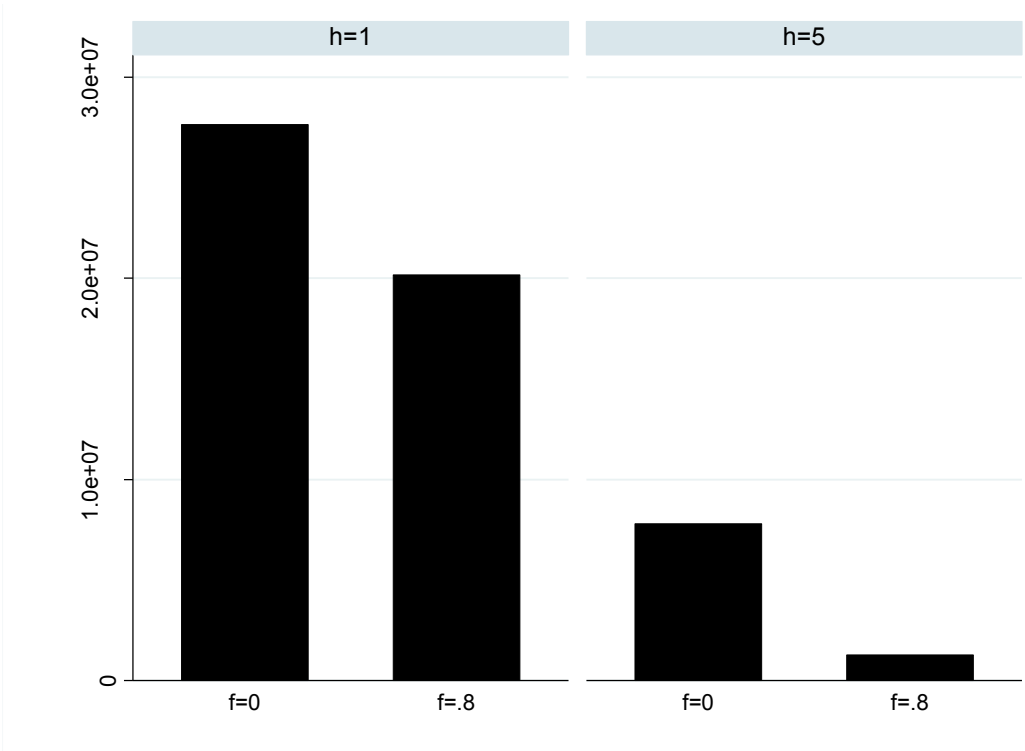


Figure 13 shows that the second experiment with *S*=6 also replicated the results of the paper. First, the Figure confirms Proposition 1b (see Figure 8 of the paper). We found stronger subgroup polarization in the conditions with strong homophily (*h*=5). Second, Figure 13 shows that under strong homophily (see the right panel of the Figure) the increase in *polarization* was stronger when initial congruency was high. Similar to the results of the main paper (see Figure 7 of the main paper), however, this effect was not found in the conditions with weak homophily (*h*=1). As we discussed in the main paper, this effect was caused by a ceiling effect.

Counter intuition, the simulations that we report in the paper suggest that a strong faultline and strong homophily speed up the convergence process (see Figure 9 in the paper). To test if this result can also be replicated with *S*=6, we conducted a third experiment. In this experiment, we varied faultline strength (*f*=0 vs. *f*=.8) and homophily strength (*h*=0 vs. *h*=5) in a 2-by-2 design. In all runs, we assumed a strong initial congruency (*w*=.9) and *S*=6. Figure 14 reports the result of this experiment. Again, the figure confirms the results of the paper: the stronger the faultline and the stronger homophily, the faster consensus was reached.

FIGURE 14

Average number of events until the teams arrived at an overall consensus over *f*, by *h* (*S*=6; *K*=1, *w*=.9; 100 runs per bar)



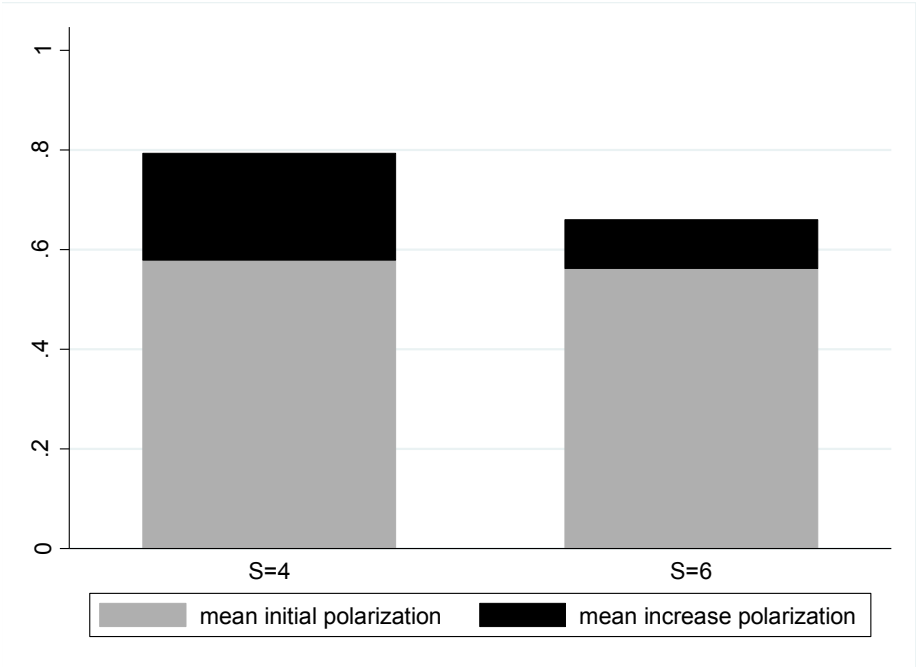
To sum up, the three new experiments show that the core effects of the paper could be replicated also when one assumes that the agents use more than 4 arguments to form their opinion.

