

# Supplementary Material

## Why Echo Chambers are Useful

Ole Jann

Nuffield College, University of Oxford

Christoph Schottmüller

University of Cologne and TILEC

November 29, 2021

### Contents

<b>1</b>	<b>Bipolar polarization</b>	<b>2</b>
1.1	Segregation and full information as equilibrium . . . . .	2
1.2	Welfare optimal room allocation . . . . .	4
1.3	Welfare optimal room assignment as equilibrium . . . . .	10
<b>2</b>	<b>Limit results to give lower bounds on polarization</b>	<b>11</b>
2.1	Uniformly distributed biases . . . . .	12
<b>3</b>	<b>Symmetrically single peaked bias distribution</b>	<b>15</b>
<b>4</b>	<b>Example: Too much segregation in equilibrium</b>	<b>16</b>
<b>5</b>	<b>Uncertainty</b>	<b>17</b>
5.1	Main results and intuition . . . . .	17
5.2	Detailed analysis and proofs for the model with uncertainty . . . . .	20
5.2.1	Preliminary analysis . . . . .	20
5.2.2	Proofs . . . . .	24
<b>6</b>	<b>Public information</b>	<b>26</b>
<b>7</b>	<b>Signaling</b>	<b>30</b>
7.1	Binary signal . . . . .	30
7.2	Continuous signal . . . . .	32
<b>8</b>	<b>Verification</b>	<b>33</b>
<b>9</b>	<b>States correlated within bias groups</b>	<b>35</b>

<b>10 Alternative signal technologies</b>	<b>36</b>
10.1 Larger signal and state space . . . . .	37
10.2 Continuum of signals . . . . .	39
10.3 Single state . . . . .	43
<b>11 Follower model</b>	<b>48</b>
<b>12 Mediated talk</b>	<b>55</b>
<b>13 Alternative empirical approach</b>	<b>57</b>
<b>14 Documentation of data collection</b>	<b>59</b>
14.1 Tweets of members of Congress . . . . .	59
14.1.1 Collection . . . . .	59
14.1.2 Generating relative monogram/bigram frequencies . . . . .	59
14.2 Main sample and reply sample . . . . .	59
14.3 Scoring . . . . .	60
14.3.1 Sentiment scores . . . . .	61
14.3.2 Other variables of interest . . . . .	61

## 1. Bipolar polarization

Here we consider the bipolar case, i.e. the case where a player's bias  $b_i$  is either 0 or  $b > 0$ . We denote by  $n_0$  ( $n_b$ ) the number of people with bias 0 ( $b$ ) and without loss of generality we let  $n_0 \geq n_b$ . Recall that a player's expected payoff in a given room configuration can be written as

$$U_i = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n-1)n] + (1/4 - p(1-p)) \left[ \zeta_i + \alpha \sum_{j \neq i} \zeta_j \right]. \quad (1)$$

where  $\zeta_i$  is the number of pieces of information player  $i$  has access to (through truthful messages or by observing his signal).

### 1.1. Segregation and full information as equilibrium

Now we ask the question when segregation, i.e. all players with bias  $b_i = 0$  choosing room 1 and all players with bias  $b_i = b$  choosing room 2, is an equilibrium. Clearly, every player is truthtelling in the most informative messaging equilibrium in this case and player  $i$ 's expected payoff is

$$U_i^{fs} = -\alpha n_{-i} b^2 - 1/4 [n + \alpha(n-1)n] + (1/4 - p(1-p)) [(1-\alpha)n_i + \alpha n_i^2 + \alpha n_{-i}^2] \quad (2)$$

where  $n_i$  is the number of players with the same bias as player  $i$  and  $n_{-i}$  is the number of players with the other bias.

To analyze when segregation is an equilibrium consider the incentives of player  $i$  to unilaterally deviate in his room choice. There are three relevant cases: (i) after the deviation everyone including  $i$  still sends truthful messages, (ii) after the deviation  $i$  will babble but the players in the room he is switching to still send truthful messages and (iii)  $i$ 's room switch leads to babbling by all players in the room he is switching to. We analyze  $i$ 's incentives to switch as well as conditions for each of these cases in turn.

First, truthful messages after switch. This occurs if and only if  $b \leq p - 1/2$ . If  $i$ 's bias group is the (weakly) smaller group, then  $i$  will benefit from a room switch in this case, see equation (2). If  $i$ 's group is the larger group, he will not. Hence, segregation does not exist as equilibrium if  $b$  is less than  $p - 1/2$ .

Second,  $i$  babbles after the switch but everyone else remains truthtelling. This happens if and only if  $p - 1/2 < b \leq n_{-i}(p - 1/2)$ . In this case,  $i$ 's deviation payoff is

$$U_i^d = -\alpha n_{-i} b^2 - 1/4 [n + \alpha(n - 1)n] + (1/4 - p(1 - p)) [1 + n_{-i} + \alpha n_{-i}^2 + \alpha(n_i - 1)^2]$$

which is higher than  $U_i^{fs}$  if and only if  $n_{-i} > (1 + \alpha)(n_i - 1)$ . Clearly, the members of the smaller group are the ones for who this constraint is more stringent. That is, a segregation equilibrium exists given  $p - 1/2 < b \leq n_{-i}(p - 1/2)$  if and only if the size of the bigger group is at most  $1 + \alpha$  times the size of the smaller group minus 1. Intuitively, members of the smaller group have a lot of information to gain from a switch if the larger group is very large. Hence, a segregation equilibrium only exists if the larger group is not too large.

Third, the switch leads to complete babbling in the room switched to. This will occur if and only if  $b > n_{-i}(p - 1/2)$ . In this case, it is straightforward to see that the switch is unprofitable and a segregation equilibrium exists. This yields the following result.

**Proposition 9.** *Without loss of generality assume  $n_0 \geq n_b$ . A segregation equilibrium exists if and only if one of the two following condition is met: (i)  $b > n_0(p - 1/2)$ , (ii)  $p - 1/2 < b \leq n_0(p - 1/2)$  and  $n_0 \leq (1 + \alpha)(n_b - 1)$ .*

Next we check for which parameter values a full information equilibrium, i.e. all players choosing the same room and report their signal truthfully, exists. Recall that  $n_0 \geq n_b$ . Then, truthful reporting is an equilibrium of the messaging game if and only if  $b(1 - (n_b - 1)/(n - 1)) \leq p - 1/2$ . In the full information equilibrium, each player's expected payoff is

$$U_i^{fi} = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n - 1)n] + (1/4 - p(1 - p)) [(1 - \alpha)n + \alpha n^2].$$

Unilateral deviation by player  $i$  in his room choice can lead to two scenarios: either

everyone still reports their message truthfully in the room  $i$  is deviating from (but  $i$  himself is no longer able to receive any messages) or not. In either case,  $i$ 's payoff is clearly lower after the deviation than before. Hence, a full information equilibrium exists if and only if  $b(1 - (n_b - 1)/(n - 1)) \leq p - 1/2$  which is equivalent to  $bn_0/(n - 1) \leq p - 1/2$ . This is summarized in figure 1.<sup>1</sup>

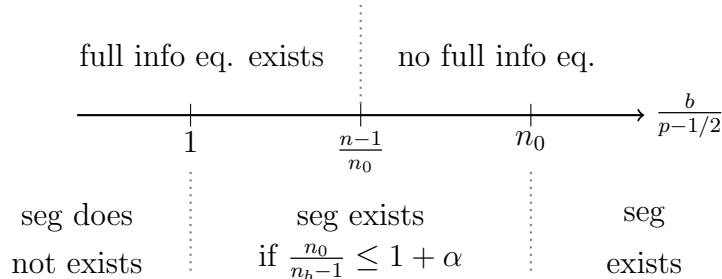


Figure 1: Segregation and full information as room choice equilibria

## 1.2. Welfare optimal room allocation

We will now consider how a planner would assign players to rooms in order to maximize the sum of players' payoffs. The planner's only tool is room assignment knowing the players' biases, i.e. the planner does not observe signals and cannot influence the messages sent or actions taken by the players.

From equation (1), it is clear that the planner's objective is equivalent to maximizing  $\sum_i \zeta_i$ , i.e. the total number of pieces of information by all players. We proceed in a number of lemmas. To avoid case distinction, we use in the following the convention that a player who is alone in a room will send a truthful message.

**Lemma 4.** *There is a welfare optimal room assignment without a room in which all players babble.*

**Proof of lemma 4:** In such a room there would be at least one player of each bias type. Splitting the room into two according to bias type would lead to weakly more transmitted pieces of information.  $\square$

**Lemma 5.** *Assume there are two rooms  $R_1$  and  $R_2$  such that in equilibrium players with bias  $x \in \{0, b\}$  send truthful messages in both rooms. Then players with bias  $x$  will also send truthful messages in a merged room  $R_1 \cup R_2$ .*

**Proof of lemma 5:** As players send truthful messages by assumption in room  $R_i$  for  $i \in \{1, 2\}$ ,

$$\frac{b}{p - 1/2} n_{R_i, -x} \leq n_{R_i} - 1$$

<sup>1</sup>Note that  $(n - 1)/n_0 \leq n_0$  as  $n_0 \geq n/2$  and  $n \geq 2$ .

has to hold where  $n_{R_i,-x}$  is the number of players in room  $R_i$  who do not have bias  $x$ . Summing this inequality over the two rooms  $i = 1, 2$  yields

$$\frac{b}{p-1/2} (n_{R_1,-x} + n_{R_2,-x}) \leq n_{R_1} + n_{R_2} - 2$$

which is sufficient for

$$\frac{b}{p-1/2} (n_{R_1,-x} + n_{R_2,-x}) \leq n_{R_1} + n_{R_2} - 1.$$

The latter inequality states that it is optimal for players with bias  $x$  to send truthful messages in a room  $R_1 \cup R_2$ .  $\square$

**Corollary 2.** *The welfare maximizing room assignment will not include rooms  $R_1$  and  $R_2$  such that either*

- *both  $R_1$  and  $R_2$  are both populated exclusively by players with bias  $x \in \{0, b\}$ , or*
- *both  $R_1$  and  $R_2$  are populated by players with both biases and all players send truthful messages, or*
- *$R_1$  and  $R_2$  are populated by players with both biases and only the players with bias  $x \in \{0, b\}$  send truthful messages, or*
- *$R_1$  is exclusively populated by players with bias  $x \in \{0, b\}$  and  $R_2$  is populated by players with both biases but only players with bias  $x$  send truthful messages in  $R_2$ .*

**Proof of corollary 2:** In each of these cases, merging the two rooms maintains truthtelling incentives for those that originally sent truthful messages by lemma 5. As the truthful messages are received by more players, merging clearly increases the planner's objective.  $\square$

**Lemma 6.** *The welfare maximizing room assignment will not include rooms  $R_1$  and  $R_2$  such that both rooms contain players of each bias and all players in  $R_1$  send truthful messages while only players of bias  $x \in \{0, b\}$  send truthful messages in  $R_2$ .*

**Proof of lemma 6:** First, we consider the case where more players are in room  $R_1$ , i.e.  $n_{R_1,x} + n_{R_1,-x} \geq n_{R_2,x} + n_{R_2,-x}$ . Note that the number of pieces of information generated in those two rooms is  $n_{R_1}^2 + n_{R_2,x} (n_{R_2,x} + n_{R_2,-x})$ . Note that  $n_{R_2,x} > n_{R_2,-x}$  as otherwise  $x$  would not be truthtelling in  $R_2$  while  $-x$  is not. Consider now an alternative room assignment that differs from the original one in the way that  $n_{R_2,-x}$  players of each bias are moved from  $R_2$  to  $R_1$ . Denote everything after the change using  $\tilde{\cdot}$ . That is,  $\tilde{R}_2$  will contain  $n_{R_2,x} - n_{R_2,-x}$  players of bias  $x$  and none of bias  $-x$  while  $\tilde{R}_1$  will contain  $n_{R_1,x} + n_{R_2,-x}$  players of bias  $x$  and  $n_{R_1,-x} + n_{R_2,-x}$  of type  $-x$ . The crucial result is that

all players in room  $\tilde{R}_1$  find truthtelling optimal: As truthtelling was optimal for players of both biases by assumption in  $R_1$ ,

$$\frac{b}{p-1/2} \leq \frac{n_{R_1} - 1}{\max\{n_{R_1,x}, n_{R_1,-x}\}}.$$

Now note that

$$\frac{n_{R_1} - 1}{\max\{n_{R_1,x}, n_{R_1,-x}\}} \leq \frac{n_{R_1} - 1 + 2n_{R_2,-x}}{\max\{n_{R_1,x} + n_{R_2,-x}, n_{R_1,-x} + n_{R_2,-x}\}}$$

as the latter fraction is increasing in  $n_{R_2,-x}$  and therefore truthtelling is still optimal in  $\tilde{R}_1$ . Clearly, truthtelling is also optimal in  $\tilde{R}_2$  as only players with bias  $x$  are left there. Consequently, the total number of pieces of information in  $\tilde{R}_1$  and  $\tilde{R}_2$  is  $(n_{R_1} + 2n_{R_2,-x})^2 + (n_{R_2,x} - n_{R_2,-x})^2$  which, by  $n_{R_1} \geq n_{R_2} > n_{R_2,x}$ , is strictly greater than the number of pieces of information in rooms  $R_1$  and  $R_2$ . Hence, the planner prefers the room assignment  $\tilde{R}_1, \tilde{R}_2$  over the room assignment  $R_1$  and  $R_2$  (keeping room assignment for players not in those rooms fixed).

Second, consider the case where more players are in room  $R_2$ , i.e.  $n_{R_1,x} + n_{R_1,-x} < n_{R_2,x} + n_{R_2,-x}$ . In this case, we argue that merging the two rooms to  $R_1 \cup R_2$  will yield more information than keeping them separate. By lemma 5, players with bias  $x$  will still be truthtelling in room  $R_1 \cup R_2$ . Assume that players with bias  $-x$  will not tell the truth in  $R_1 \cup R_2$  (otherwise merging the two rooms is clearly optimal). Note that this implies  $n_{R_1,x} + n_{R_2,x} > n_{R_1,-x}, n_{R_2,-x}$  as players with bias  $x$  tell the truth in  $R_1 \cup R_2$  and players with bias  $-x$  do not. The number of pieces of information generated in  $R_1 \cup R_2$  is then  $(n_{R_1,x} + n_{R_2,x})(n_{R_1,x} + n_{R_2,x} + n_{R_1,-x} + n_{R_2,-x})$ . This is greater than the number of pieces of information generated in  $R_1$  and  $R_2$  separately as

$$\begin{aligned} & n_{R_1,x}^2 + 2n_{R_1,x}n_{R_2,x} + n_{R_1,x}n_{R_1,-x} + n_{R_1,x}n_{R_2,-x} + n_{R_2,x}^2 + n_{R_2,x}n_{R_1,-x} + n_{R_2,x}n_{R_2,-x} \\ & > n_{R_1,x}^2 + 2n_{R_1,x}n_{R_1,-x} + n_{R_1,-x}^2 + n_{R_2,x}^2 + n_{R_2,x}n_{R_2,-x} \\ & \Leftrightarrow 2n_{R_1,x}n_{R_2,x} - n_{R_1,x}n_{R_1,-x} + n_{R_1,x}n_{R_2,-x} + n_{R_2,x}n_{R_1,-x} - n_{R_1,-x}^2 > 0 \\ & \Leftrightarrow -n_{R_1,-x}n_{R_1} + n_{R_1,x}n_{R_2} + n_{R_2,x}n_{R_1} > 0 \end{aligned}$$

which holds true by  $n_{R_1} < n_{R_2}$  and  $n_{R_1,x} + n_{R_2,x} > n_{R_1,-x}, n_{R_2,-x}$ . Consequently, the planner would prefer  $R_1 \cup R_2$  to  $R_1$  and  $R_2$  separately.  $\square$

**Lemma 7.** *In a welfare maximizing room allocation, the following cannot occur: There are three rooms  $R_0$  populated only of players with bias 0,  $R_b$  populated only with players of bias  $b$  and  $R_m$  populated with players of both biases.*

**Proof of lemma 7:** If  $R_m$  induces babbling, it is clearly better to assign the bias 0 ( $b$ ) players in there to  $R_0$  ( $R_b$ ) instead. If only players with bias  $x \in \{0, b\}$  send truthful

messages in  $R_m$ , then it is clearly better to merge this room with  $R_x$  which maintains truthtelling incentives for players with bias  $x$ , see lemma 5.

