The diversity puzzle

Mäs, Michael

II. Negative influence and demographic faultlines[3]

Abstract
In this chapter, we will discuss the first approach to the polarization problem. We will develop a formal model that includes the assumption of negative influence and demonstrate that this model can indeed explain opinion polarization. Recently, several models that assume negative influence have been developed. We will contribute to this literature and show how demographic attributes might interfere with opinion dynamics generated by negative influence. In particular, we will discuss Lau and Murnighan's work on the effects of demographic faultlines. Lau and Murnighan proposed that polarization is the more likely the stronger demographic differences between group members correlate across various dimensions.
Computational experiments will demonstrate that the central claims of Lau and Murnighan's theory are consistent with the model. Furthermore, we will show that the model highlights a new structural condition that may give managers a handle to temper the negative effects of strong demographic faultlines. We will call this condition the *timing of contacts*. Computational analyses will reveal that negative effects of strong faultlines critically depend on *who* is *when* brought in contact with *whom* in the process of social interactions in the team. More specifically, we will demonstrate that faultlines have hardly negative effects when teams are initially split into demographically homogeneous subteams that are merged only when a local consensus has developed.

II.1. Introduction

The demographic diversity of a work team is seen as one of the major determinants of its performance. While managers as well as diversity researchers emphasize that diverse teams benefit from their large variety of social and human capital resources, (e.g. Chatman et al. 1998), many studies also highlight that this benefit comes at a potentially large cost. Diverse teams may be less socially cohesive than homogeneous teams and social cohesion, in turn, can be an important antecedent of performance (e.g. Jehn and Bezrukova 2004; Jehn, Northcraft and Neale 1999). Milliken and Martins concluded that "diversity thus appears to

[3] This chapter has been published together with Andreas Flache (first author) under the title "How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams" in Computational and Mathematical Organization Theory (2008:14/1). The article is freely available online (www.springerlink.com/content/0l03554v37118683)

be a double-edged sword" (Milliken and Martins 1996: 403), reflecting the mixed research evidence that produced both positive as well as negative effects of demographic diversity on team performance (for comprehensive reviews about theoretical and empirical research see: Bowers, Pharmer and Salas 2000; Milliken and Martins 1996; Pelled 1996; Stewart 2006; Webber and Donahue 2001; Williams and O'Reilly 1998).

The mixed effects of diversity have been attributed to the simultaneous operation of both positive effects on a team's human capital and negative effects on team cohesion (Reagans and Zuckerman 2001). However, Lau and Murnighan (1998; 2005) have questioned that demographic diversity is necessarily a threat for team cohesion. In Lau and Murnighan's view, cohesion suffers in a diverse group only to the extent that the distribution of demographic attributes across group members generates a *strong demographic faultline*. „Group faultlines increase in strength as more attributes are highly correlated, reducing the number and increasing the homogeneity of resulting subgroups. In contrast, faultlines are weakest when attributes are not aligned and multiple subgroups can form" (Lau and Murnighan 1998: 328). To give an example, a faultline is strong in a team consisting of two Caucasian, highly educated women and two African-American men with low level of education. In this case, all three demographic dimensions along which team members differ (race, sex, educational level) split the team along the same line. The faultline would be weaker if, for example, the two highly educated team members would be one man and one woman. The core prediction (see Lau and Murnighan 1998: 331) is that stronger demographic faultlines increase the potential for dissensus between team members and thus put performance under pressure. The theory also implies that the direct effects of diversity on performance are positive due to larger human and social capital in diverse teams. Subsequent empirical research has provided partial support for the proposed negative effects of strong faultlines (e.g. Lau and Murnighan, 2005; Molleman, 2005; Thatcher, Jehn and Zanutto, 2003) and has identified organizational design features that interact with the effects of faultline strength on team outcomes, such as team empowerment strategies or the use of knowledge management systems in team learning (Gibson and Vermeulen 2003)

In a nutshell, the theory of faultlines (Lau and Murnighan 1998: 332-333) is based on two main mechanisms: First, it is assumed that team members prefer to interact with those team members who are similar with respect to a salient demographic attribute. This corresponds to the prominent notion that homophily (Lazarsfeld and Merton 1954) is a strong force in social interactions (McPherson, Smith-Lovin and Cook 2001). Which

demographic attribute is salient in a certain work situation changes from situation to situation. Secondly, if actors choose to interact they are assumed to exert social influence (Festinger, Schachter and Back 1950) upon each other. Lau and Murnighan seem to assume furthermore that demographically similar actors tend to hold similar opinions even prior to interaction. Based on psychological research on opinion formation in groups (Isenberg 1986; Vinokur and Burnstein 1978) the authors propose that interactions between demographically similar actors reinforce the opinions they hold prior to interaction and, in the process, increase dissensus with demographically dissimilar group members. In other words, demographically similar interaction partners become more convinced of their respective opinions, because they tend to agree in opinion and they learn new arguments that are in line with their opinion. But only in teams with a strong faultline, the same team members interact again and again so that the opinions of the demographic subgroups become increasingly distinct at the expense of lower cohesion of the team as a whole. By contrast, in teams with weak faultlines, group members repeatedly interact with colleagues with a large variety of demographic characteristics and opinions, such that no self reinforcing dynamic towards an emergent subgroup split can develop.

While applications of faultline theory clearly demonstrate its relevance for both researchers and managers, neither Lau and Murnighan's original elaboration nor subsequent extensions have fully explicated the mechanisms that may underlie faultline effects. Both the transparency of Lau and Murnighan's theory as well as its deductive power can benefit considerably from a formal deduction of their central claims and an analysis of the precise combination of assumptions that is needed to derive them. In a previous paper we proposed a formal model of faultline effects that allows to generate hypotheses in line with previous informal reasoning (Flache and Mäs 2008b). We could also show that it is not even necessary to assume that opinions and demographic characteristics of team members are correlated already prior to interaction[4]. In the present paper, we move one step further and argue that the model also implies a remedy against negative effects of strong faultlines that has hitherto been overlooked in the literature. We propose that the effects of strong faultlines may critically depend on *who* is *when* brought in contact with *whom* in the process of social interactions in the team. More generally, it may depend on the *timing of contacts* between team members whether strong faultlines have negative effects on team cohesion. To be precise, we use "timing of contacts" here in the

[4] We explain this point below in our elaboration and discussion of the formal model.

sense of Moody's (2002) concept of "relationship timing". Broadly, relationship timing defines the sequence within which social interactions occur in given network of interactions. Consider for example an opinion formation process between three members of a work team, two of whom agree with each other and totally disagree with the third one. One possible timing of contacts might be that all three group members are brought together to discuss the issue. In this case, social influence occurs simultaneously in all three dyads in the network. Another sequence might be that only one of the two majority members discusses the issue with the minority member and after each meeting, the two majority members come together again. Obviously, in the first sequence the deviant might influence the positions of both other team members at the same time, while in the second sequence, he can directly influence only one of them, while the other one may bring his colleague "back into line" after each encounter with the deviant.

Effects of the timing of contacts on the outcome of group discussions have been demonstrated in experimental research by Kameda and Sugimori (1995). These authors manipulated the sequence within which in a group discussion minority members encountered majority members and found that this affected the chances for consensus in the overall group. More recently theoretical analyses have shown that the diffusion dynamics of, e.g., knowledge or diseases in social networks may critically depend upon the timing of network contacts (Gibson 2005; Moody 2002). For example, whether an infectious disease can spread in a chain A-B-C from A to C critically depends upon whether B was infected by A before or after being in contact with C. The idea that timing matters has not yet been theoretically elaborated for the study of opinion dynamics. However, we believe that the diffusion of opinions may be similarly affected by relationship timing than the diffusion of infection or information. The key reason why we expect the timing of contacts to be important for the group dynamics in diverse teams is the inherent path dependence of the process of social interactions between team members. For example, early contacts between group members who are strongly dissimilar both in terms of their opinions and their demographic characteristics may trigger negative and hostile interactions between the interactants. This, in turn, may lead them to adopt extreme positions on some issues. If these "radicalized" actors interact subsequently with demographically similar "friends", this may entail "bandwagon dynamics" in which the friends of the early conflict partners are socially influenced to adopt similarly extreme positions. The stronger are demographic faultlines, the more such a dynamic would project the demographic faultline onto an emergent faultline in the opinion space, with the result

that communication between team members and thus group cohesion and team performance may severely suffer. Clearly, this downward spiral might be avoided when contacts between team members are arranged in such a way that opposed "extremists" are initially isolated from each other and are instead exposed to interactions with demographically similar in-group members who are more moderate in their opinions. Then, the likely consequence is that initial extremists also become more moderate and initial moderates from different demographic subgroups move towards each other in the opinion space.

It may be a plausible idea that the timing of contacts modifies faultline effects, but Lau and Murnighan's original theory is not precise enough to generate testable predictions about the exact conditions under which this mechanism may work. We use and extend in the present analysis the formal model proposed by Flache and Mäs (2008b) to elaborate our reasoning why timing matters and under which conditions. In section 2, we describe the formal model and its extension to accommodate timing effects. Section 3 contains a description of the simulation experiments and results. In section 4, we discuss results and offer conclusions.

II.2. The Model

The model consists of four main elements, the formalization of the dynamics and elementary mechanisms of *social interactions and influence* between team members, the operationalization of *demographic faultlines*, the model of the *timing of contacts* and, finally, *aggregate outcome measures* that capture the dependent variables we are interested in.