In general, the model generates weaker subgroup polarization the higher *S* is. To illustrate this, we compared in Figure 15 results from the experiments of the paper and the new experiments (see above). In particular, we compared the simulations with a strong faultline (*f*=.8), strong homophily (*h*=5) and strong initial congruency (*w*=.9).

The Figure shows that under S=6 the average increase in *polarization* was .12 scale points smaller than in the experiments from the paper where we assumed S=4. This difference is statistically significant (t=-6.89).

FIGURE 15

Average maximal opinion polarization over S, (w=.9, K=1, h=5; f=.8; 500 runs under S=4, 100 runs under S=6)



We think that this difference is caused by two main forces. First, the more arguments the agents use, the smaller is the impact of a single argument on the agents' opinions (see equation 1) and the more arguments are needed to develop a maximally extreme opinion. This makes subgroup polarization less likely. Second, the more arguments the agents consider, the longer it takes before an agent drops an argument that is not in line with her opinion. Consider an agent who holds only pro arguments and happens to learn a con argument. In our model, it will take the agent at least S interactions to drop this argument again. Thus, the higher S is, the longer it will take this agent to readopt an extreme opinion. This makes interaction with others that provide this agent with further con arguments more likely. It also makes it more likely that this agent reports the con argument to others that also use mostly pro arguments. This, in turn, will make them develop more moderate opinions and decreases subgroup polarization.

In this section we collected results of additional analyses in order to demonstrate the robustness of our results against changes of core parameter values. In particular, we demonstrated that the results of the main paper could be replicated when two (instead of one) opinion dimensions were assumed and when agents were assumed to base their opinions on six (instead of four) arguments. Obviously, further tests of robustness are needed. In particular, it would be a stronger test of robustness if one could demonstrate that there are *no* values of K and S under which results would change fundamentally, an endeavor which requires analytical approaches and goes beyond the abilities of agent-based computational modeling. Analytical solutions, however, seems to be out of reach, considering for instance the nonlinearity of central

model mechanisms (see the implementation of homophily) and the influence of randomness (e.g. selection of interaction partners and of the transmitted argument). Future modeling research is needed to study which model implications also follow from more abstract but analytically tractable models of social influence dynamics in work teams with strong and weak faultlines.

5. Random dropping of arguments

In the paper, we developed a formal model of Lau and Murnighan's informal theory. It is essential to build a formal model which does not include assumptions that Lau and Murnighan did not make and that can critically affect the model outcomes. Otherwise, it would not be possible to study the implications of those assumption that Lau and Murnighan do make explicit.

However, in a range of details, Lau and Murnighan (1998) did not specify their theory precisely enough for the specification in a formal computational model. Accordingly, we had to choose our own specification of the corresponding assumptions. In particular, we had to include assumptions about the way how agents memorize and forget arguments they learn from others. Our implementation assumes that agents consider only the most recent arguments for opinion formation and neglect arguments that are not sufficiently recent. This implementation has two central advantages. First, the recency effect is well supported by empirical research (Brown and Chater 2001) and is also central in recently developed cognitive theories (Brown et al. 2007; Sederberg et al. 2008) and formal models of social influence (Mark 2003, Carley 1991).

The second advantage of the recency assumption is that it does not crucially influence the implications of the model. To demonstrate this, we have implemented an alternative dropping procedure. Instead of dropping the least recent argument, we implemented that a random argument is dropped[2]. We conducted a computer simulation experiment to test whether the central results of the paper can be replicated with random argument dropping. In this experiment, we studied 500 runs under weak faultlines (f=0), strong faultlines (f=.8) and maximal faultline strength (f=1). In all runs we imposed strong homophily (h=5) and strong initial congruency (w=.9).

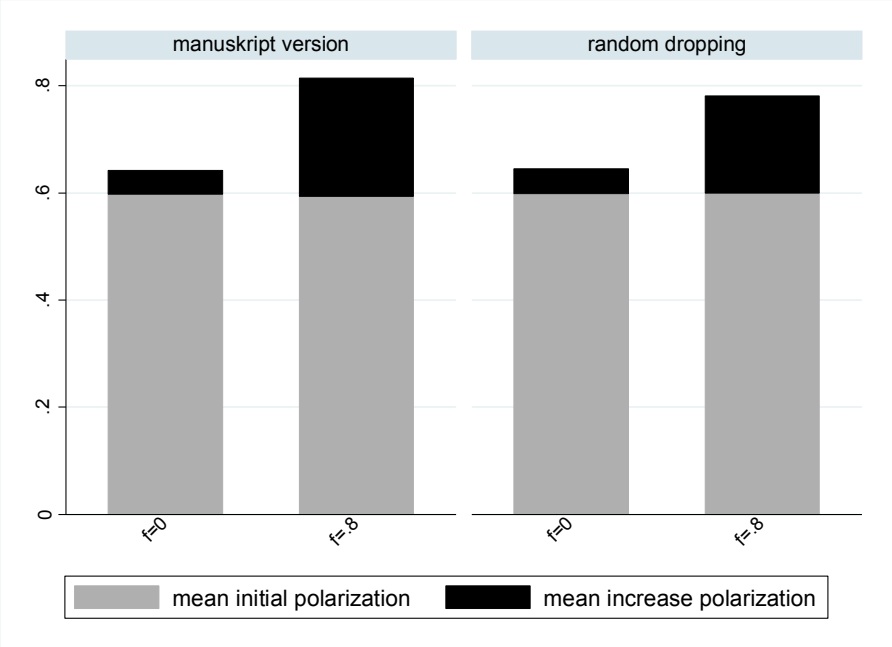
It turned out that the model predictions are very similar under random dropping and recency-based dropping. The results of the runs with maximal faultline strength (f=1) did not differ from those reported in the paper. In the paper version, 49.6 % (see Figure 2 in the paper) of the runs ended in a perfect split. In the model with random dropping, 55 % of the runs ended with this result. The difference is not significant (z=1.71).