Hence, we only have to consider the case where players with both biases send truthful messages in  $R_m$ . For concreteness assume there are more players in  $R_0$  than in  $R_b$  denoted by  $n_0 \geq n_b$  (the reverse case is analyzed analogously). We consider two cases in turn. First, consider  $n_0 \geq n_m + n_b$ . We claim that merging  $R_m$  and  $R_0$  will then lead to more information than keeping these rooms separate. By lemma 5, player with bias 0 will still be truthtelling in  $R_0 \cup R_m$  and the number of pieces of information generated by rooms  $R_b$  and  $R_0 \cup R_m$  is  $n_b^2 + (n_0 + n_m)(n_m + n_0)$  which is strictly higher than the number of pieces of information generated by  $R_0$ ,  $R_b$  and  $R_m$ , i.e.  $n_b^2 + n_0^2 + n_m^2$ , as  $n_0 \geq n_m + n_b > n_m$ .

Second, consider  $n_0 < n_m + n_b$ . The following change in the room allocation creates more information: Move  $n_b$  players from  $R_b$  and  $n_b$  players from  $R_0$  to  $R_m$ . This leaves no one in  $R_b$ ,  $n_m + 2n_b$  in  $R_m$  and  $n_0 - n_b$  players in  $R_0$ . As in the proof of lemma 6, the move maintains truthtelling incentives for players in  $R_m$ . The number of pieces of information generated after the move is  $(n_m + 2n_b)^2 + (n_0 - n_b)^2 = n_m^2 + 4n_b n_m + 5n_b^2 + n_0^2 - 2n_0 n_b$  which is higher than the number of pieces of information generated by  $R_0$ ,  $R_b$  and  $R_m$  without the move, i.e.  $n_0^2 + n_m^2 + n_b^2$ , by  $n_0 < n_b + n_m$ .  $\square$

**Corollary 3.** *The welfare optimal room assignment consists of at most two rooms and at most one room in which players of both biases are present.*

**Proof of corollary 3:** This follows from the combination of corollary 2 and lemmas 4, 6 and 7.  $\square$

Corollary 3 (and the preceding lemmas) leave the following possibilities for welfare optimal room assignment:

- segregation: each bias group has its own exclusive room.
- full integration: one room for all players
- mix: one room exclusively with players of bias  $x$  and one room with both bias types in which either
  - only players of type  $-x$  send truthful messages
  - all players send truthful messages.

The idea behind the mix situation is the following: If there are more of players with bias  $-x$  than players with bias  $x$  and assigning all player to one room would lead to babbling, then it can be optimal not to separate completely but to assign some players of the minority  $x$  to the majority room  $-x$ . The  $x$  players will babble there but they receive a lot of information (truthful messages of all the  $-x$  players). If such a situation is optimal, then clearly it must be the case that assigning one more minority player to the mixed room would lead to babbling. This is the first mixed assignment possibility.

The second mixed assignment possibility refers to a situation where the minority bias group would babble if all players were in one room while the majority would send truthful messages. In this case, taking some players of the majority to a separate room can restore truthtelling by both groups in the mixed room and might be optimal.

As should be clear from the discussion above, the mixed scenario is only optimal if the group sizes differ. With equal group size either segregation or full integration is optimal. Which of the two is optimal depends on whether the bias difference  $b$  is sufficiently small to obtain truthtelling in a fully integrated room.

**Proposition 10.** *Let the number of players  $n$  be strictly greater than 2 and let the two bias groups be of equal size, i.e.  $n_b = n_0 = n/2$ . Then assigning all players to the same room is welfare optimal if  $b/(p - 1/2) \leq 2(n - 1)/n$ . Otherwise, segregation, i.e. assigning all players with bias 0 to a room  $R_1$  and all players with bias  $b$  to a room  $R_2$  is welfare maximizing.*

**Proof of proposition 10:** If  $b/(p - 1/2) \leq 2(n - 1)/n$ , then a fully integrated room leads to truthtelling and is therefore clearly optimal. Therefore, consider  $b/(p - 1/2) > 2(n - 1)/n$ . One fully integrated room leads then to babbling by all players and full integration is therefore worse than segregation. Note furthermore that the alternatives with a mixed room are not feasible with  $n_b = n_0$  and  $b/(p - 1/2) > 2(n - 1)/n$ : A mixed room containing all players of bias  $x$  will lead to babbling of players with bias  $-x$  by  $b/(p - 1/2) > 2(n - 1)/n$ . Hence, the scenario of a mixed room in which all players tell the truth is not feasible. A mixed room containing all players of bias  $x$  and babbling of those players in the mixed room with bias  $-x$  might be feasible is, however, clearly worse than segregation as it creates less pieces of information:  $n/2(n/2 + n_{m,-x}) + (n/2 - n_{m,-x}) < 2(n/2)^2$  for  $n_{m,-x} > 0$ . All other room assignments cannot be optimal by corollary 3.  $\square$

Taking the previous proposition and the earlier results on the existence of segregation and full integration equilibrium together yields the following result.

**Proposition 11.** *Let  $n_0 = n_b = n/2$  and  $n > 2$ . Then the welfare optimal room assignment is an equilibrium if  $\alpha \geq 2/(n - 2)$ . If  $\alpha < 2/(n - 2)$ , then the welfare optimal room assignment is still an equilibrium if  $b/(p - 1/2) \notin (2(n - 1)/n, n/2)$ .*

Now turn to the case where  $n_0 > n_b$ . If  $b/(p - 1/2) \leq (n_0 + n_b - 1)/n_0$ , then full information, i.e. all players in one room and every player sends a truthful message, is feasible and a single room is obviously welfare optimal.

If  $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$ , then players with bias  $b$  would babble in a single room. It is important to notice that segregation cannot be optimal in this situation: segregation leads to  $n_b^2 + n_0^2$  pieces of information while in one single room there would still be  $n_0 * (n_0 + n_b)$  pieces of information which is greater than  $n_b^2 + n_0^2$  by



$n_0 > n_b$ .<sup>2</sup> Consequently, there are two options for the welfare maximal room assignment. Either all players are in one single room and babbling by players with bias  $b$  is tolerated or some players with bias 0 are assigned to a separate room in order to balance the mixed room and restore truthtelling incentives for the bias  $b$  players. The maximal number of bias 0 players that can remain in the mixed room without inducing babbling by bias  $b$  players is  $n_{m0} = \lfloor (n_b - 1)/(b/(p - 1/2) - 1) \rfloor$ . Furthermore, truthtelling by the majority group clearly also requires that players with bias 0 are (weakly) in the majority. Hence, consider for now  $n_{m0} \geq n_b$ . The number of pieces of information in the scenario with  $n_{m0}$  bias 0 players and all bias  $b$  players in one room and  $n_0 - n_{m0}$  bias 0 players in a separate room is  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2$ . Whether this is higher or lower than the number of pieces in a fully integrated room, i.e.  $n_0(n_0 + n_b) + n_b$ , depends on the parameters. In particular, if  $b/(p - 1/2)$  is close to the lower boundary (and  $n_b$  is not too small), two rooms will be optimal as  $n_{m0}$  will be relatively high. If  $b/(p - 1/2)$  is close to the upper boundary, however,  $n_{m0}$  will be low and one room with babbling by the minority will be optimal. Finally, if  $n_{m0} < n_b$ , then babbling by one group occurs in any mixed room and therefore it is optimal to have one integrated room in which the minority players babble.

Finally, we consider  $b/(p - 1/2) > (n_0 + n_b - 1)/n_b$ . In this case, even the bias 0 players will babble if all players are in one room. This implies that putting some bias zero players in an own separate room will no longer help: All the remaining bias zero players would have even higher incentives to babble and it would be more informative to fully separate the two bias groups. Consequently, only two options remain: Either the just mentioned segregation or enough bias  $b$  players are assigned to an own separate room to restore truthtelling incentives for the bias 0 players in the mixed room. The maximum number of bias  $b$  players that can remain in the mixed room without destroying truthtelling by bias 0 players is  $n_{mb} = \lfloor (n_0 - 1)/(b/(p - 1/2) - 1) \rfloor$ . This yields  $n_0(n_0 + n_{mb}) + n_{mb} + (n_b - n_{mb})^2$  pieces of information which can, depending on the parameters, be higher or lower than the  $n_0^2 + n_b^2$  pieces of information created by segregation. What is important to notice is that for  $b/(p - 1/2) > n_0$  we get  $n_{bm} = 0$  and in this case segregation is always welfare optimal.

We summarize the preceding discussion in the following proposition.

**Proposition 12.** *Let  $n_0 \geq n_b$ . The welfare maximizing room allocation is as follows:*

1. *For  $b/(p - 1/2) \leq (n - 1)/n_0$ , all players are in one room (where their messages are truthtelling).*
2. *For  $(n - 1)/n_0 < b/(p - 1/2) \leq (n - 1)/n_b$ , let  $n_{m0} = \lfloor (n_b - 1)/(b/(p - 1/2) - 1) \rfloor$ .  
Then,*

---

<sup>2</sup>Similarly, it is not optimal to put only some bias  $b$  players into their own room: As  $n_b \leq n_0$ , they would have less information there than they would have if they were babbling in one single room.

- (a) all players in one room (where those with  $b_i = 0$  are truthtelling and those with  $b_i = b$  are babbling) is optimal if  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 \leq n_0(n_0 + n_b) + n_b$  and  $n_{m0} \geq n_b$ ,
- (b) one room with  $n_0 - n_{m0}$  players with  $b_i = 0$  and another room with all other players (in which everyone is truthtelling) is optimal.

3. For  $b/(p - 1/2) > (n - 1)/n_b$ , let  $n_{mb} = \lfloor (n_0 - 1)/(b/(p - 1/2) - 1) \rfloor$ . Then,

- (a) one room with  $n_b - n_{mb}$  players with  $b_i = b$  and another room with all other players is optimal if  $n_0(n_0 + n_{mb}) + n_{mb} + (n_b - n_{mb})^2 \geq n_0^2 + n_b^2$
- (b) full segregation is optimal otherwise.

Note that an increase in  $b/(p - 1/2)$  will decrease both  $n_{m0}$  and  $n_{mb}$ . This implies that we can simply step through the different cases of proposition 12 as  $b/(p - 1/2)$  increases. However, depending on parameter values, in particular  $n_0$  and  $n_b$ , some of the cases may be skipped. However, the first and last case will never be skipped as they always occur for sufficiently low, respectively high, values of  $b/(p - 1/2)$ .

### 1.3. Welfare optimal room assignment as equilibrium

It remains to check when the welfare optimal room assignments correspond to equilibria of the game. If full information is feasible, i.e. if  $b/(p - 1/2) \leq (n_0 + n_b - 1)/n_0$ , then it is clearly an equilibrium. For  $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$ , the welfare optimal room assignment is definitely an equilibrium if it is optimal to keep all players in the same room (and have the smaller group babbling): The point is that this can only be optimal if isolating one player of the larger group would not lead to truthtelling in a room with all other players. (If this was the case, then isolating one player of the larger group would be optimal.) But this implies that any deviation in room choice from the situation with one big mixed room will lead to less information for the deviating player and less (or the same) information for the other players. Following (1), such a deviation will therefore reduce the deviating player's expected payoff. Hence, the welfare maximizing room assignment is an equilibrium if  $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$  and  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 \leq n_0(n_0 + n_b)$ . For the case  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 > n_0(n_0 + n_b)$  where two rooms are optimal – one mixed room with truthtelling by all players and one with only players of the larger group – the relevant question is whether the  $n_0 - n_{m0}$  players in the room only for bias 0 players would want to deviate to the mixed room. Note that this deviation will destroy truthtelling by the bias  $b$  players. However, a unilaterally deviating player will obtain information from  $n_{m0}$  other players which is more information than the  $n_0 - n_{m0} - 1$  truthful messages he obtains when not deviating.<sup>3</sup> It is therefore

<sup>3</sup>To see this, first note that  $n_{m0} \geq n_b$  as no truthtelling mixed room exists if bias  $b$  players were not truthtelling in a perfectly balanced, i.e. same number of players from each bias group, room. Then note

unsurprising that the welfare optimal room assignment will only be an equilibrium if  $\alpha$  is sufficiently high. Using (1), the deviation will not be profitable if and only if it decreases  $\zeta_i + \alpha \sum_{j \neq i} \zeta_j$ . This is the case if and only if

$$\begin{aligned} n_{m0} + 1 + \alpha \left( (n_0 - n_{m0} - 1)^2 + (n_{m0} + 1)(n_b + n_{m0}) \right) \\ \leq n_0 - n_{m0} + \alpha \left( (n_0 - n_{m0} - 1)(n_0 - n_{m0}) + (n_{m0} + n_b)^2 \right). \\ \Leftrightarrow \alpha \geq \frac{2n_{m0} - n_0 + 1}{n_0 - 2n_{m0} - 1 + n_b^2 + n_b(n_{m0} - 1)}. \end{aligned} \quad (3)$$

Hence, if  $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$  and  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 > n_0(n_0 + n_b)$  as well as  $n_{m0} \geq n_b$ , the welfare optimal room assignment is an equilibrium if and only if (3) holds.

For  $b/(p - 1/2) > (n_0 + n_b - 1)/n_b$ , the welfare optimal room assignment can be either a mixed room combined with a room exclusively for the smaller group or segregation. In the first case it is straightforward to see that this is an equilibrium. Recall that moving one more player from the separate room to the mixed room would lead to babbling by all players in the mixed room. Clearly, such a deviation is not profitable. In the second case, we already showed before that total separation is an equilibrium if either  $b/(p - 1/2) \geq n_0$  or  $\alpha \geq (1 + n_0 - n_b)/(n_b - 1)$ .

We summarize the prior discussion in the following proposition.

**Proposition 13.** *Let  $n_0 \geq n_b$ . The welfare optimal room allocation is an equilibrium in cases (1), (2a) and (3a) of proposition 12. The welfare optimal room allocation is also an equilibrium*

- in case (2b) if

$$\alpha \geq \frac{2n_{m0} - n_0 + 1}{n_0 - 2n_{m0} - 1 + n_b^2 + n_b(n_{m0} - 1)}$$

- in case (3b) if  $\alpha \geq (1 + n_0 - n_b)/(n_b - 1)$ .