II.2.1. The Social Interaction and Influence Dynamics

The main endogenous outcome variable of our model is the *distribution of work related opinions* in the team, because following previous work (Mason 2006: 234; Molleman 2005: 175-176; Pfeffer 1985) we assume that consensus at least on fundamental issues seems a necessary precondition for effective teamwork, while opinion polarization on these issues may be a major obstacle to good team performance. The theoretical assumptions of homophily and social influence identify a clear causal link between team cohesion, consensus on work related opinions and the strength of demographic faultlines in a team. Broadly, the stronger are faultlines in the team, the less likely it is that team members in different subgroups influence each other sufficiently to generate a consensus on work related opinions on the level of the team as a whole, and the more likely it is that the influence processes result in

polarization rather than consensus. At the same time, the combined assumptions of homophily and influence link the degree of consensus closely to the level of cohesion in the team. We assume that only when team members agree on important issues, they have good social relations with each other which, in turn, generates social cohesion.

With this approach, we deliberately exclude from our analysis variables which also may affect performance but which are not or at least much less directly causally related to faultline strength (Lau and Murnighan 1998), like the size of the team's pool of human and social capital.

We assume that the effects of faultlines on opinion polarization (and poor team performance) are generated by the interplay of the four fundamental social mechanisms *homophily*, *social influence, heterophobia and rejection*. According to homophily[5], the more similar two actors are with respect to salient opinions or demographic characteristics, the more they like each other and the more they interact (Brass et al. 2004; Byrne 1971; Harrison and Carroll 2002; Homans 1951; Kandel 1978; Lazarsfeld and Merton 1954; McPherson, Smith-Lovin and Cook 2001; Rogers and Bhowmik 1970). According to social influence, if two persons interact they adapt their opinions (Abelson 1964; Brass et al. 2004; Kerr and Tindale 2004). But homophily and social influence alone do not suffice to explain why groups with strong faultlines exhibit a tendency towards extreme and over time increasing opinion differences between a small number of opposed and demographically dissimilar factions in the team (cf. Early and Mosakowski 2000)[6]. To address this pattern with our model, we followed previous research and complemented the mechanisms of homophily and social influence with their negative counterparts of *heterophobia* and *rejection* (Flache and Macy 2006b; Jager and Amblard 2005; Kitts 2006; Macy et al. 2003; Rainio 1961a; 1961b; 1962; 1965; Salzarulo 2006). Heterophobia implies that if the dissimilarity of two actors exceeds a certain threshold then the actors do not like each other (Byrne, Clore and

[5] Note that the definition of 'homophily' which we use in this chapter does not correspond to the definition used in the chapters I and II. In this chapter, homophily refers to the *preference* of individuals to interact with similar others. In the previous chapters, however, it referred to the *tendency* to interact with similar others. Such a tendency can result from the preference to do so, but also from social influence during interaction (see chapter II).

[6] Axelrod (1997) proposed to add the assumption that social influence may be entirely cut off when actors disagree beyond a certain critical level. With this assumption, homophily and social influence can stabilize differences between subgroups (Axelrod 1997; Flache and Macy 2006b; Flache, Macy and Takács 2006; Hegselmann and Krause 2002; Weisbuch, Deffuant and Amblard 2005). However, this explanation is not readily applicable to work groups, where there is little room to entirely avoid social interaction with dissimilar others. Moreover, Axelrod's assumptions can at best explain why differences between subgroups persist over time, but not why groups may increasingly polarize in the course of team interaction, as described by Earley and Mosakovski (2000).

Smeaton 1986; Chen and Kenrick 2002; Pilkington and Lydon 1997; Rosenbaum 1986a; Rosenbaum 1986b; Smeaton, Byrne and Murnen 1989). Rejection states that actors have a tendency to change their attributes in a way to become more dissimilar to interaction partners they do not like (Abelson 1964; Kitts 2006; Salzarulo 2006; Tsuji 2002).

It is important to note that Lau and Murnighan do not directly assume rejection. They propose instead that increasing opinion differences between dissimilar actors result from a self-reinforcing dynamic that is triggered by an initial correlation between demographic attributes and opinions. We avoided this assumption, because what we aim to explain is that the strength of the demographic faultline leads to opinion polarization along this faultline. If we already assume in the model that demographic attributes are correlated with the opinions then it is not surprising that the model predicts exactly this as an outcome. In our previous work (Flache and Mäs 2008b), we could show that the model sketched here suffices to reconstruct the main regularities predicted by faultline theory. Hence, we argue that the assumption of an initial positive correlation between demographic attributes and opinions should be avoided in this context. However, our argument is purely theoretical. We do not claim that the dynamics that Lau and Murnighan describe do not occur in real work teams.

Finally, our model distinguishes between two types of attributes on which agents can differ and which define the level of similarity between agents. Demographic attributes on the one hand are fixed and can not be changed by the dynamics of social influence and rejection. On the other hand, opinions are flexible and are subject to social influence and rejection. Previous computational studies based on similar sets of assumptions have already demonstrated how demographic differences can lead to the emergence of cultural niches in demographic space such that demographically dissimilar actors also hold dissimilar or even radically opposing opinions (Macy et al. 2003; Mark 2003). However, these studies did not address the effects of faultline strength in the demographic distribution.

Technically, each of the N team members is represented as an agent i characterized by D *fixed* (a_{id}^{fix}) and K *flexible attributes* (a_{ik}^{flex}), where d and k refer to the d'th and k'th fixed and flexible attribute, respectively. The fixed attributes correspond to the demographic characteristics, the flexible ones represent the agent's work related opinions. For simplicity, we assume that demographic attributes and opinions are equally salient. Moreover, we focus on clearly distinguishable demographic attributes, expressed by the assumption that demographic attributes are dichotomous and can take either the value -1 or +1

($a_{id}^{fix} \in \{-1;1\}$). Opinions of the team members can instead vary continuously between -1 and +1 ($-1 \le a_{ik}^{flex} \le +1$).

A key assumption of the model is that the direction and strength of influence that an agent *i* imposes on an agent *j* does not depend directly on the opinion of *j*, but it is moderated by the sign and the strength of the interpersonal relation between *i* and *j*. To model the *interpersonal relations* between the team members we assume a directed graph where w_{ij} represents the weight of the corresponding relationship ($-1 \le w_{ij} \le +1$). If team member *i* has contact to team member *j* then the weight w_{ij} takes a nonzero value between -1 and 1. A positive weight reflects that *i* evaluates *j* positively, whereas a negative one represents a hostile relationship. If there is no contact between *i* and *j*, or *i* is indifferent between liking and disliking *j*, then the weight is 0.

Both the *K* flexible attributes and the weights of the relationships are endogenous and change in discrete time steps. In every time step, one team member is selected randomly with equal probability to update either his flexible attributes or weights. With probability 0.5, all weights of *i* are updated simultaneously. In the event that flexible attributes are updated, all flexible attributes are updated simultaneously.

Time is modeled in discrete steps. The duration of a simulation run is expressed in number of iterations. One iteration corresponds to *N* simulation steps to assure that on average each agent updates either his weights or his attributes once within an iteration. To be sure, given the asynchronous random updating of agents, an iteration does not encompass any particular length of time or synchronization of events (e.g. work days). Rather, one iteration indicates that *N* events have taken place in which agents have changed their opinions or weights.

Similar to previous models of social influence with continuous opinions (Abelson 1964; Hegselmann and Krause 2002), we assume that the change of team member *i*'s flexible attribute *k* is an aggregated result of the influences imposed by all other agents who exert influence upon *i*. Technically, the new value of the attribute, $a_{ik,t+1}^{flex}$ is obtained by adding to the old value a weighted sum of the pressures of all influential others. To model a somewhat gradual change of opinions, we also divide this weighted sum by 2. The pressure imposed by a single alter *j* "pulls" *i* towards *j*'s opinion if the weight w_{ij} is positive, and "pushes" *i* away from *j*'s opinion if the weight is negative. The magnitude of

this pressure is proportional to the distance in opinions between *i* and *j*, $a_{ik}^{flex} - a_{jk}^{flex}$. With only positive weights summing to one, this assumption would imply that the net pressure imposed on *i* moves the agent towards the weighted average of the opinions of all interactions partners. Equation 1 formalizes these assumptions.

$$a_{ik,t+1}^{flex} = a_{ik,t}^{flex} + \frac{1}{2C_t}\sum_{i \neq j} w_{ij}(a_{jk,t}^{flex} - a_{ik,t}^{flex}) \quad (1)$$

The C_t in equation (1) refers to the number of agents who are in contact with *i* at the moment influence takes place ($C_t \leq (N\text{-}1)$). We will discuss further below effects of interaction structures in which team members can interact only with a subset of other team members temporarily. To be precise, equation (1) only shows the principle model of influence. In the actual implementation, we apply a slight modification of the influence equation both to make sure that opinions do not go out of bounds and to smoothen the change of opinions when agents move towards the extreme ends of the opinion scale. Equations 1a and 1b fully specify opinion change.

$$\Delta a_{ik,t}^{flex} = \frac{1}{2C_t}\sum_{j \neq i} w_{ij}(a_{jk,t}^{flex} - a_{ik,t}^{flex}) \quad (1a)$$

$$a_{ik,t+1}^{flex} = \begin{cases} a_{ik,t}^{flex} + \Delta a_{ik,t}^{flex}(1 - a_{ik,t}^{flex}), \text{if } a_{ik,t}^{flex} > 0 \\ a_{ik,t}^{flex} + \Delta a_{ik,t}^{flex}(1 + a_{ik,t}^{flex}), \text{if } a_{ik,t}^{flex} \leq 0 \end{cases} \quad (1b)$$

The second key element of our model is the update of weights. Following previous work (Macy et al. 2003) we assume that the weight that agent *i* has towards an agent *j*, changes depending on the similarity between *i* and *j* in terms of both their demographic attributes and their opinions. More precisely, we assume that after updating, the weight adopts a level that is proportional to the current level of similarity. The new weight is negative if the average distance between *i* and *j* across all dimensions of demographic and opinion space exceeds one, i.e. half of the maximum average distance. If this average distance is exactly one, the weight is zero and otherwise it obtains a positive value. Technically,

$$w_{ij,t+1} = 1 - \frac{\sum_{d=1}^{D}\left|a_{id,t}^{fix} - a_{jd,t}^{fix}\right| + \sum_{k=1}^{K}\left|a_{ik,t}^{flex} - a_{jk,t}^{flex}\right|}{D + K} \quad (2)$$

II.2.2. Faultline strength

To disentangle the effects of the strength of demographic faultlines from effects of demographic diversity, we devised a method that allows for varying faultline strength and keeping diversity constant at the same time. More precisely, we generated different distributions of the fixed attributes in such a way that all fixed attributes were equally frequent (= all distributions generate equally diverse groups) but the correlation between the attributes differed between distributions (= the strength of the faultline differs).