Also the simulation runs with a faultline strength below 1 lead to similar results as in the paper. First, all runs ended in consensus, confirming a core message of the paper (see Proposition 2). Second, the maximal level of *polarization* reached in a run was higher when the faultline was stronger (see Figure 16). Again the differences between the results of the two models are not significant (t=-0.69)

[2] This has been suggested by one of the reviewers. We are very thankful for this advice.

FIGURE 16

Recency-based dropping vs. random dropping of arguments



Furthermore, also with random dropping consensus was reached faster in the strong faultline condition (*f*=.8) compared to the weak faultline condition (*f*=0). This is the same effect as in the paper version of the model. However, it turned out that with random dropping, this effect is significantly stronger than recency-based dropping (t-value of interaction between faultline strength and kind of dropping procedure was -13.26). This difference is plausible, since random dropping includes an additional random process in the model dynamics. This makes coordination on the same vector of arguments more difficult.

To sum up, implementing random dropping did not crucially affect the model results. The core effects of the paper could be replicated also with random dropping. One of the effects was even stronger with random dropping. On the other hand, random dropping increased the average length of the runs significantly because reaching perfect consensus is more difficult when an additional random process is implemented in the model.

We believe, however, that random dropping has implausible implications that the paper version of the model does not have. With random dropping each argument has the same dropping probability, no matter whether it is the most recent one or a very dated one. The probability that this happens equals in each interaction $1/(S+1)$. With $S=4$ this is 20% and thus very high. Recency, however, implies that it takes some time until agents drop a new argument. We therefore think that recency is more plausible and feel that it is the better assumption.

Based on suggestions of the reviewers, we also experimented with another dropping procedure, balancing-based dropping. According to cognition theories (Heider 1967), individuals strive for balanced cognitions in the sense that they prefer to consider cognitions relevant that do not contradict each other. This suggests that individuals tend to disregard those arguments that are not in line with their current

opinion. For instance, agents with a very positive (negative) opinion might tend to drop a con (pro) argument.

We implemented this in the following way. Whenever an agent adopted a new argument the computer counts the number of *pro* arguments that are relevant for this agent. Based on this number, the computer calculates the probability that one of the pro arguments is dropped. The probability that a pro argument is dropped is higher when few pro arguments are relevant. Table 1 illustrates the calculation of the dropping probabilities under the assumption that agents base their opinion on 4 arguments (S=4). Note that S+1 arguments are relevant at this moment (S old arguments plus the new argument). When only one pro argument is relevant then this argument is dropped with 99.9% probability. The more pro arguments are relevant, the smaller the probability that a pro argument is dropped.

TABLE 1

Probability of dropping a pro argument in the balancing-based dropping procedure

Number pro arguments (max=S+1)	Probability that a pro argument is dropped
0	0
1	.99
2	.66
3	.33
4	.01
5	1

Then, the computer program picks a random number between zero and one and compares it with the probability of dropping a pro argument. If the random value is smaller than the probability then a pro argument is dropped. Otherwise, a con argument is dropped. When there are several pro or con arguments that can be dropped, then the computer picks one of them randomly.

We expected that balancing-based dropping would lead to more extreme opinions. It implies that agents with a nonzero opinion will most likely move towards more extreme opinions because they tend to drop arguments in such a way that they consider either only pro or only con arguments. Over time, moderate opinions will, thus, die out. Depending on the initial distribution of opinions, two subgroups with opposing opinions will form or all agents will coordinate on the same pole of the opinion scale.

To test this expectation, we conducted simulations under a condition where faultline theory expects only weak subgroup polarization and where we found only weak polarization in the original version of the model (model where the least recent argument is dropped). We assumed a weak faultline (f=0), weak homophily (h=1) and that the opinion is initially unrelated to the demographic attributes (w=.5). We conducted 500 independent simulation runs.

It turned out that a central finding of the paper can be replicated also with balancing-based dropping: we found that all runs ended in perfect consensus. However, all runs with balancing-based dropping ended with consensus on one of the extremes of the opinion scale. In the original model version with recency-based

dropping, the opinion value of the consensus was normally distributed (average=.006; sd=.532). Furthermore, we found a very strong tendency for short-term polarization with balancing-based dropping. On average, the maximal value of polarization that was reached in the 500 runs was .92 in the balancing-based dropping version. In the manuscript version, this was only .25, what is significantly less (t=107.39). Note that the average initial degree of polarization was in both conditions .20. Thus, in the manuscript version, there was only very weak subgroup polarization. With balancing-based dropping, however, we found in the majority of the runs subgroup polarization in the short term.

In sum, the experiment revealed that balancing-based dropping is an additional mechanism that can generate subgroup polarization in the short run. In the experiment, we imposed conditions under which Lau and Murnighan's theory does not predict subgroup polarization. Nevertheless, we found very strong polarization tendencies. The contradiction between Lau and Murnighan's prediction for weak faultlines and the results that we have obtained with the assumption of balancing-based dropping has led us to discard balancing-based dropping. The key objective of our paper is to study the implications of the presence of crisscrossing agents under the assumptions that Lau and Murnighan have spelled out in their theory of faultlines. This requires that we build our analysis on a model that is capable to replicate the fundamental implications of their theory to begin with. A fundamental prediction of their theory is that faultline strength fosters polarization. However, our results suggest that balancing-based dropping practically imposes polarization as unique model outcome, regardless of the strength of demographic faultlines. Accordingly, a model that assumes balancing-based dropping would not be suitable for the task at hand.

6. Selection of interaction partners

Our model of work team dynamics is based on the assumption that team members tend to interact with those colleagues who have similar demographic characteristics and who hold similar opinions. The selection of demographically similar colleagues is a core assumption of Lau and Murnighan's theory (Lau and Murnighan 1998). However, the authors did not include the assumption of opinion-based selection in their theory. We argued in the main paper that both dimensions of selection should be included in our formal model because empirical research found support for both mechanisms (Byrne 1971; McPherson, Smith-Lovin and Cook 2001).