If the welfare optimal room allocation is not an equilibrium, then the welfare optimal equilibrium features too little segregation: In case (2b), the equilibrium has all players in a single room while the welfare optimal allocation has two rooms. In case (3b), the welfare optimal room allocation is full segregation which is not an equilibrium.

## 2. Limit results to give lower bounds on polarization

To find out how polarized biases need to be so that segregation is optimal and an equilibrium, we can consider two stylized cases. First, we will consider biases that are evenly distributed over the population. The condition  $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 > n_0(n_0 + n_b)$  is more demanding for higher  $n_0$  (keeping other parameters fixed). It is straightforward to calculate that the inequality cannot be satisfied (given  $n_b \leq n_{m0}$ ) for  $n_0 \geq 5n_{m0}/3$ . Hence,  $n_{m0} \geq 4n_0/5$  whenever this room assignment is welfare maximizing which implies  $n_{m0} > n_0 - n_{m0} - 1$ .

distributed on an interval of the real line. We can think of this case as having “zero polarization”, whereas clustering of biases around certain values exhibits positive polarization. Second, we consider biases that are tightly clustered around a central value – we could think of this as “negative polarization”.

## 2.1. Uniformly distributed biases

We will show the following result:

**Result 1.** *Let  $b_i = (i - 1) * k / (n - 1)$  for  $i = 1, \dots, n$ . Then the welfare optimal room allocation assigns either all players to one single room or all but one extreme player to the same room. Assigning all players to the same room is also an equilibrium.*

Intuitively, we can start by considering the fully integrated room, in which some people whose biases are close to the overall average tell the truth, and the rest babble and learn from the truth-tellers. Since biases are evenly distributed by assumption, there is little welfare to be gained by moving the bias average around by allocating people to another room. (This can only work because of integer effects – i.e. because changes in the average bias have discrete effects on who tells the truth – which is precisely what gives us the exceptions in the second half of the proposition.) Any room that includes only part of the players will have a shorter truth-telling interval, which (again, absent integer effects) means fewer truth-tellers. But if we cannot increase the number of truth-tellers by segregating into smaller rooms, then the fully integrated room must be welfare-optimal and also an equilibrium: Every player receives the highest possible number of truthful messages while the number of players having their own signal in addition to this number of messages is also maximal. The remainder of this section shows this result formally.

**Proposition 14.** *Let  $b_i = (i - 1) * k / (n - 1)$  for  $i = 1, \dots, n$ . Then one single room with all players is both welfare optimal and an equilibrium if either*

$$\underline{b} - \left[ k/2 - (p - 1/2) \frac{n - 1}{n} \right] \leq \frac{k}{2(n - 1)} \quad (4)$$

or

$$k(1 - 1/n)/2 - (p - 1/2) \frac{n - 2}{n - 1} > \underline{b} - \frac{k}{n - 1}. \quad (5)$$

*If neither of these two conditions holds, isolating player  $n$  in one room and all other players in one room is welfare optimal. This is only an equilibrium if*

$$\alpha \geq \frac{\lfloor \frac{n-1}{n} (2p - 1) / (k / (n - 1)) \rfloor - 1}{n - 2}. \quad (6)$$

**Proof of proposition 14:** Theorem 1 states that in the most informative equilibrium of the messaging subgame players in room  $R$  will tell the truth if and only if  $b_i \in \left[ \bar{b} - \frac{n_R - 1}{n_R} (p - \frac{1}{2}), \bar{b} + \frac{n_R - 1}{n_R} (p - \frac{1}{2}) \right]$ . If this interval covers  $[0, k]$ , then one room leads

to truthtelling by all players and one single room is clearly optimal. In the remainder of this proof, we therefore assume that this is not the case. The length of the interval  $\left[\bar{b} - \frac{n_R-1}{n_R}(p - \frac{1}{2}), \bar{b} + \frac{n_R-1}{n_R}(p - \frac{1}{2})\right]$  is  $\frac{n_R-1}{n_R}(2p - 1)$ . The number of players telling the truth in any room is consequently bounded from above by  $\lfloor \frac{n_R-1}{n_R}(2p - 1)/(k/(n - 1)) \rfloor + 1$  as the players' biases are equally spaced with distance  $k/(n - 1)$  between two consecutive players' biases. This bound may not be attained by any feasible room due to the discrete nature of the problem. More specifically, if we take the fully integrated room, then the number of truthtelling players will be either  $\lfloor \frac{n-1}{n}(2p - 1)/(k/(n - 1)) \rfloor + 1$  or  $\lfloor \frac{n-1}{n}(2p - 1)/(k/(n - 1)) \rfloor$ .

Let  $t^*$  be the maximal number of truthtelling players in any possible room. From the above, it is clear that  $t^* \in \left\{ \lfloor \frac{n-1}{n}(2p - 1)/(k/(n - 1)) \rfloor + 1, \lfloor \frac{n-1}{n}(2p - 1)/(k/(n - 1)) \rfloor \right\}$ . Suppose  $t^*$  is the number of truthtelling players if all players are in the same room. Then the number of pieces of information generated in this room is  $t^*n + n - t^*$ . We will show that in this case no other room configuration generates more pieces of information: The total number of pieces of information in  $r$  rooms is:  $\sum_R t_R n_R + n_R - t_R = \sum_R t_R (n_R - 1) + n_R \leq \sum_R t^* (n_R - 1) + n_R = t^*(n - r) + n \leq t^*n + n - t^*$ . By proposition 1, one big room with all players is then welfare optimal if this leads to  $t^*$  truthtelling players.

Next consider the situation where one integrated room with all players leads not to  $t^*$  but only to  $t^* - 1$  truthtelling players. Suppose that there is some room  $R^*$  with  $n - 1$  players in which  $t^*$  players are truthtelling. We show that in this case the room configuration  $(R^*, \{1, \dots, n\} \setminus R^*)$  is welfare optimal. This will lead to  $t^*(n - 1) + n - 1 - t^* + 1 = t^*(n - 2) + n$  pieces of information. The big integrated room leads to only  $(t^* - 1)n + n - t^* + 1 = t^*(n - 1) < t^*(n - 1) - t^* + n$  pieces of information and is therefore welfare inferior. Any other room configuration with  $r$  rooms leads to  $\sum_R t_R n_R + n_R - t_R = \sum_R t_R (n_R - 1) + n_R \leq \sum_R t^* (n_R - 1) + n_R = t^*(n - r) + n \leq t^*(n - 2) + n$  pieces of information which is also (weakly) less than  $(R^*, \{1, \dots, n\} \setminus R^*)$ . Hence, in this case  $(R^*, \{1, \dots, n\} \setminus R^*)$  is welfare optimal.

Finally, we show that the conditions in the proposition lead to either of the two just described cases. Note that in the fully integrated room  $\bar{b} = k/2$ . Hence, condition (4) states that the distance from the lowest player's bias who tells the truth to the lower boundary of the truthtelling interval is less than  $1/2$  the distance between two consecutive players' biases. By symmetry of the truthtelling interval around  $\bar{b}$  and the equal spacing of biases, this is also true for the distance of the highest bias player telling the truth and the upper boundary of the truthtelling interval. First, let (4) hold strictly. Then it is clear that shifting the truthtelling interval (by changing  $\bar{b}$ ) cannot lead to more players being truthtelling. Furthermore, the length of the truthtelling interval is strictly decreasing in the number of players in the room. Hence, in no other room can there be more truthtelling players than in the fully integrated room. This holds also if (4) holds with equality as the length of the truthtelling inequality is strictly decreasing in the number of players in the

room. Consequently,  $t^*$  is achieved by the fully integrated room and the argument two paragraphs above shows that then the fully integrated room is welfare optimal.

Now consider the case where (4) does not hold. Start from the fully integrated room. If (4) does not hold, shifting the truthtelling interval by  $k/(2(n-1))$  down (by – for now magically – reducing  $\bar{b}$  by this amount), will imply that this interval contains 1 more player than in the fully integrated room. Furthermore, the distance of this lowest truthtelling player after the shift to the lower boundary of the truthtelling interval will be less than  $k/(2(n-1))$  by the assumption that (4) did not hold. Now note that removing player  $n$  from the fully integrated room will reduce  $\bar{b}$  by exactly  $k/(2(n-1))$  (from  $k/2$  to  $(k - k/(n-1))/2$ ). But note that removing this player also implies that  $n_R = n - 1$  and therefore the length of the truthtelling interval is reduced. Condition (5) states that due to the shrinking of the interval when moving from  $n$  to  $n - 1$  players the one player whose truthtelling was gained by shifting the interval down is lost again. Furthermore, the “shrinking” occurs at the upper as well as the lower boundary to the same extent. This implies that also at the upper boundary one truthtelling player is lost due to the shrinking (while the shifting did not lose anyone as (4) was violated by assumption). Consequently, the room without player  $n$  will have one less truthtelling player than the fully integrated room if (4) is violated and (5) holds. In this case, no room with  $n - 1$  (or less) players can have more truthtelling players than the fully integrated room and therefore  $t^*$  is attained in the fully integrated room. Consequently, the fully integrated room is by the results above welfare optimal.

If neither (4) nor (5) holds, then the “shifting” argument above implies that the room allocation  $(\{1, \dots, n-1\}, \{n\})$  leads to one more truthtelling player in  $R^* = \{1, \dots, n-1\}$  than in the fully integrated room. Consequently,  $t^*$  is attained in  $R^*$  and  $(\{1, \dots, n-1\}, \{n\})$  is welfare optimal by the results above.<sup>4</sup>

In terms of equilibrium, it is immediate that no player wants to deviate from the fully integrated room by isolating himself as self-isolation leads to less information for himself and no more information for other players. The same argument applies for players in room  $R^*$  in case (4) and (5) are violated. However, the isolated player might have an incentive to join  $R^*$ : This would reduce the amount of information as only  $t^* - 1$  instead of  $t^*$  players would be truthtelling in the resulting fully integrated room reducing the number of pieces information of all other players in this room from  $t^*(n-1) + n - 1 - t^*$  to  $(t^* - 1)(n-1) + n - t^*$ . However, the deviating player would gain more information for himself, i.e. the number of pieces of information he observes is  $t^*$  instead of 1. From 6, it follows that the deviation is profitable if and only if  $\alpha < (t^* - 1)/(n - 2)$ . Note that  $t^* = \lfloor \frac{n-1}{n}(2p-1)/(k/(n-1)) \rfloor$  in the here analyzed case where one integrated room is not optimal. This gives the condition in the proposition.  $\square$

---

<sup>4</sup>It should be noted that similar arguments as above, with an upward instead of a downward shift, lead to the optimality of  $(\{1\}, \{2, \dots, n\})$  which will also attain  $t^*$  if (4) and (5) are violated.

### 3. Symmetrically single peaked bias distribution

We now move to symmetrically, single peaked distribution of biases: Assume that biases are on an equally spaced grid  $0, d, 2d, \dots, Kd$  for some  $d > 0$  and  $K \in \mathbb{N}$ . The number of players with bias  $b_i = kd$  is increasing up to  $Kd/2$  and decreasing thereafter. Furthermore, we assume that the number of players with bias  $kd$  equals the number of players with bias  $(K - k)d$  for  $k = 0, 1, \dots, \lfloor K/2 \rfloor$ .

We are going to show a result similar to the one for uniformly distributed biases in the paper. Namely, that a fully integrated room is welfare optimal and an equilibrium under certain conditions.

To state the proposition we need the following notation: Let  $\underline{k}$  be the lowest  $k$  such that  $kd \geq Kd/2 - (p - 1/2)(n - 1)/n$  and let  $\bar{k}$  be the highest  $k$  such that  $kd \leq Kd/2 + (p - 1/2)(n - 1)/n$ . Note that due to the discreteness of the grid and following theorem 1, the truthtelling interval in a fully integrated room will cover all players with  $b_i \in [\underline{k}d, \bar{k}d]$ .

**Proposition 15.** *With a symmetric, single peaked distribution of biases, one room containing all players is welfare optimal and also an equilibrium if*

$$\bar{k}d - \underline{k}d + d > (2p - 1) \frac{n - 2}{n - 1}. \quad (7)$$

**Proof of proposition 15:** Theorem 1 states that in the most informative equilibrium of the messaging subgame players in room  $R$  will tell the truth if and only if  $b_i \in \left[ \bar{b} - \frac{n_R - 1}{n_R} \left( p - \frac{1}{2} \right), \bar{b} + \frac{n_R - 1}{n_R} \left( p - \frac{1}{2} \right) \right]$ . If this interval covers  $[0, Kd]$ , then one room leads to truthtelling by all players and one single room is clearly optimal. In the remainder of this proof, we therefore assume that this is not the case. Note that – holding  $\bar{b}$  fixed – the length of the interval is increasing in  $n_R$ . If we turn to the case of one fully integrated room, then the truthtelling interval is  $\left[ Kd/2 - \frac{n - 1}{n} \left( p - \frac{1}{2} \right), Kd/2 + \frac{n - 1}{n} \left( p - \frac{1}{2} \right) \right]$  as  $\bar{b} = Kd/2$ . We will first show the result under a condition slightly stronger than (7), namely under the condition

$$\bar{k}d - \underline{k}d + d > (2p - 1) \frac{n - 1}{n}. \quad (8)$$

Condition (8) states that the length of the truthtelling interval is less than  $\bar{k} - \underline{k} + d$ . (Note that the length of the truthtelling interval is weakly larger than  $\bar{k}d - \underline{k}d$  due to the discrete grid on which biases are distributed.) This implies that the truthtelling interval would not cover more grid points if it was moved up or down while keeping its length constant. As the truthtelling interval is shorter for any other room (because of  $n_R < n$ ) and the distribution of biases is single-peaked, this implies that there is no room in which more players are truthtelling than in the fully integrated room.

The same conclusion follows if (7) holds instead of (8): (7) states that the length of the truthtelling interval in any room different from the fully integrated room (which

therefore contains at most  $n - 1$  players) is less than  $\bar{k}d - \underline{k}d + d$  which again implies that the truth-telling interval of such a room cannot cover more grid points than the fully integrated room and by single peakedness it can therefore also not contain more truth-telling players.

Let  $t^*$  be the maximal number of truth-telling players in any possible room. From the above,  $t^*$  is attained by the fully integrated room if (7) holds. In this case, the number of pieces of information generated in the fully integrated room is  $t^*n + n - t^*$ . We will show that no other room configuration generates more pieces of information: The total number of pieces of information in  $r$  rooms is:  $\sum_R t_R n_R + n_R - t_R = \sum_R t_R (n_R - 1) + n_R \leq \sum_R t^* (n_R - 1) + n_R = t^*(n - r) + n \leq t^*n + n - t^*$ . By proposition 1, one big room with all players is therefore welfare optimal if (7) holds.

In case a single fully integrated room is welfare maximal it is also an equilibrium: Unilateral self-isolation would lead to less information for the deviating player and also – by welfare optimality of the fully integrated room – to less information over all. By 1, the deviation is therefore unprofitable.  $\square$

#### 4. Example: Too much segregation in equilibrium

In the case of two bias groups, we have shown that the welfare-optimal equilibrium room allocation is either the overall welfare-optimum, or has too little segregation compared to it. If there are three or more bias groups, this is not generally true anymore – now it is possible that the welfare-optimal equilibrium involves *too much* segregation compared to the welfare-optimum. This can occur when a player wants to deviate from the welfare-optimum to another room where he can learn more and thereby destroys the truth-telling incentives of people in the room that he is leaving. The following paragraphs provide an example for such a situation.