Table II.1 shows our construction method for the prototypical case of a group with 20 members (N=20) who differ along three demographic dimensions (e.g. male / female, young / old, western / non-western ethnic background). Column 2 of the table shows that we constructed the first demographic variable (A_1) by alternately assigning the values -1 and 1 to the first $N/2$ agents beginning with the value 1 for agent 1. We did the same with the second $N/2$ agents but here we started with the value -1. The distribution of this variable is the same in all work teams.

We expressed the faultline strength by a parameter f that varies between 0.5 and 1, where f=0.5 corresponds to a situation where the demographic attributes are completely uncorrelated and f=1 imposes a perfect correlation between all demographic attributes. The first step in the construction is to impose the correlation between attribute A_1 and A_2 that corresponds to the given parameter value of f. To arrive at the values for attribute A_2, we assigned to the first $(100 \cdot f)$ % of the cases the same value as for attribute A_1. For example, for f=0.9, the first 90% of the agents (= the first 18 agents if N=20) hold the same value at attribute A_1 and A_2 (See the grey cells in column 3 of table II.1). To the rest of the agents we assigned on attribute A_2 the opposite value of what we assigned for attribute A_1.

To determine the values for attribute A_3 we used the same method with a small change. We first assigned to the first $(50 \cdot f)$ % of the cases the same value as for attribute A_1. Then we continued with the $(N/2+1)$th case and again assigned to the following $(50 \cdot f)$ % of the cases the same value as for attribute A_1. Again the rest of the cases got the opposite value than for attribute A_1. Thus for f=0.9 and N=20 the agents 1-9 and 11-19 hold the same value at attribute A_1 and A_3 (see column 4 of table II.1). This procedure makes sure, that the agents also hold at the attributes A_2 and A_3 in exactly $(100 \cdot f)$ % of all cases the same value.

Table II.1: Implementation of faultline strength

	***f* = 0.9**			***f* = 0.8**			***f* = 0.7**			***f* = 0.6**			***f* = 0.5**		
i	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
7	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	1
9	1	1	1	1	1	-1	1	1	-1	1	1	-1	1	1	-1
10	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1
13	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1
14	1	1	1	1	1	1	1	1	1	1	-1	1	1	-1	1
15	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	1	-1
16	1	1	1	1	1	1	1	-1	1	1	-1	1	1	-1	-1
17	-1	-1	-1	-1	1	-1	-1	1	-1	-1	1	1	-1	1	1
18	1	1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1
19	-1	1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	1
20	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1
Σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Correlations (*r*)

	A_2	A_3
A_1	.8	.8
A_2		.8

	A_2	A_3
A_1	.6	.6
A_2		.6

	A_2	A_3
A_1	.4	.4
A_2		.4

	A_2	A_3
A_1	.2	.2
A_2		.2

	A_2	A_3
A_1	.0	.0
A_2		.0

Table II.1 also reports the correlations between the three attributes. Note that for a given distribution all pairwise correlations between two of the three attributes are equal. The relationship between *f* and the correlation is: $r = -1 + 2f$. If *f* takes the value 1 then the three attributes are perfectly correlated (r=1) and the faultline strength is maximal. If *f* takes the value 0.5 then there is no relationship between the attributes (r=0). Thus the faultline has a minimal strength. Because of its better intelligibility we use the parameter *r* to describe the faultline strength in the following. At all levels of *f*, all variables are equally distributed in all teams.

The key advantage of our method is that it separates variation in faultline strength from variation in diversity. A more intuitive alternative approach could have been to assign attributes randomly with a given probability and a given correlation. However, for the relatively small groups we are interested in, that method would have produced considerable random variation in faultline strength between single realizations of distributions imposed

by the same level of *f*. Our deterministic approach excludes this source of random noise and thus allows us to focus in our computational experiments exclusively on effects of variation in *f*.

II.2.3. Timing of contacts

There is in principle an infinite number of ways how the sequence of who is when brought into contact with whom in a work team can be manipulated. For the sake of simplicity, we decided to focus upon *three ideal typical forms of timing* that we expect to shape the effects of faultline strengths in clearly different ways. The *first* form is the baseline scenario: *no timing*. Under "no timing" there is from t=1 on no structural restriction on the interactions between group members. I.e. all dyads are simultaneously "active" in the process of social influence, Technically, every agent can have a weight different from zero towards all other group members (C_i=(N-1), see equation 1). The *second* scenario represents the intuition that it may temper the effects of strong faultlines if in the early phases of the group process interactions are restricted to relatively homogeneous smaller subgroups (C_i<(N-1)). In other words, dyads between group members who are strongly different demographically are not active in the first phase of influence process, while dyads between similar group members are. For the sake of idealization, we represent the subgroups as isolated "caves" such that interaction in the early phase is entirely restricted within caves. All weights between agents who do not belong to the same cave are set to zero and kept at zero until caves are merged. The corresponding timing scenario of "first homogeneous caves, then complete" imposes from t=1 on isolated caves which are arranged in such a way that the demographic homogeneity within the caves is very high[7]. Then, after a critical time point t*, the boundaries between caves are eliminated. Technically, we set at t* all weights between agents who belong to different caves to the value that corresponds to their current overall similarity (see equation 2) and leave all others weights and opinions unchanged. Hence, after t* agents will be influenced by all other agents in the team and can have non-zero ties with all other group members. The *third* and final scenario is a control condition that we called "first heterogeneous caves, then complete". We wanted to know whether the formation of smaller subgroups in the initial phase also tempers faultline effects when the subgroups are not homogeneous but instead are formed randomly. Our intuition is that this scenario will not differ from the baseline, because particularly in groups with strong

[7] Details of the method for maximization of demographic homogeneity depend on the exact settings for group size and cave size and will thus be explained in the discussion of the initialization of our computational experiments (section III.3).

faultlines it is likely that demographic divisions will induce early splits and extremist opinions within each separate cave. Once this local polarization arises, the merger of caves into one large group is likely to "export" extremism and thus polarization also into the group as a whole.

II.2.4. Aggregate outcome measures

The main claims of the theory of faultlines address two relationships: First the relationship between faultline strength and the level of consensus in the team, and second the relationship between faultline strength and the degree to which divisions in opinions are associated with demographical divisions in the team. To assess whether our model can reproduce these relationships, we devise four different outcome measures, *opinion diversity* , *opinion variance, polarization* and a measure of the degree to which differences in fixed (demographic) and flexible (opinion) attributes of agents are associated with each other, called attribute-opinion covariance, *cov(fix;flex)*.

Opinion diversity is based on a count of the number of different opinion vectors present in the group as a whole, where only flexible attributes are taken into account. For normalization, we divide this number by the group size *N*. We set *opinion diversity* =0 if there is perfect consensus. Hence, $0 \leq$ *opinion diversity* ≤ 1. Clearly, both a group with high consensus and a group with perfect polarization will exhibit low *opinion diversity*. Perfect consensus implies that all agents share the same vector of opinions (*opinion diversity* = zero), whereas perfect polarization implies that there are exactly two maximally different factions in the opinion space (*opinion diversity* = $2/N$).

Opinion variance is the average standard deviation of opinions across all *K* dimensions of the opinion space. In the case of perfect consensus, we obtain *opinion variance* = 0, and in the case of perfect polarization with two equally large maximally opposed subgroups we measure *opinion variance* = 1, the highest value we ever obtained[8]. However, a high level of *opinion variance* does not necessarily indicate that the group polarizes in the opinion space. High *opinion variance* may occur if agents strongly differ from each other in all dimensions of the opinion space, but these differences are not correlated across dimensions. In that case, the group is not polarized.

[8] In this case, the average opinion in all dimensions is zero. Moreover, in all dimensions half of the group adopts an extreme opinion at +1 and the other half of the group does so at -1. Hence, on average the distance from the mean amounts to +1 in all dimensions, yielding the result of *variance* = +1.

Polarization captures the degree to which the group can be separated into a small set of factions who are mutually antagonistic in the opinion space and have maximal internal agreement. To compute *polarization*, we use the variance of pairwise 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 (averaged across all *K* subdimensions). This measure obviously adopts its lowest level of zero for the case of perfect consensus. The maximum level of polarization (*polarization*=1) is obtained when the population is equally divided between the opposite ends of the opinion scale at -1 and +1 and all opinion dimensions are perfectly correlated[9]. With uniformly distributed opinions, the polarization measure yields approximately 0.22 for *K*=1.