In the following, we study to which degree the inclusion of opinions in the selection of interaction partners affects the predictions of our model. For this purpose, we developed two new versions of the model and compared the implications of the new versions with the results that we presented in the main paper.

The first new version of the model assumes that agents select interaction partners only based on demographic attributes. Technically, we simplified equation 2 from the main paper and arrived at equation 2a:

$$sim_{i^*,j} = \frac{1}{2 \cdot (D+K)} \left(\sum_{d=1}^{D} 2 - \left| c_{i,d} - c_{j,d} \right| + \sum_{k=1}^{K} 2 - \left| o_{i,k} - o_{j,k} \right| \right) \qquad (2)$$

$$sim_{i^*,j} = \frac{1}{2D}\left(\sum_{d=1}^{D} 2 - \left|c_{i,d} - c_{j,d}\right|\right) \qquad (2a)$$

An evident consequence of this adjustment is that the model implies new equilibria in teams with a maximally strong faultline (f=1). These teams consist of two maximally different subgroups. According to equations 2a and 3 (see main paper), this implies that there is a zero probability that members of distinct subgroups will select each other as interaction partners. It is therefore possible that the system reaches a state where both demographic subgroups have developed a separate subgroup consensus with only small opinion differences between the subgroups. According to equation 2a such constellations can be equilibrium. In contrast, such a constellation is not in equilibrium in the model version where also opinions affect the selection of interaction partners (equation 2), because opinion similarity leads to interaction between the members of the demographic subgroups and triggers opinion convergence. The changes in the selection of interaction partners do not affect the equilibrium conditions for work teams that include demographic crisscrossing (f<1).

The second new version of the model assumes pure opinion-based selection of interaction partners. Technically,

$$sim_{i^*,j} = \frac{1}{2K}\left(\sum_{k=1}^{K} 2 - \left|o_{i,k} - o_{j,k}\right|\right) \qquad (2b)$$

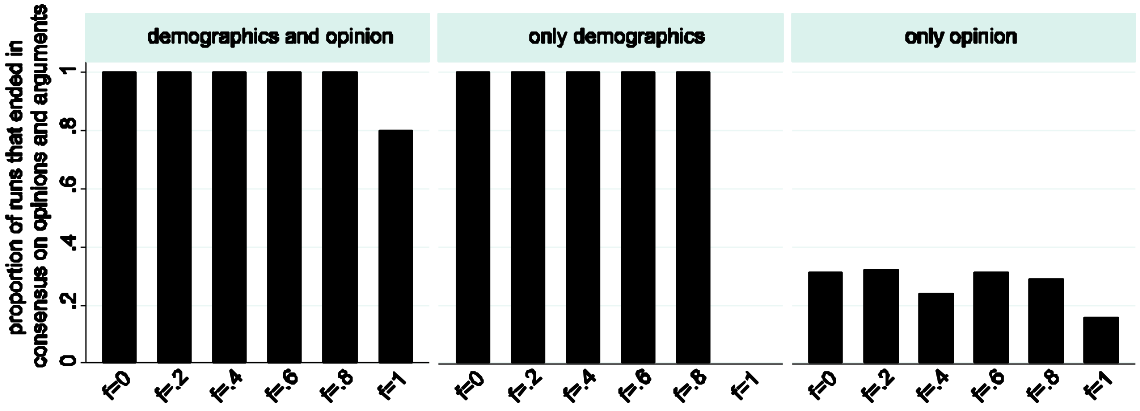
This adjustment implies that there are new equilibria. In particular, equation 2b implies that even teams which comprise crisscrossing agents but which consist of two groups with maximally different opinions will never overcome opinion differences. This, however, is not surprising because this model presumes that demographic differences do not influence the behavior of the agents.

In order to test whether the inclusion of opinions in the selection of interaction partners has critical effects on the findings of the main paper we conducted several simulation experiments with both new model versions and compared the results with otherwise identical runs (N=20, K=1, S=4, P=C=10) that were presented in the paper.

The first experiment tested the effects of demographic faultline strength (f). For this purpose, we focused on a condition where the original model version with selection based on demographic *and* opinion similarity found faultline effects. In particular, we assumed strong homophily (h=5) and strong initial congruency (w=.8). We experimentally varied the strength of the demographic faultline between f=0 and f=1 in steps of 0.2. For each new model version, we conducted 100 independent simulation runs per experimental condition.

FIGURE 17

Proportion of runs that ended in perfect consensus over *f* by selection procedure (*w*=.8, *h*=5)



Focusing on the first core finding of the main paper, Figure 17 informs about the proportion of simulation runs that ended with perfect consensus on opinions and arguments. As the left panel of the figure shows, we report in the main paper that all teams which comprised crisscrossing agents (*f*<1) arrived at consensus. The figure shows the same effect for the model version with demographics-based selection (see panel in the center of Figure 17).

In contrast to the results from the main paper, we found for the model version with demographics-based selection that not a single simulation run under perfect faultline strength (*f*=1) ended with consensus on opinions and arguments. As we have indicated above, this is because this version of the model implies that agents which belong to different demographic subgroups do not interact. In other words, each demographic subgroup develops independently a separate subgroup consensus. Hence, it is very unlikely that the two subgroups happen to develop a subgroup consensus based on the same vector of arguments.