Consider a bias configuration in which 13 people have bias  $b_1 = -1000$ , 10 people have bias  $b_2 = 0$  and 2 people have bias  $b_3 = 500$ . We can easily see that there exists no possible room with members of exactly two bias groups in which anyone tells the truth. Even in rooms that involve all three bias groups, no one with biases  $b_1$  and  $b_3$  will ever tell the truth. The only way to get anyone with bias 0 to tell the truth in a mixed room is to create a room with one person with bias  $b_1$ , two people with bias  $b_3$  and an arbitrary number of people with bias 0. This leads us to the welfare-optimal room allocation: Room 1 consists of 12 people with bias  $b_1$  and generates 144 pieces of information, and room 2 contains everybody else and generates 133 pieces of information, for a total  $\sum \zeta_i = 277$ . For low  $\alpha$ , this allocation is not an equilibrium: The person with bias  $b_1$  in room 2 can change to room 1 and have 13 pieces of information instead of 11.

Now consider the room allocation where bias groups are fully segregated: This generates  $169 + 100 + 4 = 273$  pieces of information and is also an equilibrium: No one can learn



anything by switching to another room. Hence, this is the welfare-optimal equilibrium, while the first allocation we described is welfare-optimal – which means that there is too much segregation in the welfare-optimal equilibrium. (Note that this example is generic in the sense that we could find an open ball of bias configurations around this particular bias configuration in which our conclusions remain valid.)

## 5. Uncertainty

### 5.1. Main results and intuition

Let all biases  $b_i$  be randomly and independently distributed on  $\mathbb{R}$  according to distribution  $F_i$ . Each player observes his own bias  $b_i$ , but only knows the distributions of the biases of other players. Let  $b_i^e = \int_{-\infty}^{\infty} b_i dF_i$  be the expected value of  $b_i$ . This can be thought of as a generalization of the paper’s main model, in which all biases were always identical to their expected value. When we talk about “introducing” or “adding” uncertainty in this context, we think of starting with the model in which all biases are known with certainty, and replacing each bias with a bias distribution that has the same expected value. Throughout this section, we will be comparing across distributions that have the same expected value. The following paragraphs intuitively analyze the model with uncertainty; the corresponding formal statements and analysis are in section 5.2 below.

To find the messaging equilibria within a room, we need to consider  $i$ ’s problem of choosing a message  $m_i$  after observing  $b_i$  and  $\sigma_i$ , but only knowing  $F_j$  for all  $j \in R_i$ . We can show that this problem is very similar to knowing all biases with certainty. In particular, recall that  $i$ ’s willingness to tell the truth depended only on the distance between  $b_i$  and the average of all other  $b_j$ ’s in the model with certainty. This insight applies analogously to a model in which all biases are unknown: Now  $i$  cares only about the difference between  $b_i$  and the average of all  $b_j^e$ , i.e. the expected values of other people’s bias.

A difference in describing equilibria with uncertainty arises since  $i$  may want to tell the truth for some values of  $b_i$  and not for others, and the other players are unsure about  $b_i$  when interpreting  $m_i$ . Their belief about how likely  $i$  is to tell the truth hence depends on how  $b_i$  is distributed. For each possible probability with which  $i$  tells the truth, there exists an interval around  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1}$  such that  $i$  wants to tell the truth if the realized  $b_i$  lies within this interval. Since the distribution of  $b_i$  is common knowledge, that gives us the following equilibrium condition: The beliefs of all other players about  $i$ ’s probability of truth-telling need to give rise to a truth-telling interval for  $i$  around the average of all  $b_j$  such that  $i$  wants to tell the truth with exactly the probability with which the other players believe that he tells the truth.

This translates into a slightly generalized version of theorem 1 which, for any distribution of  $b_i$ , gives us the highest probability with which  $i$  can tell the truth in any equilibrium. Intuitively, the more concentrated  $F_i$  is around  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1}$ , the higher the

probability with which  $i$  can tell the truth in equilibrium. Interestingly, only the probability mass of  $F_i$  that is sufficiently close to  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1}$  matters; whether or not  $b_i^e$  itself is close to the average or not is not directly relevant for whether  $i$  is able to tell the truth in equilibrium.

In particular, this means that we can choose any set of expected biases, regardless of how close they are to each other, and construct bias distributions such that none of the players ever wants to tell the truth to anyone in any room allocation. This means that for any bias configuration, uncertainty has the potential to completely destroy all chances of creating a room in which information is exchanged.

**Proposition 16.** *Take a set of  $n$  players with biases  $\{b_1, b_2, \dots, b_n\}$  such that there exists a room allocation in which some (or all) players tell the truth. Then there exists a set of probability distributions  $\{F_1, F_2, \dots, F_n\}$  of biases with expected values  $\{b_1, b_2, \dots, b_n\}$  such that in any room allocation of the  $n$  players, no player will tell the truth in any equilibrium. (Proof on page 24.)*

This is, of course, a very stark result. Uncertainty need not always destroy communication. It can, in fact, make communication possible where it was previously impossible, by moving probability mass of  $b_i$ 's distribution closer to the average of other biases. This effect, however, is more limited and can never lead to full truth-telling if there is no full truth-telling in a model with certain biases and identical expected values.

**Proposition 17.** *If  $b_i$  is such that there exists no equilibrium in room  $R_i$  where  $i$  tells the truth, there exists a distribution  $F_i$  with expected value  $b_i^e = b_i$  such that there exists an equilibrium in  $R_i$  where  $i$  tells the truth with positive probability. However, there exists no  $F_i$  such that  $i$  tells the truth with probability 1 in any equilibrium. (Proof on page 24.)*

While uncertainty can make some truth-telling possible where it was not possible with certainty, large amounts of uncertainty will always destroy any truth-telling and make all messages arbitrarily uninformative unless they preserve sufficient probability mass in the neighborhood of  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1}$ . Because of the large space of possible distributions and possible orderings on uncertainty, we show this result in two ways. First, we consider any continuous bias distribution and show that by “stretching” it, any equilibrium will become arbitrarily uninformative. Then we consider discrete bias distributions with bounded support, and show that any way of increasing the variance of such a distribution will likewise eventually erode all informative equilibria. In the following propositions,  $\mu_{ji}^l$  is  $j$ 's belief about  $\theta_i$ , given that  $i$  has sent the signal  $m^l$ ; the other expressions involving  $\mu$  are defined analogously.

**Proposition 18.** *Let  $F$  be a continuous distribution function that is continuous at its expected value  $b_i^e$  and symmetric around  $b_i^e$ . Let  $F^\kappa(x) = F(b_i^e + \kappa(x - b_i^e))$ , i.e.  $b_i = b_i^e$  almost surely for  $\lim_{\kappa \rightarrow \infty} F^\kappa$ . For any  $F$  and  $\varepsilon > 0$ , there exists a  $\bar{\kappa} > 0$  such that  $\mu_{ji}^h - \mu_{ji}^l < \varepsilon$  if  $F_i = F^\kappa$  and  $\kappa \leq \bar{\kappa}$ . (Proof on page 24.)*

**Proposition 19.** Fix the expected bias  $b_i^e$  of all players in a given room and a bounded support for all bias distributions  $F_i$ . Assume that there is at least one element in the support that is smaller than  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} - (2p - 1)$  and at least one element that is larger than  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + (2p - 1)$ . Then for each  $\varepsilon > 0$  there exists some  $\overline{\sigma_{F_i}}$  such that for all such  $F_i$  with  $\text{Var}(b_i) \geq \overline{\sigma_{F_i}}^2$ ,  $\mu_{j_i}^h - \mu_{j_i}^l \leq \varepsilon$ . (Proof on page 24.)

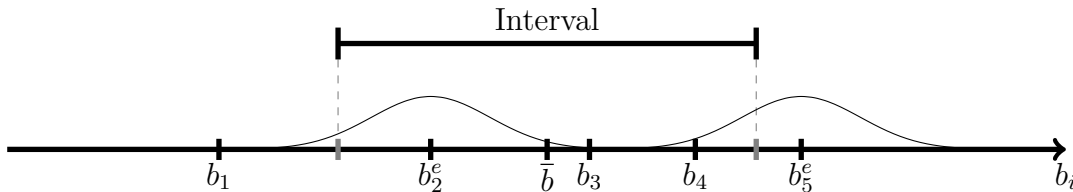


Figure 2: An illustration of propositions 17 to 19. (The biases are identical to the one in figure 1 of the paper except that  $b_2$  and  $b_5$  are now uncertain.)

Figure 2 illustrates propositions 17 to 19. The bias configuration is identical to the one in figure 1 on page 12, except that there is some mean-preserving uncertainty about the biases of players 2 and 5, whose biases are now distributed according to a bell-shaped distribution function. Under certainty, player 2 was telling the truth, but is now only telling the truth if his realized  $b_2$  falls within the interval (proposition 18). Player 5 was babbling, but will now sometimes send an informative message if his realized bias is close enough to  $\bar{b}$  (proposition 17).<sup>5</sup>

These results already contain statements about room choice with uncertainty: If truth-telling is greatly reduced or becomes impossible, there is not much to be gained from being in one room. Of course, truth-telling between people with identical bias distributions is not necessarily easier – note that proposition 16 contained no assumption that people differ in how their biases are distributed. So are the effects of uncertainty simply to make communication hard in general? Not necessarily. Consider a model where full integration is welfare-optimal and an equilibrium if biases are known. We can show that for any such model, uncertainty can cause segregation between groups to become Pareto-superior to integration, and such segregation may also be an equilibrium of the room choice game.

**Proposition 20.** Let the number of players be weakly larger than 4 and let  $b_i^e \in \{0, b\}$ , with  $b \in (0, \frac{n-1}{n}(2p-1)]$ . Let the two bias groups be of equal size, i.e.  $n_0 = n_b = n/2$ . Then in the room-choice game:

- If  $b_i = b_i^e$  with certainty, the fully integrated room is welfare-optimal and an equilibrium.

<sup>5</sup>This graphic is meant as an illustration and ignores the fact that, while the interval's length remains constant, its precise location may shift depending on the exact beliefs of the receiving players in equilibrium.

- If biases are uncertain, we can find distributions  $F_i$  that keep all  $b_i^e$  constant such that full segregation between the two bias groups is welfare-optimal. For  $\alpha \geq \frac{2}{n-2}$ , this is also an equilibrium.

(Proof on page 25.)

To illustrate this result, let us return to the example on taxation from the paper’s introduction, and assume that the world consists of liberals and conservatives. Liberals generally prefer higher taxes than conservatives, but everybody is aware that the optimal tax level depends on how bad taxes are for economic growth. If the exact political preference of each person is known, an informative exchange is possible even across party lines as long as preferences are not too different. But now assume that instead, each member of each political group is either a moderate or an extremist. It is only observable whether anyone is liberal or conservative, not whether they are extremists or moderates. Both have equal probability, so that in expectation each person is still an “average” liberal or conservative.

Consider the problem of a liberal who is unsure whether he is listening to a moderate conservative or a conservative extremist. He knows that a conservative extremist would always tell a liberal that taxes are bad for the economy, regardless of what his information is. Any statement about the damages of taxes has hence become less informative, while being more likely to be made, than if the liberal was talking to an average conservative. The same is true for a conservative listening to a liberal. Yet while discussion across party lines has become less informative, this is not true for discussion within parties: The possible biases within groups are still close enough so that both moderates and extremists want to truthfully reveal their knowledge to other members of their party. It is hence better for liberals to only talk to other liberals and for conservatives to only talk to conservatives, than for any cross-party discussion to take place – not because of inherent differences in preferences, but because of uncertainty about who one’s interlocutor is.

## 5.2. Detailed analysis and proofs for the model with uncertainty

### 5.2.1. Preliminary analysis

Similarly to the derivation of expression (16), we can write

$$\begin{aligned}
 U_i(m^h) &= \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in R_i, j \neq i} \left( b_j - b_i + \mu_{ji}^h + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right] \\
 U_i(m^l) &= \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in R_i, j \neq i} \left( b_j - b_i + \mu_{ji}^l + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right].
 \end{aligned}$$

Note that we are interested in the difference of the two expressions. Hence, while all  $b_j$ s are now unknown, this uncertainty only matters where  $b_j$  is multiplied by  $\mu_{ji}^h$  and  $\mu_{ji}^l$ , respectively. We can hence write

$$\begin{aligned}\Delta U_i(\sigma_i) &= (U_i(m^h) - U_i(m^l))/\alpha \\ &= 2(\mu_{ji}^h - \mu_{ji}^l)(n_{R_i} - 1) \left[ -\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + b_i + \mathbb{E}[\theta_i | \sigma_i] \right], \quad (9)\end{aligned}$$

which is identical to (16) except that we have substituted  $b_j^e$  for  $b_j$ .  $i$ 's problem remains virtually unchanged, except that he now considers the expected value of biases of other people within the room.

Now consider  $i$ 's messaging strategy. In the following, let

$$\begin{aligned}\lambda^h &= \Pr(m_i = m^h | \sigma_i = \sigma^h) \text{ and} \\ \lambda^l &= \Pr(m_i = m^l | \sigma_i = \sigma^l)\end{aligned}$$

i.e.  $\lambda^h$  and  $\lambda^l$  are the marginal probabilities with which  $i$  truthfully reveals his signal, averaging over all possible bias types. For example, if  $b_i$  has two possible values with equal probability and  $i$  only reveals  $\sigma^h$  truthfully for one of them, then  $\lambda^h = \frac{1}{2}$ . The resulting beliefs of player  $j$  are

$$\begin{aligned}\mu_{ji}^h &= \frac{p\lambda^h + (1-p)(1-\lambda^l)}{1 + \lambda^h - \lambda^l} \\ \mu_{ji}^l &= \frac{p(1-\lambda^h) + (1-p)\lambda^l}{1 - \lambda^h + \lambda^l}.\end{aligned}$$

We can also write the following two terms, which both appear in equation (9):

$$\begin{aligned}
\mu_{ji}^h - \mu_{ji}^l &= \frac{2p\lambda^h + 2p\lambda^l - 2p - \lambda^h - \lambda^l + 1}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} \\
&= (2p - 1) \frac{(\lambda^h + \lambda^l - 1)}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} \tag{10}
\end{aligned}$$

$$\begin{aligned}
\mu_{ji}^h + \mu_{ji}^l &= \frac{2p\lambda^h - 2p(\lambda^h)^2 - 2p\lambda^l + 2p(\lambda^l)^2 - 2(\lambda^l)^2 - \lambda^h + \lambda^l + 2\lambda^h\lambda^l + 1}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} \\
&= \frac{4p(\lambda^l)^2 - 2(\lambda^l)^2 - 4p\lambda^h\lambda^l + 2\lambda^h\lambda^l + 2p\lambda^h - \lambda^h - 2p\lambda^l + \lambda^l - 2p + 1}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} + 2p \\
&= (2p - 1) \frac{2(\lambda^l)^2 - 2\lambda^h\lambda^l + \lambda^h - \lambda^l - 1}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} + 2p \\
&= (2p - 1) \left( \frac{(\lambda^l)^2 - \lambda^h\lambda^l - \lambda^l}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} + \frac{(\lambda^l)^2 - \lambda^h\lambda^l + \lambda^h - 1}{(\lambda^h - \lambda^l + 1)(\lambda^l - \lambda^h + 1)} \right) + 2p \\
&= (2p - 1) \left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right) + 2p. \tag{11}
\end{aligned}$$

From (10), we can see that the condition  $\mu_{ji}^h \geq \mu_{ji}^l$  translates to  $\lambda^h + \lambda^l \geq 1$ . We can distinguish two cases:

- $\lambda^h + \lambda^l = 1$ . Then  $\mu_{ji}^h - \mu_{ji}^l = 0$  and  $i$ 's messages are completely uninformative.
- $\lambda^h + \lambda^l > 1$ . We will focus on this case, in which messages by  $i$  have some informative content.