To test the relationship between demographic differences and differences in opinions, we compute the attribute-opinion covariance, *cov(fix;flex)* 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 dissimilarity measures $\Delta_{i,j}^{fix}$ and $\Delta_{i,j}^{flex}$, as given by equations (4a) and (4b). These dissimilarity measures express the average distance across all dimensions for fixed attributes and flexible opinions, respectively. The resulting covariance *cov(fix,flex)* adopts values between -1 and 1. A value of zero indicates that similarity in opinions and similarity in demographic attributes are statistically unrelated. The initial values of *cov(fix,flex)* are expected to be near to zero, because opinions are initialized randomly. Changes of *cov(fix,flex)* that occur when the simulation proceeds indicate how much differences in opinions and demographic differences become aligned.

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

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

Thus *cov(fix;flex)* is calculated as given by equation (5).

[9] To see this: In 50% of all dyads the agreement is 1 (indicating maximal agreement), in 50% it is -1 (indicating maximal disagreement). The average level of agreement is zero and the average distance between the agreement in a particular dyad and the average level of agreement, i.e. the variance, yields polarization = 1.

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

II.3. Results of the computational experiments

We structured our computational analysis in two sets of experiments. In the *first set of experiments* the objective is to show that the dynamics of our model are consistent with Lau and Murnighan's (1998) informal reasoning. More precisely, we devise a fixed work team scenario and conduct ceteris paribus replications of the group dynamics that our model generates for different levels of faultline strength under the given scenario. The stylized regularity our model should produce in this set of experiments is a clear-cut negative relationship between the average level of consensus in the opinion distribution and the strength of demographic faultlines, *r*. More in particular, the model should generate both less often consensus and more often polarization as *r* increases. A second regularity that follows from the theory of faultline effects is an increasing association of opinion divisions with demographic divisions as faultlines become stronger. In other words, the stronger are the demographic faultlines, the clearer we expect subgroup splits in the simulated opinion distribution to reflect the distribution of demographic attributes.

The *second set of experiments* focuses on the effects of timing. Broadly, we expect that the negative effects of strong faultlines will be considerably tempered when homogeneous and mutually isolated subgroups are formed in a first phase, before in a second phase all group members interact with each other. We also want to test whether – as we intuit - this form of timing reduces the association between demographic and opinion differences in the team. To test these intuitions, we will conduct ceteris paribus replications of the scenario analyzed in the first set of experiments, but now with variation of the timing of contacts across the two options of "first homogeneous caves, then complete" and, "first heterogeneous caves, then complete", where the results of experiment 1 serve as the "no timing" baseline.

In both sets of experiments we use the following parameter settings. With regard to group size, we assume N=20, a size that is not too big to be unrealistic for a work team, but also large enough to allow for a sufficiently fine-grained variation in the strength of demographic faultlines (cf. Table 1). Furthermore, we assume that there are three salient demographic (fixed) attributes (D=3). As table 1 shows, the combination of 20 agents and

3 fixed attributes allows sufficient variation in the correlations between the fixed attributes of team members. Values for the demographic attributes are assigned to agents as shown in table 1, imposed by the data set we generated for the corresponding level of faultline strength *f*. For the number of flexible attributes (opinions), we choose K=4. This is the smallest number that makes polarization under strong faultlines not trivial, because with K=3 and D=4 it is still possible that two agents who maximally differ in all three demographic dimensions can have a positive relationship if they have sufficiently similar opinions. At the same time, this setting makes it hard to avoid polarization in a group with maximal faultline strength. Accordingly, K=3 and D=4 provides a particularly hard test for our conjecture that the right form of timing can prevent polarization even in groups with strong faultlines. Furthermore we assumed that initially (at the outset of t=1) all opinions of all agents are randomly drawn from a uniform distribution with full coverage of the entire opinion interval and with statistically independent dimensions of the opinion space. As a consequence, initial opinions are also statistically independent from demographic attributes. After initial opinions have been assigned, initial weights are computed on basis of overall similarity (see equation 2). In the timing experiments, initial weights between agents who do not belong to the same cave are set to zero and kept at zero until the boundaries between caves are removed.

II.3.1. Experiment 1: The effects of faultline strength

To illustrate how variation in faultline strength affects the model dynamics, we show first two typical simulation runs obtained for a setting with low faultline strength (r=0.2) and high faultline strength (r=0.8), respectively. Figure II.1 charts for both settings the dynamics of the four outcome measures for the first 120 iterations.

Figure II.1 shows dramatically different outcomes for the two different levels of faultline strength. In the weak faultline case, the simulated group quickly moves towards perfect consensus, as indicated by the rapid decline of *opinion diversity* and *opinion variance*, as well as *polarization*, from the levels given by the initial random distribution down to the theoretical minimum level of zero for all three outcome measures. The graph also shows that from the outset there is no (actually even a slightly negative) association between differences in opinions and demographic differences (see *cov(fix;flex)*). In the strong faultline case, it takes about 60 iterations until the group has moved from the random initial opinion distribution towards perfect polarization into two maximally opposed factions. Moreover,

opinion divisions and demographic divisions align almost perfectly in this case, as indicated by a level of *cov(fix;flex)* = 0.8 obtained after about 60 iterations.

Figure II.1: Change in outcome measure for typical simulation runs with weak faultline (left) and strong faultline (right). N=20, D=3, K=4. No timing of contacts.

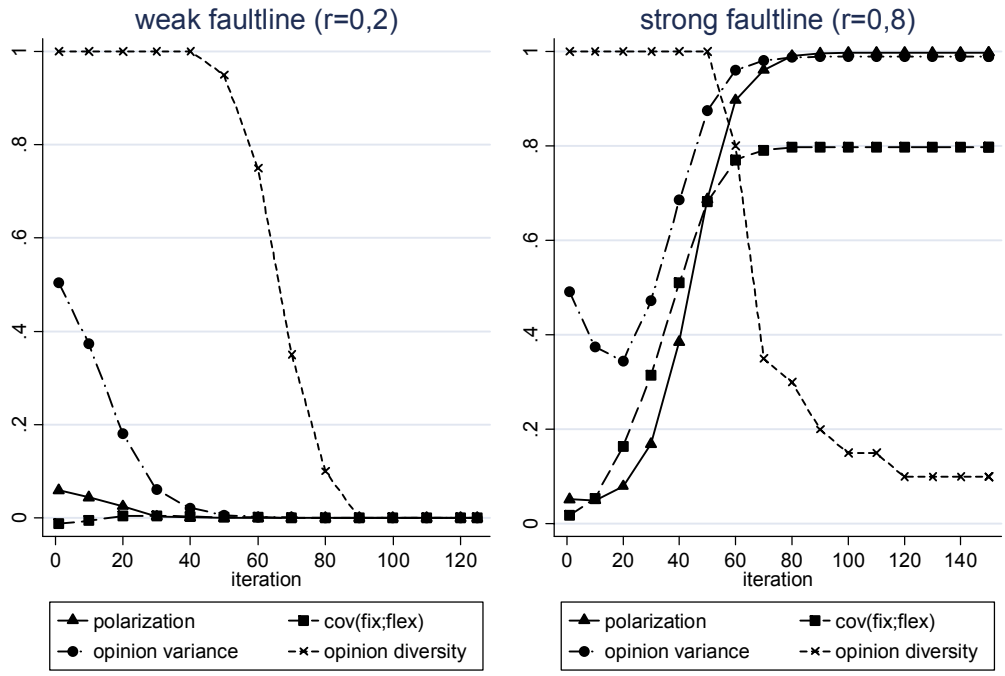


Figure II.1 shows dramatically different outcomes for the two different levels of faultline strength. In the weak faultline case, the simulated group quickly moves towards perfect consensus, as indicated by the rapid decline of *opinion diversity* and *opinion variance*, as well as *polarization*, from the levels given by the initial random distribution down to the theoretical minimum level of zero for all three outcome measures. The graph also shows that from the outset there is no (actually even a slightly negative) association between differences in opinions and demographic differences (see *cov(fix;flex)*). In the strong faultline case, it takes about 60 iterations until the group has moved from the random initial opinion distribution towards perfect polarization into two maximally opposed factions. Moreover, opinion divisions and demographic divisions align almost perfectly in this case, as indicated by a level of *cov(fix;flex)* = 0.8 obtained after about 60 iterations.

The explanation for the differences shown by Figure II.1 can be readily derived from our model assumptions. In the weak faultline scenario, demographic attributes are almost perfectly uncorrelated with each other. Hence, there are only very few pairs of agents who maximally differ on all three demographic dimensions. This makes it unlikely that negative