FIGURE 18

Number of runs that ended with respective degree of subgroup polarization; model version with selection based on demographic attributes only (*f*=1, *w*=.8, *h*=5)

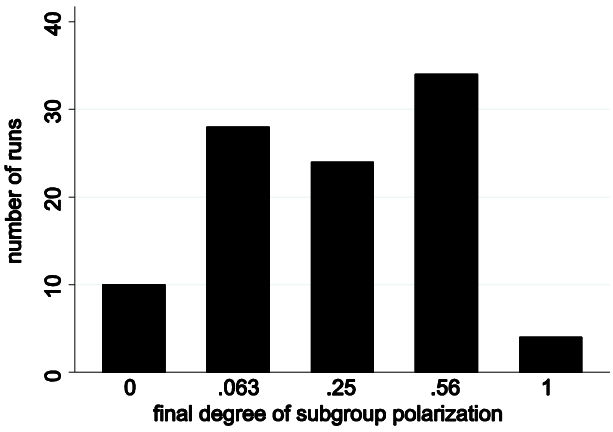


TABLE 2

Opinion distance between subgroups implied by respective degree of subgroup polarization (only if *f*=1)

Final degree of subgroup polarization	Final distance of opinions between subgroups
0	0
0.063	0.5
.25	1
.56	1.5
1	2

Showing this difference between the two model versions in more detail, Figure 18 displays the degree of subgroup polarization that we measured at the very end of the simulation runs for the condition with maximal faultline strength (*f*=1) with the model with demographics-based selection only. Table 2 represents the opinion distance between members of the different demographic subgroups that follows from the respective degree of subgroup polarization. As Figure 18 shows, ten runs ended with a subgroup polarization of zero. In these runs, the two demographic subgroups happened to develop a subgroup consensus on the same opinion value. In twenty-eight simulation runs, the final opinion differences between the two demographic subgroups was 0.5. This implies a final degree of subgroup polarization of 0.063. There were twenty-four runs with a final opinion distance of one scale point between the subgroups and thirty-four runs with a final opinion distance of 1.5 scale points. Only four runs ended in maximum opinion polarization.

In short, in the experiments with the model version that assumes demographics-based selection we could replicate the core finding of the main paper: all teams that comprised crisscrossing agents found consensus in the long run. Furthermore, our experiment showed that the model with demographics-based selection tends to reach different equilibria for teams with no crisscrossing agents.

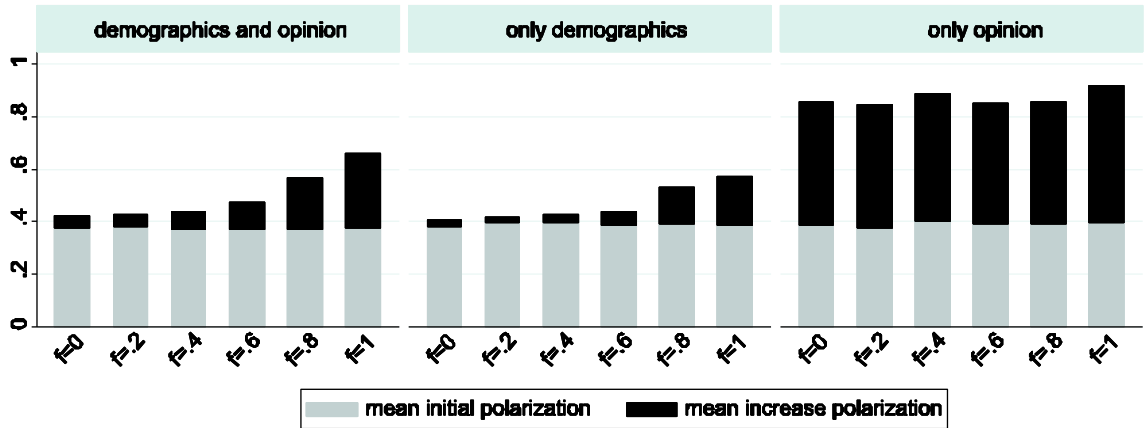
The right panel of Figure 17 focuses on the model version with opinion-based selection. One can see that fewer runs ended in consensus, compared to the model version of the main paper (left panel). This is because crisscrossing agents can not help overcoming group splits if demographic attributes are not considered in the selection of interaction partners. Likewise, this explains why the figure shows no effect of faultline strength on the number of runs that ended in consensus.

Figure 19 visualizes subgroup polarization effects of demographic faultlines in the different model versions. The left panel shows the effect reported in the main paper. The center panel shows that this effect could be replicated also with demographics-based selection of interaction partners. To be more precise, in the model version with demographics-based selection the effect of faultline strength on subgroup polarization is significantly different from zero (t=12.49) but it is significantly weaker than in the model version with selection based on demographics and opinion (t=-3.84). We think that the faultline effect is weaker because we focused here on teams with a high initial congruency (w=.8). In these teams demographic similarities overlap to a large degree with opinion similarity. As a consequence, the

effect of demographic differences is amplified if opinions are considered in the selection of interaction partners.

FIGURE 19

Average maximum opinion polarization over *f*, by model version (*w*=.8, *h*=5)

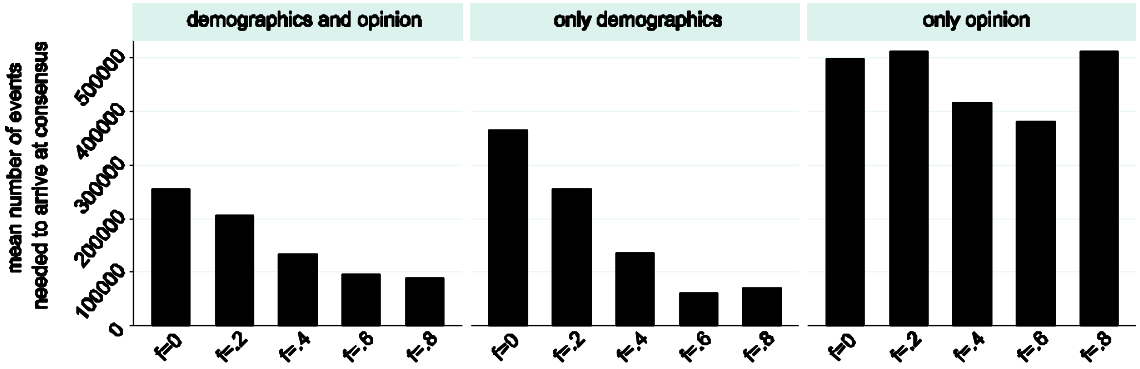


The right panel of Figure 19 shows that there was no significant effect of faultline strength when opinion-based selection was considered only (*t*=1.17). This is not surprising because in this version of the model demographic differences do not influence agents' decisions. However, the figure also shows that subgroup polarization increased much more in simulation runs with opinion-based selection than in the runs with the model versions of the main paper (*t*=40.43). This is because in the model version with selection based on demographics and opinions demographic similarities lead to frequent interaction between agents who hold similar demographic attributes but relatively dissimilar opinions. These interactions likely lead to more moderate opinions. The model version where selection is only based on opinions, however, implies that interactions between agents with dissimilar opinions are unlikely. As a consequence, agents are less often exposed to counter arguments. This, in turn, results in stronger subgroup polarization tendencies.