We can intuitively see that if  $i$ 's messages are believed to contain some information about  $\sigma_i$ ,  $i$  should never want to misrepresent  $\sigma^h$  if  $b_i$  is high compared to the average bias of other players (and vice versa if  $b_i$  is low). In fact, we can show the following result:

**Lemma 1.** *Assume that  $\lambda^h + \lambda^l > 1$ . Then  $i$  always strictly prefers to truthfully reveal (i)  $\sigma^h$  if  $b_i \geq \mathbb{E} \left[ \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} \right]$  and (ii)  $\sigma^l$  if  $b_i \leq \mathbb{E} \left[ \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} \right]$ .*

*Proof.* Consider case (i) and assume that the opposite was true, i.e.  $\Delta U_i(\sigma^h) \leq 0$  for some  $b_i \geq \mathbb{E} \left[ \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} \right]$ . Then, since  $(\mu_{ji}^h - \mu_{ji}^l) > 0$  by assumption and  $b_i \geq \mathbb{E} \left[ \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} \right]$ , it must be that  $\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \mathbb{E}[\theta_i | \sigma_i] > 0$  or  $\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - p > 0$ , which means  $\left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right) > 0$ . But we know that  $\lambda^h - \lambda^l - 1 < 0$  and  $\lambda^h - \lambda^l + 1 > 0$  from  $\lambda^h + \lambda^l > 1$ , which implies that  $\left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right) < 0$ . We can analogously prove (ii).  $\square$

Now we can consider which conditions need to be in place for an equilibrium to exist in which  $i$  tells the truth with probabilities  $\lambda^h$  and  $\lambda^l$ . To be clear: We are still considering pure equilibria, since  $i$  has a strict preference for lying or telling the truth for any  $b_i$  except for non-generic boundary cases. However, given  $F_i$  (the distribution of  $b_i$ ), we can determine how often  $i$ 's messages will be truthful once we have established for which  $b_i$   $i$

wants to tell the truth and for which he wants to lie. We can think of  $\lambda^h$  and  $\lambda^l$  as the marginal probabilities of truth-telling by  $i$ .

**Lemma 2.** *There exists an equilibrium in which  $i$  truthfully reveals  $\sigma^h$  with marginal probability  $\lambda^h$  and truthfully reveals  $\sigma^l$  with marginal probability  $\lambda^l$  if and only if*

$$1 - F_i \left( \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + \left( p - \frac{1}{2} \right) \cdot \left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right) \right) \leq \lambda^h$$

and

$$F_i \left( \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + \left( p - \frac{1}{2} \right) \cdot \left( \frac{\lambda^h - 1}{\lambda^h - \lambda^l - 1} + \frac{\lambda^h}{\lambda^h - \lambda^l + 1} \right) \right) \geq \lambda^l.$$

Both inequalities hold with equality if  $F_i$  is continuous at the argument.

*Proof.* From equation 9 we get that  $\Delta U_i(\sigma_i) \geq 0 \Leftrightarrow$

$$b_i - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \geq \frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \mathbb{E}[\theta_i | \sigma_i].$$

Recall that  $\mathbb{E}[\theta_i | \sigma_i = \sigma^h] = p$  and  $\mathbb{E}[\theta_i | \sigma_i = \sigma^l] = 1 - p$ . We can make use of the expression for  $\mu_{ji}^h + \mu_{ji}^l$  that we have derived in (11) to get  $\Delta U_i(\sigma^h) \geq 0 \Leftrightarrow$

$$b_i - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \geq \left( p - \frac{1}{2} \right) \cdot \left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right)$$

and  $\Delta U_i(\sigma^l) \leq 0 \Leftrightarrow$

$$b_i - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \leq \left( p - \frac{1}{2} \right) \cdot \left( \frac{\lambda^h - 1}{\lambda^h - \lambda^l - 1} + \frac{\lambda^h}{\lambda^h - \lambda^l + 1} \right).$$

In an equilibrium, the beliefs of the receivers of  $m_i$  must be correct on average. In this case, this means that it must be sufficiently likely for  $b_i$  to fulfill either of the two inequalities, which gives us the conditions from the proposition. If  $F_i$  is continuous at the argument, correct beliefs require that the inequalities hold with equality. If it is not, there could potentially be mixed equilibria in which for the borderline type,  $i$  mixes between different messages and beliefs are correct on average.  $\square$

Note that that  $\left( \frac{\lambda^h - 1}{\lambda^h - \lambda^l - 1} + \frac{\lambda^h}{\lambda^h - \lambda^l + 1} \right) - \left( \frac{\lambda^l}{\lambda^h - \lambda^l - 1} + \frac{\lambda^l - 1}{\lambda^h - \lambda^l + 1} \right) = 2$ . Lemma 2 consequently describes conditions on the distribution function  $F$  at two points that are  $2p - 1$  apart. In particular if  $F_i$  is continuous at these two points the conditions state that probability mass in the interval between these two points has to equal  $\lambda^l + \lambda^h - 1$ . More importantly, the conditions can be used to show that player  $i$  babbles in a given room if  $F_i$  does not have enough probability mass around the average bias of the other players in the room. To be precise, if  $F_i$  has no probability mass in  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \pm (2p - 1)$ , then the conditions

of lemma 1 imply  $\lambda^l + \lambda^h = 1$  and therefore uninformative messages.<sup>6</sup>

### 5.2.2. Proofs

#### Proof of proposition 16 on page 18.

Without loss of generality, let  $b_1$  and  $b_n$  be the smallest and largest biases respectively. We can represent each bias as the expected value of a distribution that only places density on the values  $b_1 - (2p - 1)$  and  $b_n + (2p - 1)$ . For this set of distributions  $\{F_1, F_2, \dots, F_n\}$ , the conditions of lemma 2 imply  $\lambda^h + \lambda^l = 1$ , and hence there exists no equilibrium in which any of the players tells the truth.  $\square$

#### Proof of proposition 17 on page 18.

We can construct a distribution  $F_i$  that has positive density on  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1}$ , which means that the conditions of lemma 2 imply that there exists an equilibrium in which a message by  $i$  is informative.

To achieve full truth-telling (i.e.  $\lambda^h = \lambda^l = 1$ ), lemma 2 implies we would have to be able to construct an  $F_i$  that only has density inside the interval  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \pm (p - \frac{1}{2})$ . However, this would contradict our starting assumption that if  $b_i$  is  $b_i^e$  for sure, there exists no equilibrium in which  $i$  tells the truth.  $\square$

#### Proof of proposition 18 on page 18.

By the symmetry of  $F$ , all  $F^\kappa$  have the same expected value. We can find a  $\bar{\kappa}$  small enough so that  $F^\kappa$  has less than  $\varepsilon' > 0$  probability mass within  $\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} \pm (2p - 1)$  for any  $\kappa \leq \bar{\kappa}$ . Then it follows from lemma 2 that there exists no equilibrium for which  $\lambda^l + \lambda^h > 1 + \varepsilon'$ . The result follows now from the continuity of (11) and the fact that  $\mu_{ji}^h - \mu_{ji}^l = 0$  if  $\lambda^h + \lambda^l = 1$ .  $\square$

#### Proof of proposition 19 on page 19.

Let the lower (upper) bound of the support be  $\underline{b}_i$  ( $\bar{b}_i$ ). Note that by assumption  $\underline{b}_i \leq \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} - (2p - 1)$  and  $\bar{b}_i \geq \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + (2p - 1)$  which implies by lemma 1 that player  $i$  sends uninformative messages in equilibrium. Now fix  $\underline{b}^\varepsilon = \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} - (2p - 1)$  and  $\bar{b}^\varepsilon = \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + (2p - 1)$ . This implies that  $\mu_{ji}^h - \mu_{ji}^l \leq \varepsilon$  whenever the probability that  $b_i \geq \bar{b}^\varepsilon$  plus the probability that  $b_i < \underline{b}^\varepsilon$  is more than  $1 - \varepsilon'$  for some  $\varepsilon' > 0$  (by lemma 1

<sup>6</sup>To be precise, both points at which  $F_i$  is evaluated in lemma 1 lie in the interior of the interval  $[\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} - (2p - 1), \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + (2p - 1)]$  and therefore  $F_i$  will be continuous at both points and equal to the same value if there is no probability mass in this interval. As the conditions in lemma 1 then hold with equality, they imply  $\lambda^h + \lambda^l = 1$  which in turn implies  $\mu_{ji}^h - \mu_{ji}^l = 0$ .



and the continuity of  $\mu_{ji}$  in  $\lambda^h$  and  $\lambda^l$ ). Let  $\overline{\sigma_{F_i}^2}$  be defined by

$$\overline{\sigma_{F_i}^2} = (1-\varepsilon') \left( \frac{\bar{b}_i - b_i^e}{\bar{b}_i - \underline{b}_i} (b_i - b_i^e)^2 + \frac{b_i^e - \underline{b}_i}{\bar{b}_i - \underline{b}_i} (\bar{b}_i - b_i^e)^2 \right) + \varepsilon' \left( \frac{\bar{b}^\varepsilon - b_i^e}{\bar{b}^\varepsilon - \underline{b}^\varepsilon} (b^\varepsilon - b_i^e)^2 + \frac{b_i^e - \underline{b}^\varepsilon}{\bar{b}^\varepsilon - \underline{b}^\varepsilon} (\bar{b}^\varepsilon - b_i^e)^2 \right).$$

Any distribution with variance above  $\overline{\sigma_{F_i}}$  has to have more than  $\varepsilon'$  probability mass above  $\bar{b}^\varepsilon$  or below  $\underline{b}^\varepsilon$  as  $\overline{\sigma_{F_i}}$  is the variance of the distribution maximizing variance under the constraint that only  $1 - \varepsilon'$  probability mass is outside the interval  $[\underline{b}^\varepsilon, \bar{b}^\varepsilon]$ . Consequently, any distribution with variance above  $\overline{\sigma_{F_i}}$  will lead to  $\mu_{ji}^h - \mu_{ji}^l \leq \varepsilon$ .  $\square$

### Proof of proposition 20 on page 19.

Fix 0 and a  $b > 0$ . Consider the distributions putting probability  $1/2$  on  $-(p - 1/2)$  and  $1/2$  on  $p - 1/2$  instead of 0 for sure and  $1/2$  on  $b - (p - 1/2)$  and  $1/2$  on  $b + (p - 1/2)$ . Under segregation everyone is (just!) truthtelling. In any room including at least 1 player with another bias than the own one, a bias 0 ( $b$ ) player will however lie if his bias is the lower (higher) element of the support:

Take for example a player with bias  $b + p - 1/2$  that got a low signal. Then  $\Delta U(\sigma^l) > 0$  can be written as  $b + p - 1/2 - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} > (\mu_{ji}^h + \mu_{ji}^l)/2 - (1 - p)$ . The right hand side of this inequality is bounded from above by  $p - 1/2$  because  $\mu_{ji}^h \leq p$  and  $\mu_{ji}^l = 1 - p$  by lemma 1 according to which  $\lambda^h = 1$ . As  $b - \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} > 0$ , the claim follows.

To compute welfare under a non-segregated scenario, we need to compute  $\mathbb{E}[(\mu_{ij} - \theta_j)^2]$ . Take, for example, a player  $j$  with biases in  $\{b - p + 1/2, b + p - 1/2\}$ . We showed that this player always sends the high signal if  $b_i = b + p - 1/2$  if at least one player of the other group is in his room. The most informative messaging strategy of such a player in such a room is therefore truthtelling when  $b_i = b - p + 1/2$  and sending the high message otherwise. This implies  $\lambda^h = 1$  and  $\lambda^l = 1/2$  and therefore  $\mu_{ij}^h = (1 + p)/3$  and  $\mu_{ij}^l = 1 - p$ . In this case,

$$\begin{aligned} \mathbb{E}[(\mu_{ij} - \theta_j)^2] &= \frac{1}{2} \left[ \frac{1}{2} \left\{ p \left( \frac{1+p}{3} - 1 \right)^2 + (1-p)(-p)^2 \right\} + \frac{1}{2} \left\{ p(1-p)^2 + (1-p) \left( \frac{1+p}{3} \right)^2 \right\} \right] \\ &\quad + \frac{1}{2} \left[ \frac{1}{2} \left( \frac{1+p}{3} - 1 \right)^2 + \frac{1}{2} \left( \frac{1+p}{3} \right)^2 \right] \\ &= \frac{1}{4} \left[ (1+p) \frac{p^2 - 4p + 4}{9} + (1-p)p^2 + p(1-p)^2 + (2-p) \frac{1 + 2p + p^2}{9} \right] \\ &= \frac{1}{4} \left[ \frac{2}{3} + \frac{4}{3}p - \frac{4}{3}p^2 \right]. \end{aligned}$$

Following the derivations of player  $i$ 's utility in a room that contains players of both groups, see the proof of proposition 1, we can write player  $i$ 's utility if all players are in the same fully integrated room – and choose the best possible messaging strategy

corresponding to  $\lambda^h = 1$  ( $\lambda^h = 1/2$ ) and  $\lambda^l = 1/2$  ( $\lambda^l = 1$ ) for players with expected bias  $b_i^e = b$  ( $b_i^e = b$ ) – as

$$U_i^{int} = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - [n + \alpha(n-1)n]/4 + (1/4 - p(1-p))(1 + \alpha(n-1)) \\ + (1/4 - [2/3 + p4/3 - p^24/3]/4) [n-1 + \alpha \sum_{j \neq i} \{n-1\}]$$

while his expected payoff under full segregation is

$$U_i^{seg} = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - [n + \alpha(n-1)n]/4 + (1/4 - p(1-p))(n/2 + \alpha(n-1)n/2).$$

$U_i^{seg}$  exceeds  $U_i^{int}$  if and only if

$$(1/4 - p(1-p))(1 + \alpha(n-1))(n/2 - 1) \geq (1/4 - [2/3 + p4/3 - p^24/3]/4) [n-1 + \alpha(n-1)^2] \\ \Leftrightarrow (1 - 4p + 4p^2)(1 + \alpha(n-1))(n/2 - 1) \geq (1/3 - p4/3 + p^24/3) [n-1 + \alpha(n-1)^2] \\ \Leftrightarrow 3(1 + \alpha(n-1))(n/2 - 1) \geq n-1 + \alpha(n-1)^2 \\ \Leftrightarrow \frac{3}{2}(1 + \alpha(n-1)) \frac{n-2}{n-1} \geq 1 + \alpha(n-1) \\ \Leftrightarrow \frac{n-2}{n-1} \geq \frac{2}{3}$$

which is true for  $n \geq 4$ . As the payoffs do not differ across players in each of the two scenarios, welfare is higher under segregation than under integration given that  $n \geq 4$ .