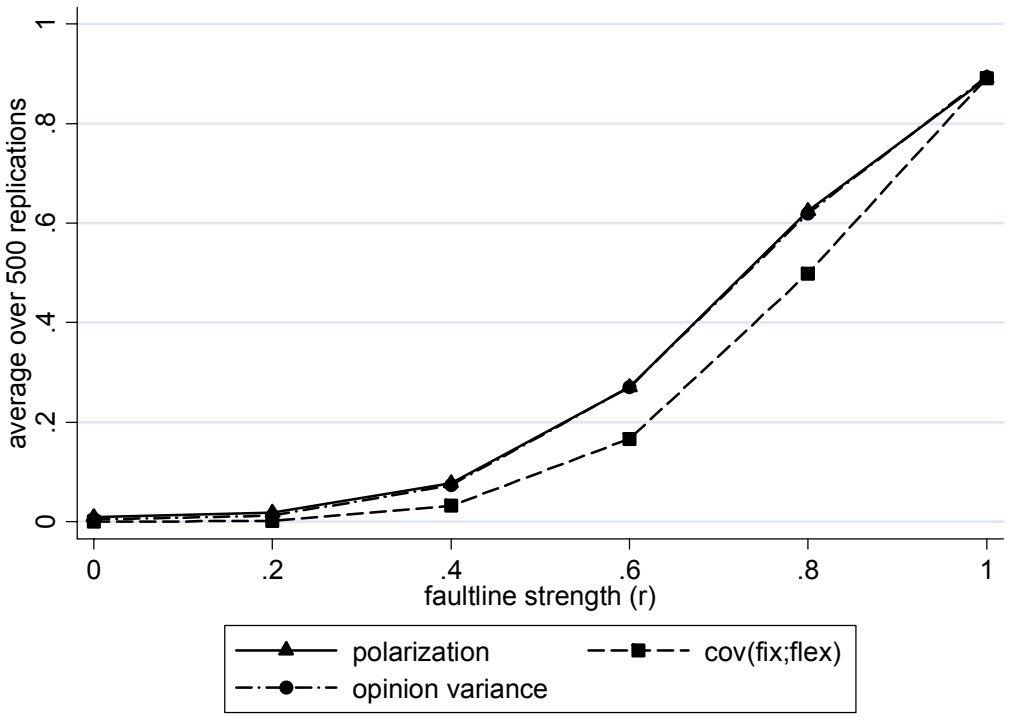
ties ($w_{ij}<0$) arise in the initial configuration. In addition, if some negative ties arise, then they will most likely be between agents who are, in turn, embedded into a large number of positive ties with the same colleagues. As a consequence, positive social influence prevails and rejection hardly ever occurs in the social interactions between agents. If some agents are "pushed" to reject some enemies' opinions, then they are at the same time "pulled in" by many more friends so that the net change of their opinion is more likely towards the group average than towards the extreme ends of the opinion scale. A similar reasoning explains why the outcome for the strong faultline case is so different. In the strong faultline case, demographic differences are maximal within a large fraction of the dyads in the team. In these dyads only relatively small opinion differences in the initial configuration suffice to generate a negative relationship between the interactants. Moreover, these negative relationships tend to segregate the two major subgroups in demographic space so that most agents have the same enemies than their friends have. This entails a quick self reinforcing dynamic towards opinion polarization. Most agents move towards whatever is the current average opinion profile in their (demographic) in-group and they distance themselves from whatever is the current average opinion profile in the (demographic) out-group. The result is a coordinated movement of all agents that soon leads to convergence of their opinions on two opposite poles that align with the demographic faultline in the group.

For statistical reliability, we conducted a large number of replications of this simulation experiment and varied faultline strength across the entire interval between r=0 and r=1.0 in steps of 0.2. Figure II.2 reports the average of the outcome measures we obtained after iteration 1000, over 500 replications per condition. We do not report *opinion diversity* in Figure II.2, because final states are almost always either perfectly polarized or exhibit perfect consensus so that the variation of *opinion diversity* across conditions is extremely small. To make it easier to distinguish the different outcome measures in the Figure, we used lines to connect the data points for the six different levels of r that we simulated but we did of course not obtain results for r-values other than those shown in table II.1.

Figure II.2 clearly confirms that our model generates the stylized regularities predicted by Lau and Murnighan's theory of faultlines. All three outcome measures consistently increase with higher levels of faultline strength. More specifically, the average outcomes of almost zero for *opinion variance*, *polarization* and *cov(fix;flex)* when demographic dimensions are entirely unrelated (r=0) indicate that virtually all simulated groups have reached almost perfect consensus in this condition. By contrast, with maximal faultline

strength (r=1.0) groups almost always polarize maximally, as indicated by an average *polarization* and an average *opinion variance* at the same level. The correspondingly high value of *cov(fix;flex)* in this condition shows that it is the demographic faultline along which the group also splits in the opinion space. The consistent increase of the outcome in between these two extremes shows that - for the given set of conditions (N=20, D=3, K=4) - our model clearly implies that higher faultline strength is associated with less consensus, more polarization and a stronger association between demographic and attitudinal differences, as predicted by Lau and Murnighan's theory. A further striking feature of Figure II.2 is that average polarization and opinion variance take almost the same values for all conditions. The reason for this is explained in more detail in Flache and Mäs (2008b). It is shown there that the model tends to generate in almost every replication of the experiment either nearly perfect polarization or nearly perfect consensus. The effects of faultline strength reported in Figure II.2 mainly reflect a shift in the distribution of these two outcomes. Accordingly, in a single run polarization and opinion variance take in equilibrium almost always either both the value of zero (consensus and no polarization) or of +1 (maximal variance and maximal polarization).

Figure II.2: Effect of faultline strength on outcome measures, averages over 500 replications per conditions, outcomes measured after 1000 iterations per replication N=20, D=3, K=4. No timing of contacts.



II.3.2. Experiment 2: Effects of timing of contacts

The design of our second experiment mirrors that of experiment 1, except that we replicate all conditions for the two different forms of timing, "first homogeneous caves, then complete" and "first heterogeneous caves, then complete". For the conditions that impose temporary caves, we set the size of caves to N_C = 5. This cave size is chosen because with N=20, it allows to easily generate demographically homogeneous caves. With N_C = 5 and the 50/50 distribution of demographical attributes that we use in all demographic dimensions, it is always possible to collect within one cave those 25% of the agents in the group who are equal on at least the first two of their three demographic attributes. For the condition "first homogeneous caves, then complete", we generate the corresponding caves as follows. In a first step, we lexicographically order the set of agents based on their three fixed attributes. Thus, the first five agents in this ordered set have attributes -1,-1 on dimensions d=1 and d=2 respectively, the subsequent five agents have attributes -1,+1 and so forth. In the second step, we match these relatively homogeneous subgroups of five generated in step 1 with the caves of size five.

Table II.2: Initialization of homogeneous caves

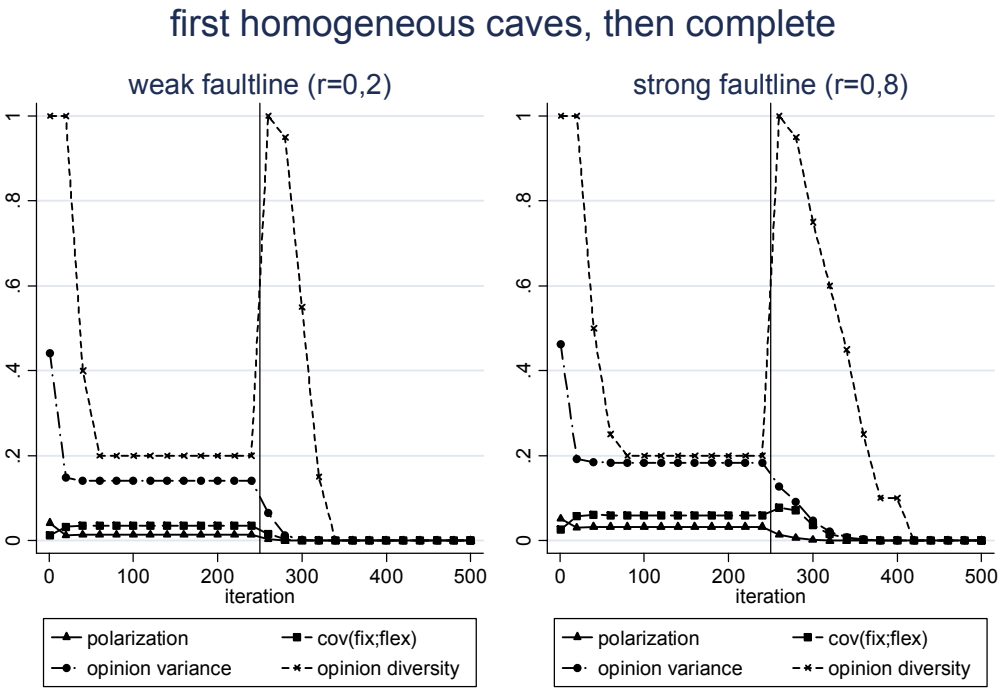
	***r* = 0.8**			***r* = 0.6**			***r* = 0.4**			***r* = 0.2**			***r* = 0**		
*i**	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3	attr. A_1	attr. A_2	attr. A_3
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1
6	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	1	-1
8	-1	-1	-1	-1	-1	1	-1	1	-1	-1	1	-1	-1	1	-1
9	-1	-1	1	-1	1	-1	-1	1	-1	-1	1	1	-1	1	1
10	-1	1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	1
11	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1
12	1	1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1
13	1	1	1	1	1	-1	1	-1	1	1	-1	1	1	-1	-1
14	1	1	1	1	1	1	1	1	-1	1	-1	1	1	-1	1
15	1	1	1	1	1	1	1	1	1	1	1	-1	1	-1	1
16	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	-1
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

*Note that actors' numbers *i* in this table do not correspond to those in table II.1

The result of this procedure is shown in table II.2 which shows the composition of the caves depending on the strength of the faultline. If the correlation between the demographic variables is perfect (a case not included in this table) then there are 4 perfect homogeneous caves: two where all actors hold on all attributes the value +1 and two where all actors hold on all attributes the value -1. The grey cells in table 2 indicate that for other cases some caves are not perfectly homogeneous. While this can not be avoided under the assumptions that $N = 20$ and $N_C = 5$, the table also shows that our method generates a high level of homogeneity within caves.

Our method assures that in the condition "first homogeneous caves, then complete", there are almost no negative weights within caves in the initial condition, regardless of the level of faultline strength. Finally, we assumed that in the conditions with caves, the caves are merged in iteration t^*=250. This choice of the critical time point assured that the dynamics within caves had practically settled down to equilibrium before the caves were joined.