Figure 20 illustrates that we could replicate the counter-intuitive effect of faultline strength on the time until convergence by using just selection based on demographic similarity. In the main paper, we found that teams with a strong demographic faultline arrived at consensus faster than teams with a weak faultline (see left panel of Figure 20). The center panel shows that this effect was even stronger for the model version with demographics-based selection (*t*=-7.05), supporting our reasoning that demographic crisscrossing causes the counter intuitive effect. Our reasoning finds further support in the right panel of Figure 20, which shows that there was no significant effect of faultline strength (*t*=0.45) in the model where demographic differences played no role.

FIGURE 20

Average number of events until the teams arrived at equilibrium over *f*, by model version (*w*=.8, *h*=5)



The second simulation experiment tested whether we could replicate the effects of initial congruency *w* on subgroup polarization. For this purpose, we focused on those conditions where the model version with selection based on demographics and opinions found congruency effects. In particular, we assumed strong homophily (*h*=5) and a strong demographic faultline (*f*=1). We experimentally varied the initial congruency between *w*=.5 and *w*=.9 in steps of 0.1. For both new model versions, we conducted 100 independent simulation runs per experimental condition.

FIGURE 21

Average maximum opinion polarization over *w*, by model version (*f*=1, *h*=5)

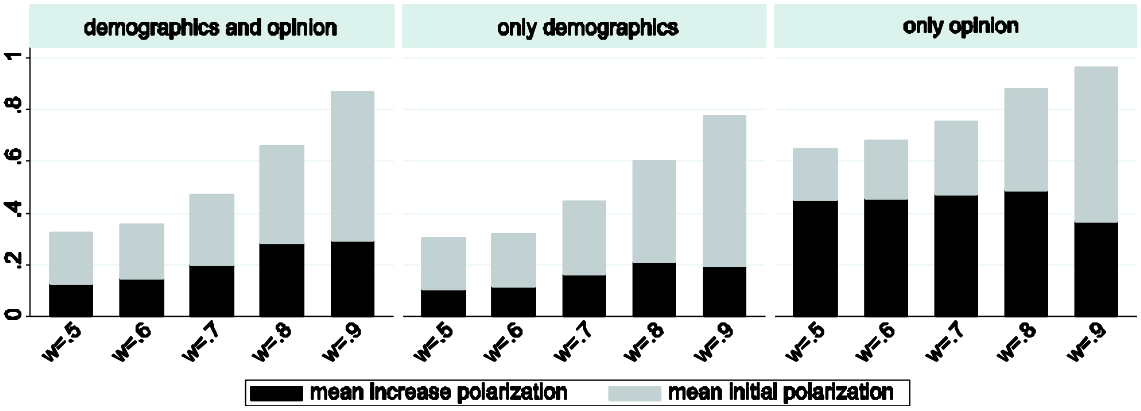
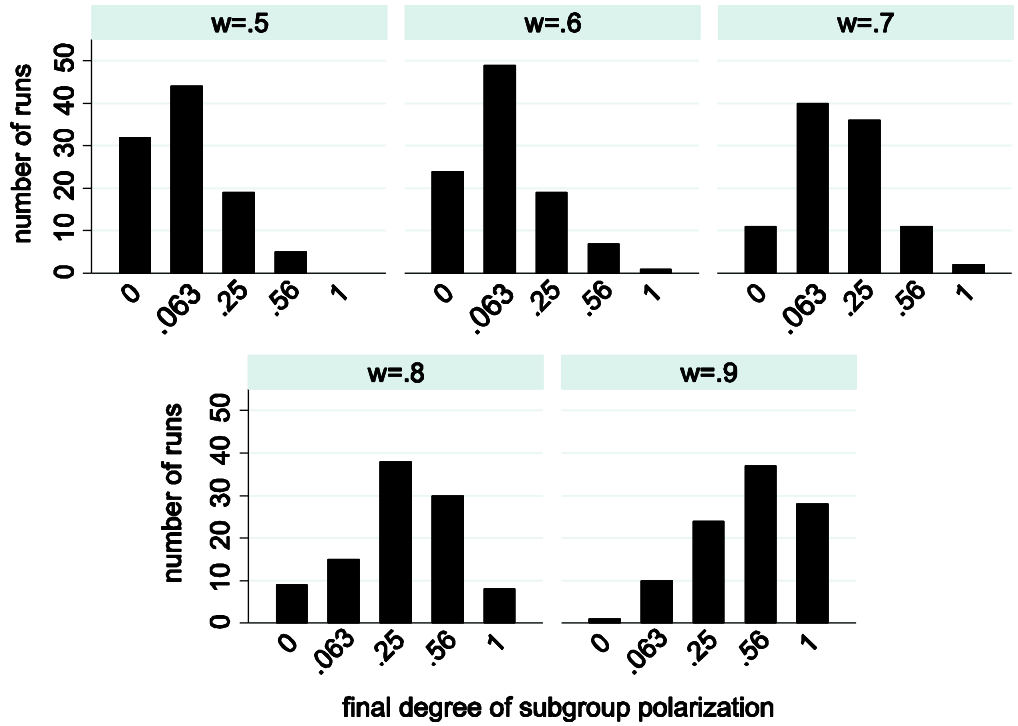


Figure 21 informs about the effects of initial congruency on subgroup polarization. The left panel shows the effect that we reported in Figure 7 of the main paper: stronger initial congruency leads to a stronger increase in polarization. The center panel shows that we found a similar effect for the model with demographics-based selection. This effect was significant (*t*=5.31) but it was also significantly weaker (*t*=-3.05) than the effect found with the model version of the main paper.