To see that other room configurations cannot improve welfare, start from full segregation. Moving  $k$  players from room 1 to room 2 will lead to less information for the remaining players in room 1. Suppose nevertheless that this move was welfare increasing. Then players in the new room 2 must have better information than under segregation. Note that by assumption the most informative strategy players could possibly adopt in the new room is  $\lambda^h = 1$  ( $\lambda^h = 1/2$ ) and  $\lambda^l = 1/2$  ( $\lambda^l = 1$ ) for players with expected bias  $b_i^e = b$  ( $b_i^e = b$ ). Assume that this strategy is an equilibrium in the new room 2 (if it is not, this step increases the welfare gain over segregation). But then it is clearly optimal to move the remaining players from room 1 to room 2 as well (if this strategy remains an equilibrium): This improves information for all players. But this would imply  $U_i^{int} > U_i^{seg}$  which contradicts what we showed above.  $\square$

## 6. Public information

Here we add a public information component. Our main interest is in the comparative statics of the weight of this public information component, i.e. if more information becomes publicly available, how will communication be affected?

We consider in this extension a state of the world  $\theta = \tau\theta_0 + (1 - \tau)\sum_{i=1}^n \theta_i$ ; that is, we add – compared to the main model of the paper – an element  $\theta_0 \in \{0, 1\}$  which receives a weight  $\tau$ . As for all other  $\theta_i$ , there is also binary a signal  $\sigma_0$ . This signal has accuracy  $p_0 > 1/2$ , i.e.  $Pr(\sigma_0 = \sigma^h | \theta_0 = 1) = Pr(\sigma_0 = \sigma^l | \theta_0 = 0)$ , and is observed by all players. Consequently, all players share the same belief about  $\theta_0$  which is denoted by  $\mu_0$ . Everything else is as in the main model of the paper.<sup>7</sup>

The optimal choice of action is now

$$a_i^* = b_i + \mathbb{E}[\theta] = b_i + \tau\mu_0 + (1 - \tau)\sum_{j=1}^n \mu_{ij}.$$

Note that the proof of lemma 1 goes through with straightforward adaptations. In particular,

$$U_i(m_i) = \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in R_i, j \neq i} \left( a_j(m_i, m_{-i, R_i}, \sigma_j) - b_i - \tau\theta_0 - (1 - \tau)\sum_{k=1}^n \theta_k \right)^2 \middle| \sigma_i \right].$$

which leads to

$$\begin{aligned} \Delta U_i(\sigma_i) &= (U_i(m^h) - U_i(m^l))/\alpha \\ &= - \sum_{j \in R_i, j \neq i} \mathbb{E} \left[ (1 - \tau)^2 \mu_{ji}^{h^2} - (1 - \tau)^2 \mu_{ji}^{l^2} \right. \\ &\quad \left. + 2(1 - \tau)(\mu_{ji}^h - \mu_{ji}^l) \left( b_j - b_i + \tau(\mu_0 - \theta_0) + (1 - \tau)\sum_{k \neq i} (\mu_{jk} - \theta_k) - (1 - \tau)\theta_i \right) \middle| \sigma_i \right] \\ &= -2(1 - \tau)(\mu_{ji}^h - \mu_{ji}^l) \sum_{j \in R_i, j \neq i} \left[ (1 - \tau)\frac{\mu_{ji}^h + \mu_{ji}^l}{2} + b_j - b_i - (1 - \tau)\mathbb{E}[\theta_i | \sigma_i] \right] \\ &= 2(1 - \tau)(\mu_{ji}^h - \mu_{ji}^l)(n_{R_i} - 1) \left[ -(1 - \tau)\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} + b_i + (1 - \tau)\mathbb{E}[\theta_i | \sigma_i] \right]. \end{aligned}$$

Using this expression instead of (16) in the paper the proof of lemma 1 applies and we can concentrate on pure strategy equilibria.

A result similar to theorem 1 in the paper now follows immediately from the expression above:

---

<sup>7</sup>One way to interpret the weights is the following: Suppose there is a continuum of  $\tilde{\theta}$  of unit length, say  $[0, 1]$ . Each  $\tilde{\theta} \in [0, 1]$  is either 0 or 1 and agents try to match the average  $\tilde{\theta}$  (plus their bias) with their action. For a set  $\Theta_0 \subset [0, 1]$  of measure  $\tau$ , a public signal is available (there could be several public signals which are then aggregated; the only thing that matters is that everyone has the same expectation about the value of the average  $\tilde{\theta}$  in  $\Theta_0$ ). Each player  $i$  has a private signal about the average value of the  $\tilde{\theta} \in \Theta_i$  where  $\Theta_i$  has measure  $(1 - \tau)/n$  and we assume that all  $\Theta_i$  are pairwise disjoint. The comparative static with respect to  $\tau$  answers then the question: What happens if information/signals that used to be privately held by some expert are now publicly available?

**Theorem 3.** Let  $\bar{b} = \frac{\sum_{k \in R} b_k}{n_R}$  be the mean bias of players in room  $R$ . In the most informative equilibrium in this room, a player  $i$  tells the truth if and only if

$$b_i \in \left[ \bar{b} - \frac{n_R - 1}{n_R} \left( p - \frac{1}{2} \right) (1 - \tau), \bar{b} + \frac{n_R - 1}{n_R} \left( p - \frac{1}{2} \right) (1 - \tau) \right]$$

and babbles otherwise.

**Proof.** Consider the difference between lying and truth-telling for player  $i$ , i.e.  $\Delta U_i$  as derived above. For every non-babbling player  $\mu_{ji}^h = p$  and  $\mu_{ji}^l = 1 - p$  (as we can concentrate on pure strategy equilibria) which implies that the necessary equilibrium condition  $\Delta U_i(\sigma^h) \geq 0$  simplifies to

$$\begin{aligned} b_i - \frac{1}{n_R - 1} \sum_{j \in R_i, j \neq i} b_j &\geq \left( \frac{1}{2} - p \right) (1 - \tau) \\ \frac{n_R}{n_R - 1} b_i - \frac{1}{n_R - 1} \sum_{k \in R_i} b_k &\geq \left( \frac{1}{2} - p \right) (1 - \tau) \\ b_i &\geq \bar{b} - \frac{n_R - 1}{n_R} \left( p - \frac{1}{2} \right) (1 - \tau). \end{aligned}$$

If this inequality does not hold, player  $i$  will not use the truthful strategy in the most informative equilibrium and therefore he will babble in the most informative equilibrium.

We can analogously solve for  $\Delta U_i(\sigma^l) \leq 0$  and get the interval in the theorem.  $\square$

Theorem 3 implies that a higher weight on public information reduces the length of the truthtelling interval. That is, public information crowds out communicated private information in a given room. This implies that also under the welfare optimal room allocation less private information will be communicated for higher  $\tau$ .<sup>8</sup>

**Proposition 21.** Let  $\tau^h > \tau^l$ . In the welfare optimal room allocation the total amount of communicated information is (weakly) less under  $\tau^h$  than under  $\tau^l$

**Proof.** Take the welfare optimal room allocation under  $\tau^h$ . Using the same room allocation under  $\tau^l$  will create at least as much information as under  $\tau^h$  by theorem 3. (Adapting the room allocation may increase the number of communicated pieces of information further.)  $\square$

---

<sup>8</sup>Note that welfare – for a given  $\tau$  – is, as in the paper, proportional to the number of communicated pieces of information. The derivation goes through with the obvious adaptations leading to

$$\begin{aligned} W &= -\alpha \sum_{i=1}^n \sum_{j \neq i} \{ (b_j - b_i)^2 \} - \frac{1}{4} (1 - \tau)^2 n^2 [1 + \alpha(n - 1)] \\ &\quad + (1 - \tau)^2 \left( p - \frac{1}{2} \right)^2 (1 + \alpha(n - 1)) \sum_i \zeta_i - n(1 + \alpha(n - 1)) p_0 (1 - p_0) \tau^2. \end{aligned}$$

Note, however, that welfare is not necessarily decreasing in  $\tau$ . The positive effect of more public information counteracts the negative effect of less private information. The overall effect is generally ambiguous.<sup>9</sup>

To think about segregation, let us focus on the binary bias case, i.e.  $b_i \in \{0, b\}$  for  $i = 1, \dots, n$ . Recall that in this setting the welfare optimal room allocation has to fall into one of the following four categories, see section 1:

1. full integration (either with everyone truthtelling or only the majority)
2. full segregation
3. a mixed room in which only majority players are truthtelling and one room with some minority players
4. a mixed room in which everyone is truthtelling and an extra room with some majority players.

The detrimental effect of higher  $\tau$  on communication, now immediately implies that an increase in  $\tau$  leads to more segregation in one of the following ways: First, full segregation might become optimal for higher  $\tau$  (as truthful communication in a mixed or fully integrated room is possible to a lesser degree). Second, less minority players can remain in a mixed bias room in which only majority players are truthtelling (as otherwise majority players babble). Third, less majority players can remain in a mixed room in which all players are truthtelling (as otherwise minority players lose their incentive to be truthful).

The main intuition behind these results is that more public information implies less influence of  $i$ 's message on  $j$ 's decision because  $i$  holds less relevant information privately. With less influence lying is less costly as  $j$  will “overshoot” less (when communicating a high message instead of a low message). This intuition also suggests that for sufficiently high  $\tau$  communication between players of different biases is impossible and therefore segregation by bias is welfare optimal and an equilibrium. The following result states this formally for generic configurations of biases  $\mathcal{B}$  (not necessarily binary). Assume that  $\mathcal{B}$  is generic in the sense that player  $i$ 's bias is not the average of other players' biases (whose biases are distinct from  $b_i$ ).<sup>10</sup>

**Theorem 4.** *For generic  $\mathcal{B}$ , there exists a  $\bar{\tau} < 1$ , such that full segregation based on biases is both welfare optimal and an equilibrium if  $\tau \geq \bar{\tau}$ .*

---

<sup>9</sup>To construct an example where welfare is locally decreasing in  $\tau$  it is enough to choose parameter values such that truthtelling in a fully integrated room is just possible, i.e. some player is indifferent between truthtelling and not. Marginally increasing  $\tau$  in this situation will discretely lower the  $\sum_i \zeta_i$  in the welfare function while affecting all other terms continuously. Hence, a slightly higher  $\tau$  will decrease welfare.

<sup>10</sup>More precisely, the assumption is that  $b_i \neq \sum_{b_j \in \mathcal{B} \setminus \{b_i\}} \frac{\tilde{n}_{b_j}}{\sum_k \tilde{n}_{b_k}} * b_j$  for any  $\tilde{n}_{b_j} \in \{0, 1, \dots, n_{b_j}\}$  where  $n_{b_j}$  is the number of players with bias  $b_j$ .

**Proof.** Theorem 3 together with our genericity assumption implies that for a given room in which at least two players differ in their bias there is a  $\bar{\tau}_R < 1$  such that babbling is the unique equilibrium of the messaging game in this room if  $\tau \geq \bar{\tau}_R$ . By the finiteness of the set of players, the number of possible room configurations is finite and therefore  $\max_R \bar{\tau}_R$  exists and is strictly less than 1. Take  $\bar{\tau} = \max_R \bar{\tau}_R$ . Then babbling is the unique equilibrium of all non-segregated rooms and it is clear that segregation maximizes the number of communicated pieces of information and therefore welfare. Also no player wants to deviate from a segregated room to another room as this would lead to babbling by all players in the room he deviates to (and also deprives the players of the segregated room of his truthful message).  $\square$

## 7. Signaling

### 7.1. Binary signal

This section explores a setting where players in a given room have an additional option: they can not only send a cheap talk message but also send a costly signal. The costly signal could be to search for a link to some document supporting the stated opinion or to spend some time to carefully state the argument one may want to make. The message space is therefore  $\{m^l, m^h, \bar{m}^l, \bar{m}^h\}$  where  $m^l$  and  $m^h$  are costless cheap talk messages as before and the messages  $\bar{m}^h$  and  $\bar{m}^l$  are costly messages. Sending such a costly message deducts  $c > 0$  from the sending player's payoff.

For now, take the room allocation as given and consider the choice of messages. Theorem 1 implies that players with  $b_i \in [\bar{b} - (p - 1/2)(n_R - 1)/n_r, \bar{b} + (p - 1/2)(n_R - 1)/n_R]$  can communicate truthfully through costless cheap talk. Hence, the possibility of sending costly messages is irrelevant for them in the most informative equilibrium. We will try to construct an equilibrium in which some players with  $b_i$  outside this interval are able to communicate their information through a costly message. For concreteness, let  $b_i > \bar{b} + (p - 1/2)(n_R - 1)/n_R$  in the following. Now consider the following strategy of  $i$  and beliefs of  $-i$ . Player  $i$  sends message  $\bar{m}^h$  if  $\sigma_i = \sigma^h$  and message  $m^l$  otherwise. The beliefs of player  $j \in R$ ,  $j \neq i$  are  $\mu_{ji}^h = \mu_{ji}^l = \mu_{ji}^{\bar{l}} = 1 - p$  and  $\mu_{ji}^{\bar{h}} = p$ . In words, player  $j$  believes that  $i$  received a low signal unless  $i$  sends the costly high message  $\bar{m}^h$ . These beliefs are consistent with Bayes' rule given  $i$ 's strategy and it remains to check whether  $i$ 's strategy is optimal given these beliefs. First, consider  $\sigma_i = \sigma^h$ . Equation 16 in the proof of lemma 1 implies that  $i$  prefers sending the costly message  $\bar{m}^h$  to sending a cheap

talk message if and only if

$$\begin{aligned}
& 2\alpha(2p-1)(n_R-1) \left[ -\frac{1}{2} - \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1} + b_i + p \right] \geq c \\
\Leftrightarrow b_i & \geq \frac{c}{2\alpha(2p-1)(n_R-1)} - \left( p - \frac{1}{2} \right) + \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1} \\
& \Leftrightarrow b_i \geq \bar{b} - \frac{n_R-1}{n_R} \left( p - \frac{1}{2} \right) + \frac{c}{2\alpha(2p-1)n_R}.
\end{aligned}$$

Hence,  $i$ 's strategy is only optimal if  $b_i$  is sufficiently high (relative to  $c$ ). The reason is that players with higher  $b_i$  suffer, due to the concavity of the utility function, more from other players taking a too low action. Hence, they are willing to pay more, i.e. tolerate a higher  $c$ , for increasing the beliefs and therefore the actions of the other players in the room. Note that  $b_i > \bar{b} + (p-1/2)(n_R-1)/n_R$  implies the condition above if  $c \leq 2\alpha(2p-1)^2(n_R-1)$ .