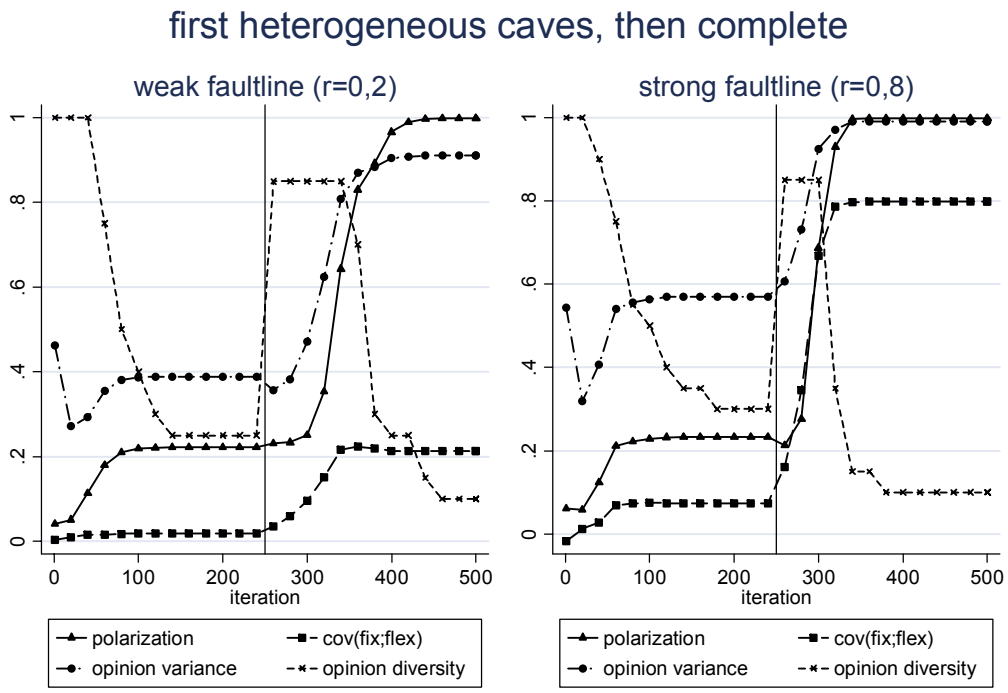
Figure II.3: Change in outcome measure for typical simulation runs with timing "first homogeneous caves, then complete", for weak faultline (left) and strong faultline (right). N=20, D=3, K=4.



We start again with an illustration of the timing effects by a comparison of typical model dynamics as we obtained them for the two different forms of timing with caves,

crossed with low faultline strength (r=0.2) and high faultline strength (r=0.8), respectively. Figures II.3 and II.4 chart for all four settings the dynamics of the four outcome measures until equilibrium. Both figures show for both weak and strong faultlines the type of dynamic that we encountered most frequently in the replications we ran for the corresponding condition (cf. Figure II.6). The corresponding "no timing" baseline is visualized by the time charts in Figure II.1.

Figure II.4: Change in outcome measure for typical simulation runs with timing "first heterogeneous caves, then complete", for weak faultline (left) and strong faultline (right). N=20, D=3, K=4.



The Figures show that the merger of the caves at time t*=250 dramatically changes group dynamics under both forms of timing. But already in the first phase of the simulated group process there are remarkable differences between homogeneous and heterogeneous caves. The results illustrate that the simulated group dynamics in homogeneous caves exhibit a very strong tendency towards perfect consensus both for weak and for strong faultlines before the merger occurs. *Polarization* drops to nearly zero in this phase and the measures for *opinion variance* and *opinion diversity* approach low levels (about 0.1 and 0.2, respectively). The explanation for this pattern is that homogeneous caves generate local convergence within the caves. Between the agents in a homogeneous cave, there are almost no negative ties. Accordingly their opinions converge towards the average of the randomly chosen initial local opinion distributions. In other words, in each cave the agents reach consensus

on moderate opinions. All local initializations are drawn from the same random distribution. As a consequence, the remaining *opinion diversity* and *opinion variance* between the caves is also relatively small. The *opinion diversity* of 0.2 at *t*=250* in both subfigures of Figure II.3 show that exactly four different opinion vectors remain after the early phase, one per cave. The corresponding low *opinion variance* (about 0.1) indicates that the differences between the caves are also very small. By contrast, Figure II.4 shows that in heterogeneous caves groups tend to develop a higher level of *polarization* already within the caves, at both levels of faultline strength. The *polarization* measure increases in both conditions to about 0.2 and *opinion variance* moves to 0.4 (weak faultline) and 0.6 (strong faultline). *Opinion diversity* takes in both runs the value 0.25 before the merger. This indicates that there are 5 different opinion vectors in the team what shows that there is perfect consensus in 3 caves and perfect polarization in one[10]. The local polarization is triggered by the relatively high proportion of negative within-cave ties that is generated due the high likelihood that demographically strongly dissimilar agents are matched within the same cave by the random assignment procedure.

The different developments within the caves set the stage for the dynamics that unfold after merger. A comparison of Figures II.3 and II.4 shows that after *t**=250, groups move to perfect consensus when caves were homogeneous, while perfect polarization is the outcome when initially caves were heterogeneous. Under homogeneous caves, all caves reached consensus on moderate opinions. As a consequence there are virtually no negative ties in the overall group at the point when caves are merged. Accordingly, social influence is overwhelmingly positive and all agents move towards and converge upon the current average group opinion. By contrast, with heterogeneous caves, the dynamics of rejection drove the agents of one cave to the very extreme ends of the opinion dimensions already before the caves are joined. After the merger these extremists exert influence on all team members, with many of whom they have negative ties due to their large opinion differences. As a consequence, agents sufficiently disagree with each other within many dyads, to generate a large proportion of negative ties within the group as a whole at the point when the caves are connected. The result is that in the runs shown by Figure II.4, the previously uncoordinated local division lines merge after *t**=250 into a developing global opinion division, as exhibited by the maximum level of polarization (1.0) shown for the

[10] It is also possible that this result obtains when the opinions in more than one cave polarized and the opinion vectors in the different caves happened to be equal. We checked to make sure that this was not the case in the runs that are reported in Figure V.s III.3 and III.4.

final state in both subgraphs of Figure II.4. The dynamics of the attribute-opinion association *cov(fix;flex)* in Figure II.4 also show that this division occurs mainly along demographical differences when faultlines are strong, whereas the division is only weakly related to demographical differences when faultlines are weak.

For statistical reliability, we conducted again a large number of replications of this simulation experiment and varied faultline strength across the entire interval between r=0 and r=1.0 in steps of 0.2. Figure II.5 reports the results we obtained in both timing conditions for the outcome measures of *polarization* (a), *opinion variance* (b) and *cov(fix;flex)* (c). For comparison, we also include the baseline results of "no timing" in the figures. Results are averages based on 500 replications per condition, where we measured the outcomes after 1000 iterations.

Figure II.5 confirms the patterns exhibited by the typical simulation runs shown in Figures II.3 and II.4. Overall, we find that the indicators for polarization and its association with demographic differences are dramatically lower when the "right" form of timing is chosen (homogeneous caves) than in any of the two alternative cases (no timing or heterogeneous caves). The results also support our intuition that the "right" form of timing strongly tempers the negative effects of faultline strength that we found in the baseline condition of "no timing". As part (a) of Figure II.5 shows, without timing, average *polarization* increases from zero at r=0 (no faultline) to almost the theoretical maximum of 1 at r=1 (maximally strong faultline). With homogeneous caves, there is only a slight increase of *polarization* between those two extremes, from zero at r=0 to 0.2 at r=1. An inspection of the measure of association between fixed and flexible attributes (part (c) Figure II.5) reveals that the formation of homogeneous caves also greatly reduces the degree to which opinion differences in the team align with demographic differences. While the association measure increases for no timing from *cov(fix;flex)* about 0 at r=0 to about 0.6 at r=1, the association measure increases only slightly under homogeneous caves, from *cov(fix;flex)* about 0 at r=0 to about 0.2 at r=1.

Figure II.5: Effect of timing and faultline strength on average *polarization* (a), average *opinion variance* (b) and average association between demographic differences and opinion differences (c), based on 500 replications per conditions, outcomes measured after 1000 iterations per replication N=20, D=3, K=4.