Figure 22 focuses on the model version with demographics-based selection. The figure depicts for each experimental condition the distribution of the degree of subgroup polarization reached in equilibrium (see Table 2 for the interpretation of the polarization values). One can see that the simulation runs under weak initial congruency tended to reach equilibria with mainly small opinion differences between the demographic subgroups. With a strong initial congruency, however, more runs ended with bigger opinion differences.

FIGURE 22

Number of runs that ended with respective degree of subgroup polarization by *w*; model version with selection based on demographic attributes only (*f*=1, *h*=5)



In sum, Figures 21 and 22 show that the model version with demographics-based selection implies that initial congruency leads to more subgroup polarization, the same effect that we found with the model where selection is based on opinions and demographic attributes. Interestingly, Figure 21 shows that the effect is weaker for the model with demographics-based selection only. This suggests that the effect of initial congruency that we report in the main paper is the result of two processes. First, initial congruency leads to an initial opinion bias within both subgroups. As we explained in detail in the main paper, this bias forms the basis for the reinforcement of opinions in the process of argument exchange and, therefore, leads to opinion polarization.

Second, when selection is based on both demographics *and* opinions, strong initial congruency has an additional effect: it decreases the probability that agents with relatively extreme opinions interact with those team members who hold the same demographic attributes but different opinions. As a consequence, these agents likely develop even more extreme opinions. If selection is based only on demographics, however, extreme agents likely interact with those members of their subgroup who hold moderate opinions and therefore also develop more moderate opinions. As a consequence, there is weaker subgroup polarization.

The right panel of Figure 21 depicts the effect of initial congruency on subgroup polarization when selection is based only on opinions. Strikingly, the increase in subgroup polarization is under all experimental conditions stronger than in the other versions of the model. This is caused by the fact in this model version agents select interaction partners only based on opinion similarity and therefore interact less likely with those team members who are demographically similar but hold dissimilar opinions. Such interactions likely provide agents with counter arguments and

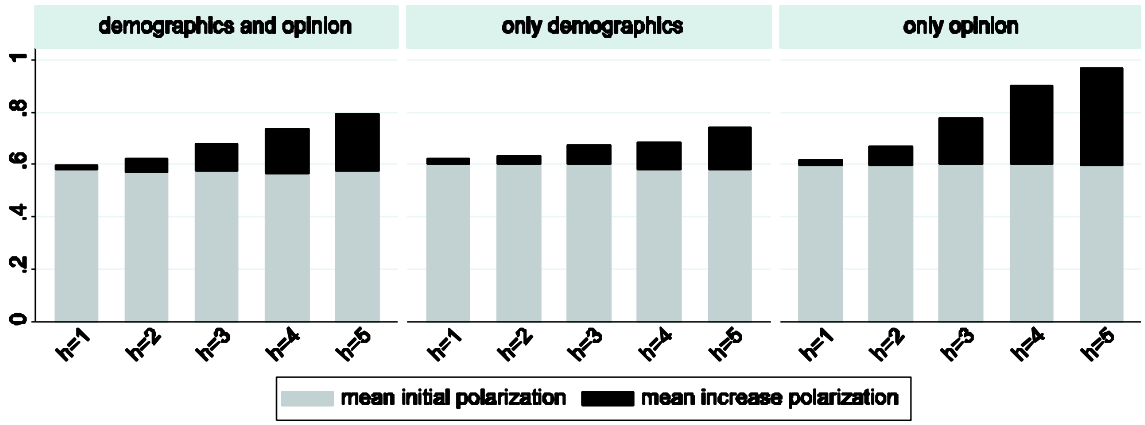
therefore lead to a decrease in subgroup polarization. As a consequence, we found less subgroup polarization for the model versions where selection is based on demographic similarity.

The black section of the bar for w=0.9 shows that there was a relatively weak increase in subgroup polarization in those runs which started with a very high initial congruency. As the size of the complete bar shows, however, this is due to a ceiling effect. Under this experimental condition, the initial degree of subgroup polarization (see the gray section of the bar) was so high that an increase in subgroup polarization like in the conditions with weaker congruency was logically impossible. For the remaining experimental conditions (w<.9), the right panel of Figure 21 shows that there was no effect of congruency on the increase of subgroup polarization (t=1.15).

The third simulation experiment was conducted to test whether assuming different selection procedures changes the effects of homophily strength (h). For this purpose, we focused on those conditions where the model version with selection based on demographics and opinion implies homophily effects. In particular, we assumed strong initial congruency (w=.9) and a strong demographic faultline (f=.8). We experimentally varied homophily strength between h=1 and h=5 in steps of 1. For both new model versions, we conducted 100 independent simulation runs per experimental condition.

FIGURE 23

Average maximal opinion polarization over h, by model version (f=.8, w=.9)



The left panel of Figure 23 reports the effect of homophily strength on subgroup polarization that we reported in the main paper (Figure 8 of main paper): stronger homophily leads to a stronger increase in subgroup polarization. The other two panels show that the same effect was found with the other two versions of the model. However, compared to the original version of the model, the homophily effect is significantly weaker (t=-4.48) under demographics-based selection and significantly stronger under opinion-based selection (t=10.83).

Thus, the strength of the homophily effect depends on the relative impact of opinions in the selection of interaction partners. This is because the exchange of arguments leads to intensified opinions and subgroup polarization only when agents tend to interact with team members who hold similar opinions. The demographic attributes are correlated with the opinion (w=.9), but this correlation is not perfect. As

a consequence, when demographic attributes have a relatively weak influence on the selection of interaction partners, homophilious selection of interaction partners results more likely in interactions of agents with similar opinions and, therefore, in stronger subgroup polarization.