Second, consider  $\sigma_i = \sigma^l$ . Equation 16 in the proof of lemma 1 implies that  $i$  prefers sending a cheap talk message to sending the costly message  $\bar{m}^h$  if and only if

$$\begin{aligned}
& 2\alpha(2p-1)(n_R-1) \left[ \frac{1}{2} - p + b_i - \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1} \right] \leq c \\
\Leftrightarrow b_i & \leq \frac{c}{2\alpha(2p-1)(n_R-1)} + \left( p - \frac{1}{2} \right) + \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1} \\
& \Leftrightarrow b_i \leq \bar{b} + \frac{n_R-1}{n_R} \left( p - \frac{1}{2} \right) + \frac{c}{2\alpha(2p-1)n_R}.
\end{aligned}$$

This condition is satisfied if  $b_i$  is not too high. The reason why players with very high  $b_i$  are unable to signal with costly messages is that they would happily pay the cost  $c$  in order to induce a high belief even if their signal is low. If  $b_i$  is lower, however, players are only willing to do so if their signal is high where a low action (induced by low beliefs resulting from cheap talk) would hurt the player more compared to the situation where his signal is low.

This implies that the above stated strategy and beliefs constitutes an equilibrium of the messaging game if  $b_i \in [\bar{b} - (p-1/2)(n_R-1)/n_R + c/(2\alpha n_R(2p-1), \bar{b} + (p-1/2)(n_R-1)/n_R + c/(2\alpha n_R(2p-1))]$ .

An analogous argument yields that it is an equilibrium for  $b_i \in [\bar{b} - (p-1/2)(n_R-1)/n_R - c/(2\alpha n_R(2p-1), \bar{b} + (p-1/2)(n_R-1)/n_R - c/(2\alpha n_R(2p-1))]$  to use the strategy of sending message  $\bar{m}^l$  if  $\sigma_i = \sigma^l$  and message  $m^h$  if  $\sigma_i = \sigma^h$  together with beliefs  $\mu_{ji}^h = \mu_{ji}^l = \mu_{ji}^h = p$  and  $\mu_{ji}^l = 1 - p$ .

In conclusion, we have to distinguish two cases. First,  $c \leq 2\alpha(2p-1)^2(n_R-1)$ . Then, there is an equilibrium in which players with bias very close to  $\bar{b}$  will engage in truthful cheap talk, players with bias moderately close to  $\bar{b}$  will truthfully communicate using

costly messages for signals in the direction of  $b_i - \bar{b}$ , and finally players with  $b_i$  far away from  $\bar{b}$  will babble.

Second,  $c > 2\alpha(2p - 1)^2(n_R - 1)$ . In this case, there is an equilibrium in which players with bias very close to  $\bar{b}$  will engage in truthful cheap talk, players with  $b_i$  somewhat further away from  $\bar{b}$  or very far away from  $b_i$  will babble, but there are some biases in between at which players engage in meaningful communication via costly signaling.

Above we specified certain beliefs and strategies and checked when these strategies and beliefs constitute an equilibrium of the messaging game. One might wonder whether other equilibria with signaling are possible. This is indeed the case, however, none of these equilibria generates more information than the one we constructed (while the usual babbling logic implies that there are many less informative equilibria). To see this note a few peculiarities of the equilibrium above: First, we chose the most extreme beliefs,  $p$  and  $1 - p$ , possible in order to maximize the incentives to engage in costly signaling. Second, we let players buy the signal only if  $\sigma_i$  is in line with their bias relative to  $\bar{b}$ . Put differently, a player with  $b_i > \bar{b}$  will buy the costly signal only if  $\sigma_i = \sigma^h$ . Due to the concavity of the utility function this is the  $\sigma_i$  value for which  $i$  has the highest willingness to pay for increasing the other players' beliefs. Any other equilibrium will give lower incentives to engage in signaling and will therefore reduce the range of  $b_i$  for which costly signaling is optimal.

In terms of welfare inducing a babbling player to signal information increases the number of pieces of information by  $n_{R_i} - 1$ . Consequently, welfare increases by  $(p - 1/2)^2(1 + \alpha(n - 1))(n_{R_i} - 1)$ . The costs are  $c$ . Welfare increases therefore by signaling if  $c < (p - 1/2)^2(1 + \alpha(n - 1))(n_{R_i} - 1) = (2p - 1)^2(1 + \alpha(n - 1))(n_{R_i} - 1)/4$  (see the expression for welfare in the proof of proposition 1). This implies that the most informative equilibrium is no longer necessarily the welfare maximal equilibrium because information through signaling comes at a cost and a player's trade off between informing other players and incurring the cost  $c$  differs from the trade off a welfare maximizing planner faces. Note that welfare maximizing and most informative equilibrium coincide if  $c$  is either very low or very high and only differ in an intermediate range. As a consequence, theorem 2 still holds in this setup (the proof goes through with minor adaptations): in sufficiently polarized societies full segregation is welfare optimal and in sufficiently homogenous societies full integration is optimal.

## 7.2. Continuous signal

This subsection adapts the previous one by allowing for a continuous signal. That is, the sender can choose how much effort he wants to put into drafting the message. We equate this effort with its costs  $c$  which are observable by the receiver. The sender can choose any level of  $c$  in some interval  $[0, \bar{c}]$ . Following the signaling literature, we focus on the least cost separating equilibrium. "Separating" refers in our setup to informative



communication. In our empirical application effort can be interpreted as formulating the message well or searching for evidence in the form of links.

The message space in this variation consists of a family of two messages  $\{m^l(c), m^h(c)\}$  indexed by the effort cost. Cheap talk messages occur for  $c = 0$ . Theorem 1 implies that players with  $b_i \in [\bar{b} - (p - 1/2)(n_R - 1)/n_R, \bar{b} + (p - 1/2)(n_R - 1)/n_R]$  can communicate truthfully through costless cheap talk and therefore the least cost separating strategy for them is truthful cheap talk.

For players with  $b_i \in (\bar{b} + (p - 1/2)(n_R - 1)/n_R, \bar{b} + (p - 1/2)(n_R - 1)/n_R + \bar{c}/(2\alpha n_R(2p - 1))]$ , there exists a  $c_{b_i} \in (0, \bar{c}]$  such that

$$b_i = \bar{b} + (p - 1/2) \frac{n_R - 1}{n_R} + \frac{c_{b_i}}{2\alpha n_R(2p - 1)}$$

$$\Leftrightarrow c_{b_i} = 2\alpha(2p - 1)(n_R - 1) \left[ \frac{1}{2} - p + b_i - \frac{\sum_{j \in R, j \neq i} b_j}{n_R - 1} \right].$$

By the arguments in the previous subsection,  $c_{b_i}$  is the lowest cost level at which player  $i$  can credibly signal. Analogously, for players with  $b_i \in (\bar{b} - (p - 1/2)(n_R - 1)/n_R, \bar{b} - (p - 1/2)(n_R - 1)/n_R - \bar{c}/(2\alpha n_R(2p - 1))]$ ,  $c_{b_i} \in (0, \bar{c}]$  equals

$$b_i = \bar{b} - (p - 1/2) \frac{n_R - 1}{n_R} - \frac{c_{b_i}}{2\alpha n_R(2p - 1)}$$

$$\Leftrightarrow c_{b_i} = 2\alpha(2p - 1)(n_R - 1) \left[ \frac{1}{2} - p - b_i + \frac{\sum_{j \in R, j \neq i} b_j}{n_R - 1} \right].$$

Players with  $|b_i - \bar{b}| > (p - 1/2)(n_R - 1)/n_R + \bar{c}/(2\alpha n_R(2p - 1))$  no credible signaling is possible (see the previous subsection).

In conclusion, the model implies that players with  $b_i$  close to the average will communicate by cheap talk, players with  $b_i$  intermediately away from  $\bar{b}$  will signal and extremists will babble. The signaling effort is increasing in  $|b_i - \bar{b}|$  up to some point and drops to zero afterwards.

## 8. Verification

This sections extends the model by allowing players to communicate their signal verifiably at a cost  $c > 0$ . In other words, players have the choice to either send a costless cheap talk message or to verifiably communicate their true signal at cost  $c$  to all players in their room. Players in other rooms receive neither cheap talk nor verifiable messages.

Let the room allocation be given. Theorem 1 implies that players with  $b_i \in [\bar{b} - (p - 1/2)(n_R - 1)/n_R, \bar{b} + (p - 1/2)(n_R - 1)/n_R]$  can communicate truthfully through cheap talk and therefore verification is unnecessary for them. We will therefore concentrate on players with bias  $b_i$  outside this interval. Let, for concreteness,  $b_i > \bar{b} + (p - 1/2)(n_R - 1)/n_R$ .

We now try to construct an equilibrium in which  $i$  uses verification in order to credibly transmit information. By  $b_i > \bar{b} + (p - 1/2)(n_r - 1)/n_R$ , player  $i$  has greater incentives to verify a high signal than a low signal. To maximize informativeness, it is therefore optimal to choose the belief of the other players in case  $i$  sends a cheap talk message to be  $\mu_{ij}^{nv} = 1 - p$ . That is, player  $j$  believes with probability 1 that  $i$  received the low signal whenever  $i$  sends a cheap talk message. The fully informative equilibrium that we try to establish is then that  $i$  verifies his signal whenever it is high and sends a cheap talk message when his signal is low. Given the belief  $\mu_{ji}$ , this is informationally equivalent to truthful communication. It is obvious that, given  $\mu_{ji}$ ,  $i$  will not verify a low signal but it needs to be checked whether verifying a high signal is optimal. Using equation 16 in the proof of lemma 1 and given  $\mu_{ji}$ , the utility of verifying minus the utility of sending a cheap talk message given  $\sigma_i = \sigma^h$  is greater than  $c$  if and only if

$$2\alpha(2p - 1)(n_{R_i} - 1) \left[ -\frac{1}{2} - \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} + b_i + p \right] \geq c$$

$$\Leftrightarrow b_i \geq \frac{c}{2\alpha(2p - 1)(n_{R_i} - 1)} - \left( p - \frac{1}{2} \right) + \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1}.$$

This means that  $b_i$  has to be large enough relative to  $c$  to make verification worthwhile. Clearly, verification is not optimal if  $c$  is excessively large. The option of sending a cheap talk message inducing belief  $\mu_{ji} = 1 - p$  is less attractive for players with higher  $b_i$  due to the concavity of the quadratic loss function. Consequently, players with a higher  $b_i$  are willing to tolerate higher costs of verification.

If  $b_i$  is below the threshold above, then there is no equilibrium in which  $i$  verifies his signal. By choosing  $\mu_{ji} = 1 - p$  and having verification only when the signal is high, we maximized the incentives of  $i$  to verify. Hence,  $i$  will never verify if he finds verification suboptimal above. The following result summarizes the derivation above

**Lemma 8.** *The most informative equilibrium in room  $R$  consists of the following strategies:*

- *truthful cheap talk if  $b_i \in [\bar{b} - (p - 1/2)(n_r - 1)/n_r, \bar{b} + (p - 1/2)(n_r - 1)/n_r]$*
- *verifying  $\sigma^h$  and cheap talk in case of  $\sigma^l$  if  $b_i > \bar{b} + (p - 1/2)(n_R - 1)/n_r$  and  $b_i \geq \frac{c}{2\alpha(2p-1)(n_R-1)} - p + \frac{1}{2} + \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1}$*
- *verifying  $\sigma^l$  and cheap talk in case of  $\sigma^h$  if  $b_i < \bar{b} - (p - 1/2)(n_R - 1)/n_r$  and  $b_i \leq -\frac{c}{2\alpha(2p-1)(n_R-1)} + p - \frac{1}{2} - \frac{\sum_{j \in R, j \neq i} b_j}{n_R-1}$*
- *babbling else.*

Note that for  $c$  sufficiently small no player will babble. As we considered only players that could not truthfully communicate with cheap talk, the condition

$$c \leq 2\alpha(2p - 1)^2(n_R - 1)$$

is sufficient for ensuring that no player babbles. If this condition is violated, the structure is that “centrists” (those with  $b_i \in [\bar{b} - (p - 1/2)(n_R - 1)/n_r, \bar{b} + (p - 1/2)(n_R - 1)/n_r]$ ) communicate truthfully through cheap talk, “moderate extremists” babble and “strong extremists” verify.

In terms of welfare inducing a babbling player to verify information increases the number of pieces of information by  $n_{R_i} - 1$ . Consequently, welfare increases by  $(p - 1/2)^2(1 + \alpha(n - 1))(n_{R_i} - 1)$ . The costs are  $c$ . Consequently, welfare increases by verification if  $c < (p - 1/2)^2(1 + \alpha(n - 1))(n_{R_i} - 1) = (2p - 1)^2(1 + \alpha(n - 1))(n_{R_i} - 1)/4$  (see the expression for welfare in the proof of proposition 1). This implies that the most informative equilibrium is no longer necessarily the welfare maximal equilibrium because information through verification comes at a cost.<sup>11</sup>

The implications of verification for room allocation is that bigger rooms can be optimal. Clearly, any room allocation will produce at least as much information with verification as without. Since players that were babbling under pure cheap talk may now communicate information by means of verification, it can make sense to have people with a larger spread of biases in one room.

## 9. States correlated within bias groups

This section considers a variation of the model in which players with a similar bias have similar information. This feature implies that communication across bias groups is even more desirable from a welfare perspective. However, we will show that for the same reasons as in the paper such communication is infeasible in equilibrium if bias differences are large.

We will focus on a model setup with two bias groups, i.e.  $\mathbb{B} = \{0, b\}$ . Without loss of generality let players  $i = 1, 2, \dots, n_0$  have bias  $b_i = 0$  and players  $i = n_0 + 1, \dots, n_0 + n_b$  have bias  $b_i = b$ . We will introduce similarity of information by assuming that  $\theta_i$  and  $\theta_j$  are positively correlated if either  $i, j \in \{1, \dots, n_0\}$  or  $i, j \in \{n_0 + 1, \dots, n_0 + n_b\}$ . However, we maintain the assumptions that (i)  $\theta_i$  and  $\theta_j$  are uncorrelated if  $i \in \{1, \dots, n_0\}$  and  $j \in \{n_0 + 1, \dots, n_0 + n_b\}$ , (ii) signal  $\sigma_i$  is noisy and independent of  $\sigma_j$  and  $\theta_j$  conditional on  $\theta_i$ , (iii) that  $\theta_i \in \{0, 1\}$  and the marginals are such that  $\mathbb{E}[\theta_i] = 1/2$  (this latter assumption is for convenience of notation only). We will not be more specific about the correlation but

<sup>11</sup>In this context, it is interesting to ask when a player can be prevented from verifying his signal. Take again  $b_i > \bar{b} + (p - 1/2)(n_R - 1)/n_r$  for concreteness. Then babbling is preferred to verifying a high signal if

$$2\alpha(p - 1/2)(n_R - 1) \left[ -\frac{p + 1/2}{2} - \frac{\sum_{j \in R, j \neq i} b_j}{n_R - 1} + b_i + p \right] \leq c$$

which is equivalent to

$$b_i \leq \frac{c}{2\alpha(p - 1/2)(n_R - 1)} - \frac{1}{2} \left( p - \frac{1}{2} \right) + \frac{\sum_{j \in R, j \neq i} b_j}{n_R - 1}.$$

want to point out the two extreme cases: First, perfect correlation within bias groups. In this case, all players with the same bias receive effectively information about the same underlying variable. On the other, complete independence which is the case we analyze in the paper.

Given the flexible formulation, it is unsurprising that a closed form solution no longer exists. However, we will be able to show a result similar to the one in the main text in this setup. In a given room including players of both bias groups essentially no information transmission is possible if  $b$  is sufficiently large. Eventually, we conclude this section with some comments on a behavioral phenomenon namely correlation neglect, i.e. we discuss how our results change if players are not taking the correlation of states into account.