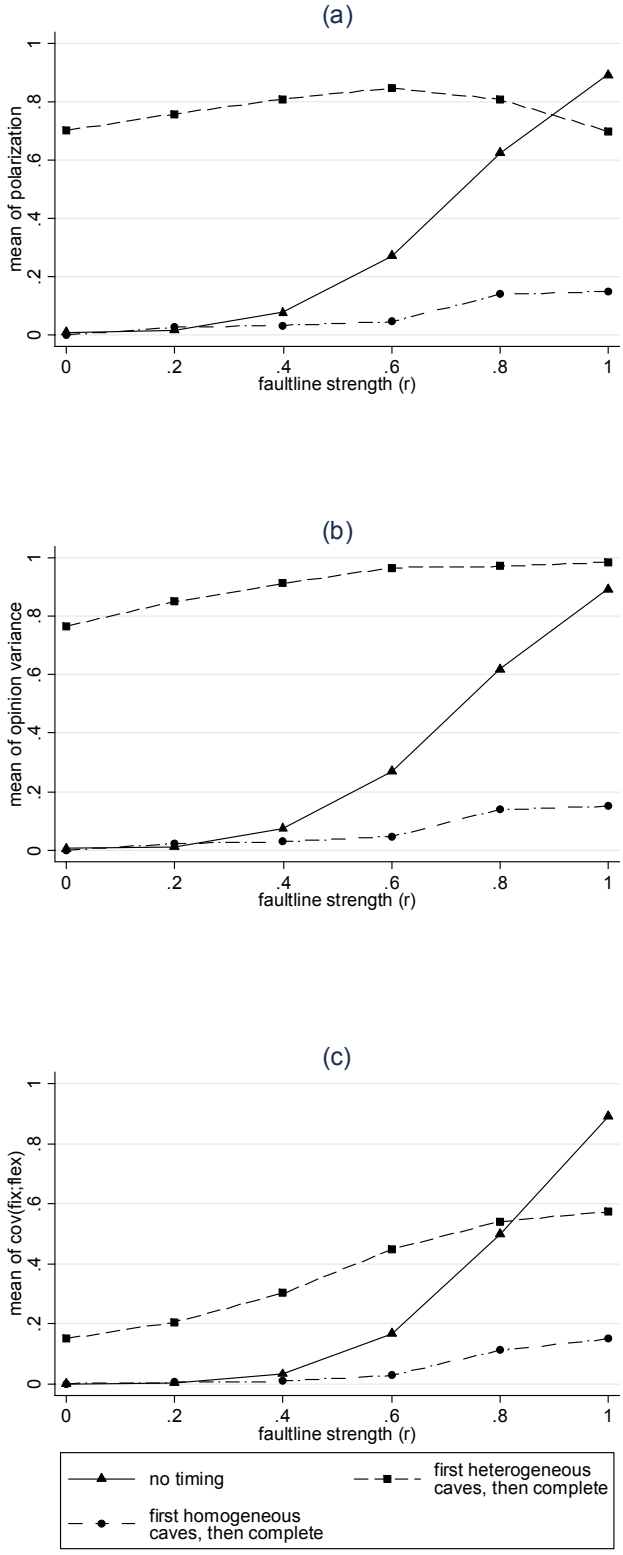


Figure II.5 shows furthermore that the effects of timing strongly depend on the "right" choice for the method of assignment of agents to caves. Broadly, while homogeneous caves generated high levels of consensus and virtually eliminated the negative effects of strong faultlines, heterogeneous caves turned out to make things even worse than they were in the baseline condition of no timing. Part (a) of Figure II.5 shows that with heterogeneous caves, *polarization* was high (about 0.6) even without demographic faultlines and stayed high at all levels of faultline strength. Correspondingly, we found at almost all levels of faultline strength a higher *opinion variance* (part (b)) and stronger association between demographic and opinion differences (part (c)) for heterogeneous caves than for any of the other timing conditions. Only for very strong faultlines (r=0.8 and r=1.0), we find that a further increase in faultline strength is related to a slight decline of the average level of polarization in initially heterogeneous caves, such that for r=1.0 the level of polarization is even somewhat lower than in the baseline condition of no timing. This decline will be explained further below, when we present a detailed analysis of the distribution of equilibrium outcomes that generated the averages reported in Figure II.5.

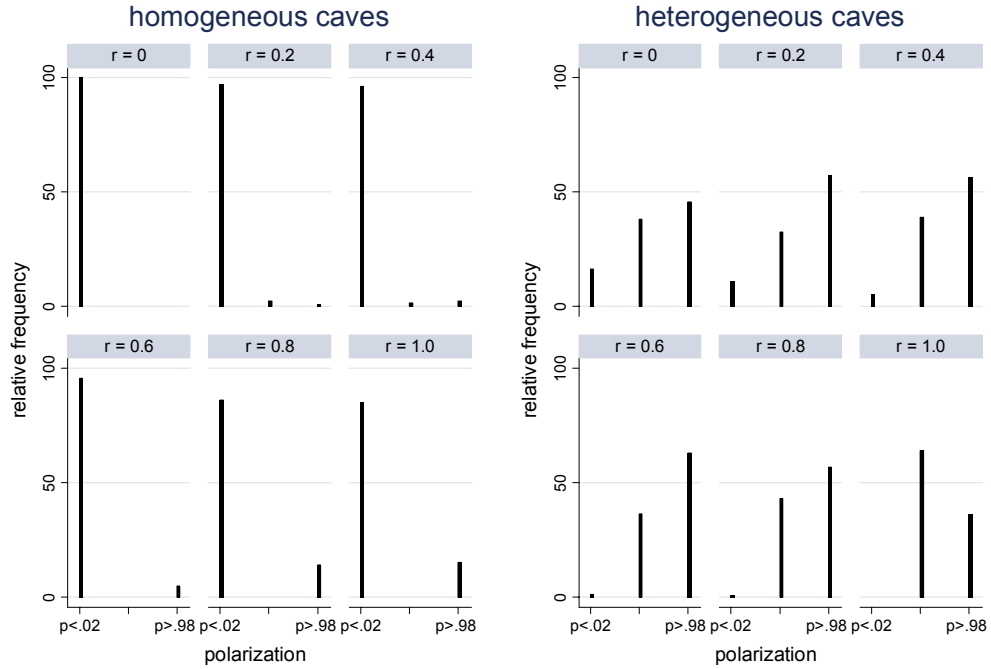
Figure II.5 shows the expected association of the timing conditions with the outcome measures, but it does not directly test our intuition that the effect of the timing conditions can be attributed to a reduction of negative ties in the early phase of the group process. For this, we checked as a first test whether the manipulations of homogeneous and heterogeneous caves affected the proportion of negative ties in the group at the time point before caves were merged (t=250) in the expected direction. While without caves (no timing) on average across all levels of r (3000 runs) 13.3% of all possible dyads were strongly negative ($w_{ij} \leq$ -0.95), there was not a single strongly negative tie[11] in any of the 12000 simulated homogeneous caves. As expected, this discrepancy between no timing and homogeneous caves became more pronounced for stronger faultlines, with a maximum level of about 47% of all possible ties in iteration 250 being strongly negative at r= 1.0 with no timing (and zero for homogeneous caves). For comparison, we found 21.9% strongly negative ties at t=250 with heterogeneous caves. These checks confirm our expectation that homogeneous caves suppress the formation of negative ties, while heterogeneous caves foster negativity compared to the baseline of no timing. We also tested whether a higher proportion of negative ties at t=250 was related to higher levels of polarization in equilibrium across all conditions of the experiment. We discovered that even a very small

[11] The number of iterations (250) was chosen large enough to assure that if a tie was negative at this point, it would also be strongly negative, i.e. have a weight w of less than -0.95.

proportion of negative ties at t=250 dramatically increased polarization in the final state. With no timing, average polarization in equilibrium across all levels of faultline strength (3000 runs) was about 0.01 if the proportion of negative ties at t=250 was less than 2.5% (947 runs). The average polarization soared to 0.997 if the proportion of negative ties was above that figure (2053 runs) in this condition. The corresponding figures for heterogeneous caves are an average polarization in equilibrium of about 0.16 if less than 2.5% of the possible ties were negative in t=250 (200 runs), and of about 0.73 with more negative ties (2800 runs). The lower level of polarization for heterogeneous caves reflects the decline of the polarization level that Figure II.5 showed for this condition at high levels of faultline strength r. An explanation for this decline can be found in a closer inspection of the distribution of equilibrium outcomes that generated the averages reported in Figure II.5. We turn now to this analysis.

Figure II.6 shows how timing affects the distribution of *polarization* in equilibrium over the 500 replications that we conducted in each of the two cave conditions in experiment 2. Overall, Figure II.6 supports the interpretation that homogeneous caves greatly increase the odds that a group ends up in consensus, even when faultlines are strong. The highest relative frequency of polarized groups that we obtained with homogeneous caves was 15% for maximally strong faultiness (r=1.0). In almost all other realizations under this timing condition the outcome was perfect consensus. That in some cases groups polarize despite homogeneous caves is a consequence of how the caves were formed. As table II.2 shows, some of the caves are not *perfectly* homogeneous. For r = 0.8, for example, actor 11 holds only on one demographic attribute the same value as the other actors in his cave. The other actors in this cave are demographically completely similar to each other. As a consequence, the likelihood is high that the randomly drawn initial opinions generate differences between actor 11 and his cave mates that are high enough to impose negative ties. Agents will then reject each other's opinions and with a high chance the cave perfectly polarizes. Thus all actors from this cave hold extreme opinions. After the merger those extremists suffice to create enough negative ties in the whole team to make it polarize as well.

Figure II.6: Distribution of polarization measure over 500 replications per condition, broken down by the six different levels of faultline strength and the two different forms of timing. *N*=20, *D*=3, *K*=4.



p < .02: perfect consensus
p > .98: perfect polarization (2 opposing subgroups)
.02 ≤ p ≤ .98: multiplex equilibrium outcome (more than 2 subgroups)

As we expected, the pattern is quite different when caves are heterogeneous. As Figure II.6 shows, we find that even with no faultlines (*r*=0) about 46% of the replications generated a perfectly polarized group and only about 16% produced perfect consensus. Like for homogeneous caves and for the baseline condition, the stronger the faultline is the less often the dynamics end with perfect consensus, but the overall level of polarization that we find for weak faultlines is much higher. This result differs clearly from the baseline condition (no timing) where about 98% of the runs ended in perfect consensus when there was no faultline. The key factor that drives this result is the relative size of the caves. With heterogeneous caves, the four different caves per group can be seen as four independent replications of the baseline condition, but each with a much smaller group size (*N*=5 as opposed to *N*=20) than in the baseline. But the smaller the caves, the more likely it is that there are at least some caves in which there is a relatively high concentration of negative ties from the outset[12]. As a consequence it is relatively likely that the opinions in at least

[12] The reason is that in smaller caves the initial random distribution of opinions produces a relatively sparse coverage of the opinion space. As a consequence, occasional initial "extremists" are likely to have larger

one of the caves polarize perfectly. If the caves are then merged the extremists pull (or push) the rest of the team to the extremes of the opinion scales. Thus the whole team polarizes.