FIGURE 24

Average number of events until the teams arrived at equilibrium over *h*, by model version (*w*=.9, *f*=.8)

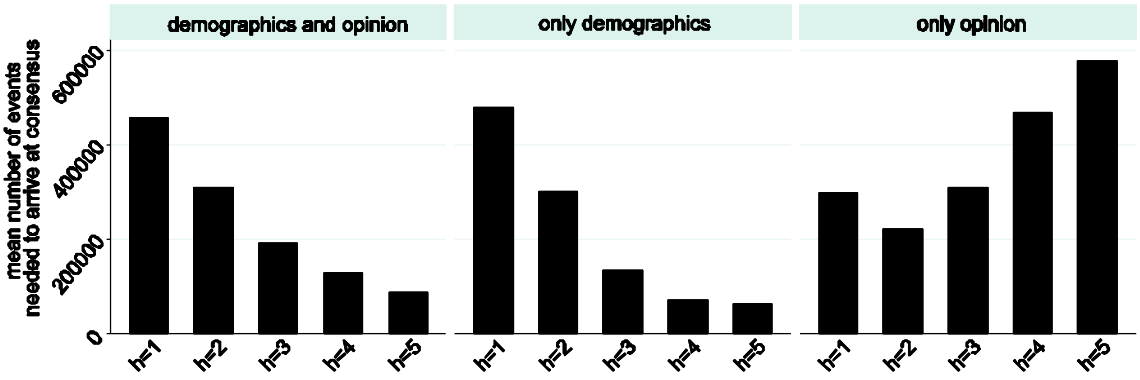


Figure 24 focuses on the counter intuitive effect of homophily strength on the length of the convergence process in teams which comprise very few crisscrossing agents. The left panel shows the effect discussed in the main paper: strong homophily speeds up the convergence process of opinions. The center panel shows that the same effect was found for the model with demographics-based selection. It turned out the homophily effects of the two model versions did not differ significantly from each other (t=-1.51).

Strikingly, we found the opposite effect of homophily strength for the model with opinion-based selection. The right panel of Figure 24 shows a strong increase in the number of events as homophily increases. This is the effect that one would expect intuitively, as strong homophily leads to less frequent communication between agents with different opinions. The effect is significant even though only 299 runs could be used for this analysis, because the remaining 201 runs ended in a perfect split into two groups, an equilibrium which is possible because demographic similarities between agents are ignored in this version of the model (see discussion of equilibria above).

The reversed effect of homophily strength in the model with opinion-based selection supports our explanation of the counter intuitive effect. In the main paper, we argued that agents coordinate quicker on the same vector of arguments when members of the demographic subgroups interact relatively seldom with crisscrossing agents. This is the case when there are only few crisscrossing agents (strong faultline) and if homophily is strong. The fact that the counter intuitive effect disappears when selection is purely based on opinion similarity demonstrates that demographic crisscrossing plays a critical role in the process that generates the counter intuitive effect because the demographic attributes are the basis of partner selection in interaction.

Summarizing the results of the three simulation experiments, we found that the effects of faultline strength, initial congruency and homophily strength could be replicated with model version which assumes only demographics-based selection of interaction partners. Most importantly, we found also with this model version that all

simulated work teams that consisted of crisscrossing agents arrived at a consensus in the long run even though there was significant subgroup polarization in the short term. Furthermore, we could replicate the counter intuitive effect of faultline strength and homophily strength with demographics-based selection. This demonstrates that our assumption that the selection of interaction partners is based on demographic similarity and opinion similarity is innocent in the sense that it does not critically affect the prediction of the model.

We did find that core findings of the paper could not be replicated with the model version where selection is based on opinion similarity only. This, however, should not be considered as a weakness of our model. In contrast, this finding supports our argumentation that the effects which we report in the main paper are caused by demographic faultlines and crisscrossing agents.

References

Brown, G. D. A., N. Chater. 2001. The Chronological Organization of Memory: Common Psychological Foundations for Remembering and Timing. Christoph Hoerl and Teresa McCormack, eds. *Time and memory: Issues in Philosophy and Psychology*. Oxford University Press. Oxford,

Brown, G. D. A., I. Neath, N. Chater. 2007. A temporal ratio model of memory. *Psychological Review.* **114**(3) 539-576.

Byrne, D. 1971. *The Attraction Paradigm*. Academic Press. New York, London.

Carley, K. 1991. A Theory of Group Stability. *American Sociological Review*. **56**(3) 331-354.

Colson, E. 1954. Social Control and Vengeance in Plateau Tonga Society. *Africa: Journal of the International African Institute.* **23**(3) 199-212.

Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences.* **24**(1) 87-185.

Evans-Pritchard, E. E. 1939. Introduction. J.G. Peristiany, eds. *The Social institutions of the Kipsigis*. Routledge. London, XIX-XXXIV.

Flap, H. 1988. *Conflict, loyalty and violence: The effects of social networks on behaviour*. Peter Lang. Frankfurt am Main.

Heider, F. 1967. Attitudes and Cognitive Organization. Martin Fishbein, eds. *Readings in Attitude Theory and Measurement*. John Wiley and Sons, Inc. New York, London, Sydney, 39-41.

Kamada, T., S. Kawai. 1989. An Algorithm for drawing general undirected graphs. *Information Processing Letters.* **31**(1) 7-15.

Lau, D. C., J. K. Murnighan. 1998. Demographic Diversity and Faultlines: The Compositional Dynamics of Organizational Groups. *Academy of Management Review.* **23**(2) 325-340.

Mark, N. P. 2003. Culture and Competition: Homophily and Distancing Explanations for Cultural Niches. *American Sociological Review* **68**(3) 319-345.

McFarland, D., S. Bender-deMoll. 2007. "SoNIA (Social Network Image Animator)."

McPherson, M., L. Smith-Lovin, J. M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology.* **27**(415-444.

Ross, E. A. 1920. *The Principles of Sociology*. Century Co. New York.

Sederberg, P. B., M. W. Howard, M. J. Kahana. 2008. A Context-Based Theory of Recency and Contiguity in Free Recall. *Psychological Review.* **115**(4) 893-912.