A second main difference between the results from the condition "first heterogeneous caves, then complete" and the other two timing conditions is that with heterogeneous caves multiplex equilibria occur much more frequently. In a multiplex equilibrium, all actors hold extreme opinions, but there are more then only two different opinion vectors in the group. A multiplex equilibrium can arise if the overall pattern of relationships and opinions is exactly balanced so that "push" and "pull" forces exerted upon agents' opinions from different groups of friends and enemies exactly neutralize each other (cf. Macy et al. 2003). Multiplex equilibria are relatively frequent under the timing condition "first heterogeneous caves, then complete" because often the opinions in more than one of the heterogeneous caves polarize before merger. If in addition the opinions of the extremists from different caves differ sufficiently then the extremists pull the moderate actors from the caves that reached a consensus to different poles of the opinion scales after the merger. Table II.3 reports the absolute frequencies of the number of different opinion vectors in teams after 1000 iterations (only for the experiments with heterogeneous caves). If there was only one opinion vector then the whole team reached a perfect consensus. As table II.3 shows, this happened in 81 of the 500 runs with minimal faultline strength. The stronger the faultline the less often the team found a consensus under this timing condition. 2 final opinion vectors indicate that the team perfectly polarized (*polarization* = 1). If there were more than 2 final opinion vectors in the team then a multiplex equilibrium was reached in which all agents hold extreme opinions but there were more then 2 subgroups in the team. This happened when the opinions in more then one of the heterogeneous caves polarized. If there is an unequal number of final opinion vectors larger than 1 then the opinions in more then one cave polarized but the opinion vectors of extremists from two different caves happened to be very similar. As a consequence the members of the two subgroups will have positive relationships. It can thus happen that after the merger these two subgroups of extremists coordinate on the same opinion vector. Table II.3 shows that the higher the faultline strength the more often multiplex equilibria occurred.

---

opinion distances and thus relatively more negative ties to other group members than in a more densely packed opinion space.

Table II.3: Number of different opinion vectors after 1000 iterations under the condition "first heterogeneous caves, then complete". (the crosstabulation shows absolute frequencies of runs)

# opinion vectors	Faultline strength						
	0	0.2	0.4	0.6	0.8	1	Σ
1	81	53	25	5	2	0	166
2	286	301	292	316	286	180	1661
3	3	4	14	3	0	0	24
4	99	105	119	128	148	205	804
5	1	2	8	5	7	3	26
6	21	24	28	26	30	53	182
7	3	4	4	2	1	0	14
8	3	6	7	12	26	59	113
9	3	1	2	3	0	0	9
10	0	0	0	0	0	0	0
11	0	0	1	0	0	0	1
Σ	500	500	500	500	500	500	3000

Multiplex outcomes are also the reason why we saw in Figure 5 that the mean of polarization decreases of r=0.8 and r=1.0 and that at r=1.0 the average level of polarization is even lower for initially heterogeneous than for no timing. Figure II.6 makes clear that this result should not be interpreted as showing that there was more perfect consensus under heterogeneous caves in these conditions. For example, in the 500 runs under the condition "first heterogeneous caves, then complete" the dynamics never produced perfect consensus. But Figure II.6 shows that in many runs under this condition more than two groups with partially opposing opinions formed. The opinions of all team members were at the poles of the respective opinion scales in all of these cases. However, the outcome was multiplex, so that there was not perfect polarization into *two* opposed subgroups in these cases. This is reflected by a value of the polarization measure that is somewhat lower than the theoretical maximum of 1.0, but still considerably above the level for consensus (0.0), which explains why on average across all runs we found a polarization level of about 0.7 in this condition.

We believe the reason that we find more multiplex outcomes with heterogeneous caves than in any of the other timing conditions lies with the uncoordinated local polarizations that under strong faultlines are likely to arise at the end of the first phase. In a four dimensional opinion space, there are only 16 possible combinations of extreme positions on every dimension. Equilibrium outcomes will arise if agents are distributed in the right way over these 16 combinations, or over smaller subsets of the combinations (e.g. with 4 or 8 different opinion vectors in the group) such that all mutual influences on agents' opinions are in balance. With heterogeneous caves, every locally polarized cave

produces at least two of those combinations with some incumbents. With five caves overall, it is not unlikely that this process generates an overall distribution in the group as a whole that is in or close to a multiplex equilibrium when all caves are merged.

II.4. Summary and Discussion

We modeled in this paper the effects of demographic faultlines on team performance. Lau and Murnighan's theory suggests that the stronger a team's demographic faultline is the less cohesive the team will be and the less likely will the team therefore be able to find a consensus with regard to work related opinions. As a consequence, teams with a strong demographic faultline tend to perform poorly. We proposed a formal computational model of this process based on four fundamental sociological mechanisms, homophily, heterophobia, social influence and rejection. We showed that the model generates results that are consistent with Lau and Murnighan's faultline theory. Our simulations demonstrate that the stronger the demographic faultline in a group the more likely will the group split up into subgroups (ceteris paribus). These subgroups' members hold opposing opinions and do not like each other.

We then used our model to show that the degree to which strong faultlines have negative effects may critically depend on the timing of contacts between group members. We tested a somewhat counter intuitive prediction: if in the first phase of the team interaction the team is separated into demographically homogeneous groups which are merged only later in the team process, then strong demographic faultlines do less often lead to opinion polarization than in a process where all group members interact with each from the outset. This result contradicts to some extent predictions of the prominent contact theory (Allport 1954; Pettigrew 1998) which states that contact improves the interpersonal relationships between demographically dissimilar actors. However, as discussed above, the effects of timing follow logically from the fundamental social mechanisms that constitute our model.

It was our main interest in this paper to show the theoretical consistence of the reasoning that implies effects of the timing of contacts in demographically diverse groups. Accordingly, we did not conduct an extensive analysis of the robustness of our results with regard to variation in other model parameters than those we have manipulated. We have shown elsewhere (Flache and Mäs 2008b) that our model reconstructs basic predictions of faultline theory also for different numbers of opinions and demographic dimensions. It is a

task for future research to conduct more extensive sensitivity analyses. At this point we see no a priori reason to expect that qualitative model results may fundamentally change for other sets of parameters, as long as the parameters of the model are chosen such that the model equations are consistent with the social mechanisms we assume.

The mechanisms we use in our model imply that timing is not the only manipulation that may avoid the negative effects of strong demographic faultlines on team cohesion. More generally, according to our model every condition that suppresses the emergence of negative ties helps to sustain group cohesion despite demographic divisions. Team building measures, emphasis on common goals or team learning may have similar effects than the right form of timing. Such measures have been proposed by previous research on the faultlines (e.g. Gibson and Vermeulen, 2003). Also previous work on agenda setting points to measures that may have similar effects (see e.g.: Levine and Plott 1977; List 2004; Plott and Levine 1978). Team managers might manipulate the sequence in which certain issues are discussed. If in a first phase only salient issues are discussed that all team members agree on, this would imply the emergence of positive interpersonal relationships between the team members. If then more controversial issues are addressed in a later phase, the prospects for finding a consensus are much better than compared to a situation where only controversial topics were addressed from the outset. Clearly, previous research on team building and agenda setting points to fruitful new applications of our model, but we also wish to emphasize that with the manipulation of the timing of contacts that we addressed in this paper our model suggests a measure that to our knowledge is new in the literature. One possible advantage of timing may be that it is a measure that organizations can implement unobtrusively, seemingly as a byproduct of functional arrangements of the workflow.

Future research should also focus on the mechanisms that produce opinion polarization. As we argued above, Lau and Murnighan's reasoning seems to critically hinge upon the assumption that there is an initial correlation between demographic attributes and opinions. In our model, this assumption is not necessary. Instead, the two negative mechanisms of heterophobia and rejection are sufficient to generate an effect of faultline strength on opinion polarization. We propose that future work should compare our model with the Lau and Murnighan reasoning on a theoretical level, to search systematically for contradicting predictions that can subsequently be submitted to empirical tests. We suggest that effects of the timing of contacts are particularly promising to compare the models empirically. We expect that the mechanisms we used and those of Lau and Murnighan

produce different dynamics under certain timing conditions. As we have shown, our mechanisms produce less polarization if first homogeneous subgroups are formed. By contrast, Lau and Murnighan's mechanism should lead to the opposite outcome. Their reasoning implies that in homogeneous groups the actors agree on opinions and that their opinions should become more extreme then. Because Lau and Murnighan assume that demographically dissimilar actors also hold opposing opinions, each of the subgroups will find a different very extreme consensus. If then the team is merged, all team members hold very extreme opinions and a consensus is very unlikely. If on the other hand in the first phase heterogeneous groups are formed then the actors in each subgroup will hold different opinions. If they then exchange the arguments their opinions are based on they may be able to convince each other. From this view, it is thus very likely that the subgroups find a consensus on moderate opinions. After the merger, the moderates will very likely find an overall consensus. Hence, the predictions of our model and the Lau and Murnighan reasoning are contradictory under certain timing conditions.

Our analysis has demonstrated how the theory of faultlines can be rigorously and formally reconstructed. We also have shown that this reconstruction can yield new, empirically testable hypotheses into the conditions and mechanisms that may temper or elicit the negative effects of demographic faultlines on team performance. Finally, our analysis suggests that the timing of contacts is a potentially fruitful governance instrument that managers may be able to use in order to avoid that the negative effects of demographic faultlines overshadow the benefits that diverse human and social capital can create for organization.