The following proposition states that the amount of information transmitted in a given room with players of both biases is less than an arbitrary  $\varepsilon > 0$  if  $b$  is large enough.<sup>12</sup>

**Proposition 22.** *Let  $R$  be a room containing at least one player with bias 0 and at least one player with bias  $b$ . For every  $\varepsilon > 0$ , there exists a  $b_\varepsilon$  such that  $\mathbb{E}_{m_{-i}, \sigma_j}[\mu_j(h) - \mu_j(l) | \sigma_i] < \varepsilon$  for every player  $i \in R$  in every equilibrium of the communication stage.*

**Proof of proposition 22:** Clearly, it is still optimal to choose action  $a_i = b_i + \mathbb{E}[\theta | \sigma_i, m_{R_i}]$  where  $m_{R_i}$  are the messages observed by player  $i$ .

For concreteness take a player  $i$  with bias  $b_i = 0$  and compare the difference in expected utility of this player when sending message  $h$  and message  $l$  (we neglect that the expectation is conditional on  $\sigma_i$  to avoid cluttering of notation):

$$\begin{aligned} \Delta U_i &= \alpha \sum_{j \neq i, j \in R_i} \mathbb{E} [a_j(l)^2 - a_j(h)^2 - 2\theta(a_j(l) - a_j(h))] \\ &= \alpha \sum_{j \neq i, j \in R_i} \mathbb{E} [\mu_j(l)^2 - \mu_j(h)^2 + 2b_j(\mu_j(l) - \mu_j(h)) - 2\theta(\mu_j(l) - \mu_j(h))] \\ &= -2\alpha \sum_{j \neq i, j \in R_i} \mathbb{E} \left[ (\mu_j(h) - \mu_j(l)) \left( \frac{\mu_j(h) + \mu_j(l)}{2} - \theta + b_j \right) \right]. \end{aligned}$$

As  $-n_0 - n_b \leq (\mu_j(h) + \mu_j(l))/2 - \theta \leq n_0 + n_b$ , choosing  $b_\varepsilon = (n_0 + n_b) + (n_0 + n_b)^2/\varepsilon$  is sufficient for  $\Delta U_i < 0$  regardless of  $\sigma_i$ .  $\square$

## 10. Alternative signal technologies

In this section, we consider three variations of the model in the paper. The first is a straightforward extensions in which we allow for more signals than just the binary signal structure considered in the paper. (We can also allow for more states but this is

<sup>12</sup>We denote here the action of player  $j$  when  $i$  sends message  $l$  by  $a_j(l)$ . This action depends also on  $\sigma_j$  and messages of other players but we suppress this dependence in the interest of readability. Similarly  $\mu_j(l)$  is  $j$ 's belief about  $\theta$  if  $i$  sends message  $l$  which again depends also on  $\sigma_j$  and other players' messages.

relatively immaterial in our setting.) The second variation considers goes a bit further by considering a continuum of signals. The third changes the signal structure such that no longer each player receives a signal about “his state”  $\theta_i$  as in the main text but instead all players receive a noisy signal about the same one-dimensional state  $\theta$ . For all variations we show that our main result that integration is optimal and an equilibrium if there is little polarization while segregation is optimal and an equilibrium if there is a lot of polarization continue to hold. The main shortcoming of the first and third variation is that for intermediate values of polarization it is no longer possible to determine the most informative equilibrium of the messaging game as we can no longer rule out that this equilibrium involves mixed strategies. The second variation allows only a closed form solution of the most informative messaging equilibrium for particular distributions, e.g. the uniform distribution. This makes each variation less tractable than the model of the main paper.

### 10.1. Larger signal and state space

Now allow for an arbitrary finite number of states, biases and signals. We keep the assumption that states and signals of different players are independent and that player  $i$  receives a signal that is partially informative about state  $\theta_i$  (but independent about all other states). We also keep the utility function, i.e. the additive structure. The message space equals the signal space and we assume that lower signals lead to a lower expected value of  $\theta_i$ . For notational simplicity let the signal be the posterior it leads to, i.e.  $\sigma_i = \mathbb{E}[\theta_i|\sigma_i]$ .

Following similar steps as in the main text, we can derive the expected utility difference between sending two messages labeled as high ( $h$ ) and low ( $l$ ). Let  $\mu_{ji}^h$  denote the expected value that  $j$  assigns to  $\theta_i$  upon receiving message  $h$  (given some equilibrium messaging strategy by  $i$ ). The expected utility difference can then, similarly to above, be derived as

$$\begin{aligned} \Delta U_i(\sigma_i) &= \sum_{j \in R_i, j \neq i} (\mu_{ji}^l)^2 - (\mu_{ji}^h)^2 + 2(\mu_{ji}^l - \mu_{ji}^h)(b_j - b_i - \mathbb{E}[\theta_i|\sigma_i]) \\ &= \sum_{j \in R_i, j \neq i} (\mu_{ji}^l - \mu_{ji}^h) [(\mu_{ji}^l + \mu_{ji}^h) + 2(b_j - b_i - \sigma_i)] \\ &= -2(n_{R_i} - 1) (\mu_{ji}^h - \mu_{ji}^l) \left[ \frac{\mu_{ji}^l + \mu_{ji}^h}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i - \sigma_i \right] \end{aligned}$$

This expression implies that a truthtelling equilibrium exists if and only if for every player  $i$  and every  $\sigma_i^l < \sigma_i^h$

$$\sigma_i^l \leq \frac{\sigma_i^l + \sigma_i^h}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i \leq \sigma_i^h$$

$$\Leftrightarrow \left| \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i \right| \leq \frac{\sigma^h - \sigma^l}{2}.$$

If we assume that all players have the same signal space, this condition is tightest for the player whose bias  $b_i$  is furthest away from the other players' average bias,  $\sum_{j \in R_i, j \neq i} b_j / (n_{R_i} - 1)$ , and for the two signals that are closest together.

It is immediate from the expression above that (i) truthtelling is impossible if bias differences are too high, (ii) adding moderates can establish truthtelling as it can move the average of the other players closer to each player's bias (e.g. consider a room with 2 people with differing biases, then adding a player with the average bias can only help). To state this formally, consider first the expected payoff of player  $i$  when choosing room  $R_i$  and expecting a given (e.g. equilibrium) room allocation:

$$\begin{aligned} & - \mathbb{E} \left[ \left( \sum_{j \in R_i^{truth}, j \neq i} (\mu_{ij} - \theta_j) + \sum_{j \notin R_i, j \in R_i^{bab}} (\bar{\mu}_j - \theta_j) \right)^2 \right. \\ & \quad + \alpha \sum_{j \in R_i, j \neq i} \left( b_j - b_i + \sum_{k \in R_i^{truth} \cup \{j\}} (\mu_{jk} - \theta_k) + \sum_{k \notin R_i, k \in R_i^{bab} \setminus \{j\}} (\bar{\mu}_k - \theta_k) \right)^2 \\ & \quad \left. + \alpha \sum_{j \notin R_i} \left( b_j - b_i + \sum_{k \in R_j^{truth} \cup \{j\}} (\mu_{jk} - \theta_k) + \sum_{k \notin R_i, k \in R_j^{bab} \setminus \{j\}} (\bar{\mu}_k - \theta_k) \right)^2 \right] \end{aligned}$$

where we denote  $\mathbb{E}[\theta_j]$  as  $\bar{\mu}_j$ , the set of players babbling in room  $R_j$  in the messaging equilibrium of the given room allocation as  $R_j^{bab}$  and the set of players sending truthful messages in room  $R_j$  in the messaging equilibrium of the given room allocation as  $R_j^{truth}$ . Note that most of the terms drop out in the expression above as signals are assumed to be independent and therefore  $\mathbb{E}[\mu_{ij} - \theta_j] = 0$  and also  $\mathbb{E}[(\mu_{ij} - \theta_j)(\mu_{ik} - \theta_k)] = 0$ . Consequently, the expression above can be rewritten as

$$\begin{aligned} & - \sum_{j \in R_i^{truth}, j \neq i} \mathbb{E} [(\mu_{ij} - \theta_j)^2] - \sum_{j \notin R_i, j \in R_i^{bab}} \mathbb{E} [(\bar{\mu}_j - \theta_j)^2] \\ & - \alpha \sum_{j \in R_i, j \neq i} (b_j - b_i)^2 - \alpha \sum_{j \in R_i, j \neq i} \sum_{k \in R_i^{truth} \cup \{j\}} \mathbb{E} [(\mu_{jk} - \theta_k)^2] - \alpha \sum_{j \in R_i, j \neq i} \sum_{k \notin R_i, k \in R_i^{bab} \setminus \{j\}} \mathbb{E} [(\bar{\mu}_k - \theta_k)^2] \\ & - \alpha \sum_{j \notin R_i} (b_j - b_i)^2 - \alpha \sum_{j \notin R_i} \sum_{k \in R_j^{truth} \cup \{j\}} \mathbb{E} [(\mu_{jk} - \theta_k)^2] - \alpha \sum_{j \notin R_i} \sum_{k \notin R_i, k \in R_j^{bab} \setminus \{j\}} \mathbb{E} [(\bar{\mu}_k - \theta_k)^2] \end{aligned}$$

As we cannot rule out mixed strategies, this expression will not simplify as neatly as in the main text. However, we can already see from here that a player's payoff is higher if another player is truthtelling than when he is babbling or mixing. This observation will be enough for our purposes.

To state our results we first introduce some notation. Let  $\underline{\sigma} = \min\{|\sigma^j - \sigma^k| : j \neq k, \sigma^j, \sigma^k \in \Sigma\}$  and  $\bar{\sigma} = \max\{|\sigma^j - \sigma^k| : j \neq k, \sigma^j, \sigma^k \in \Sigma\}$  and furthermore,  $\bar{b} = \max_i\{|nb_i - \sum_j b_j|\}$ . We will denote by  $\mathcal{B}_\eta$  the set of biases scaled by  $\eta$ ; that is, it contains all the elements  $\eta b_i$ . We will use this to talk about more spread out biases. If the set of biases is  $\mathcal{B}_\eta$  with  $\eta > 1$ , then biases are more spread out.

**Proposition 23.** *If  $\underline{\sigma} \geq 2\bar{b}/(n-1)$ , then a single room in which all players are truthtelling is both welfare maximizing and an equilibrium.*

*Let the set of biases be  $\mathcal{B}_\eta$  and fix all parameter values apart from  $\eta$ . Generically, full separation is welfare maximizing and an equilibrium if  $\eta$  is sufficiently high.*

**Proof of proposition 23:** Recall that a truthtelling equilibrium exists if and only if for all players  $i$   $\left| \sum_{k \neq i} \{b_k/(n-1)\} - b_i \right| \leq (\sigma^h - \sigma^l)/2$  for every  $\sigma^h > \sigma^l$  in  $\Sigma$ . This can be rewritten as  $|\sum_k \{b_k\} - nb_i|/(n-1) \leq (\sigma^h - \sigma^l)/2$ . The condition in the proposition ensures that this inequality holds for all players and all signals. Clearly, having all players in one room and telling the truth is welfare optimal whenever it is feasible.

If  $\left| \sum_{k \in R_i, k \neq i} \{b_k/(n-1)\} - b_i \right| > (\sigma^h - \sigma^l)/2$ , then  $i$  will not be truthful when receiving either signal  $\sigma^l$  or  $\sigma^h$ . Generically,  $\left| \sum_{k \in R_i, k \neq i} \{b_k/(n-1)\} - b_i \right| \neq 0$  for any room configuration containing players from more than one bias group. (This follows from the finiteness of players which obviously implies that the number of such room configurations is finite.) Now observe that the left hand side of the non-truthtelling inequality is scaled by  $\eta$  while the right hand side is not. That is, for  $\eta$  sufficiently high player  $i$  will report the highest (lowest) signal in  $\Sigma$  in all rooms in which  $\sum_{k \in R_i, k \neq i} b_k < n_{R_i} b_i$  ( $\sum_{k \in R_i, k \neq i} b_k > n_{R_i} b_i$ ). Put differently, any room that contains one or more players of a bias not equal to  $b_i$  will lead to totally uninformative messages by  $i$  if  $\eta$  is sufficiently high. For high enough  $\eta$ , this holds true for all players and it is then obvious that full separation is both welfare maximizing and an equilibrium.  $\square$

## 10.2. Continuum of signals

In this subsection we consider the messaging game in a setup that differs from the one in the paper by assuming that the signal space is not binary but a continuum. That is, we take the room allocation as given and analyze equilibrium messaging strategies. Room choice is considered briefly towards the end of the section. Occasionally, we will refer to a player's signal as his "type". Instead of replicating the derivation of  $\Delta U_i(\sigma_i)$  from section 10.1 we simply refer the reader at some points to it.

The main reason why our model is so tractable is that we can consider equilibrium incentives player by player. That is, player  $j$ 's messaging strategy does not affect player  $i$ 's incentives when deciding which message to take. This can be nicely seen from (12) where the only factors influencing preferences over messages are the beliefs induced by the messages, the bias distribution in the room and player  $i$ 's signal.

We will first derive a few results that apply to general finite as well as infinite signal spaces. Signals are without loss of generality viewed as posteriors, i.e.  $\sigma_i^k = \mathbb{E}[\theta_i | \sigma_i^k]$ . Similarly, messages can – in equilibrium – be equated with the beliefs they induce. The following lemma states that the support of player  $i$ 's message strategy is quite small: In fact, each type mixes at most between two messages and these messages are in some sense “adjacent”.

**Lemma 9.** *In equilibrium, the support of type  $\sigma_i^k$ 's strategy consists of at most two elements. If type  $\sigma_i^k$  mixes between two messages, then there is no message inducing a belief in between the two beliefs induced by the messages in his support.*

**Proof of lemma 9:** The indifference condition requires that  $\sigma_i$  is indifferent between any two messages in his support. Denoting the messages as  $l$  and  $h$  which lead in equilibrium to beliefs  $\mu^l$  and  $\mu^h$  (by the other players concerning  $\theta_i$ ), this indifference condition can, as derived in section 10.1, be written as

$$\frac{\mu^l + \mu^h}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i - \sigma_i = 0. \quad (12)$$

The crucial insight is that – given that  $\sigma_i$  is indifferent between  $\mu^l$  and  $\mu^h$ ,  $\sigma_i$  strictly prefers inducing any belief  $\tilde{\mu} \in (\mu^l, \mu^h)$  to either  $\mu^l$  or  $\mu^h$ . To see this note that

$$\frac{\mu^l + \tilde{\mu}}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i - \sigma_i < 0 < \frac{\tilde{\mu} + \mu^h}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i - \sigma_i$$

by the indifference condition. This implies that  $\tilde{\mu}$  is strictly preferred to  $\mu^l$  and  $\mu^h$  (see  $\Delta U_i(\sigma_i)$  as derived in section 10.1). It follows that a type can only mix between two messages  $\mu^l$  and  $\mu^h$  in equilibrium if these two beliefs are “adjacent”, i.e. there is no message inducing a belief between  $\mu^l$  and  $\mu^h$ .  $\square$

In case of mixing, each message is only used by few signal types. Furthermore, there is a standard order property in the sense that higher types send higher messages.

**Lemma 10.** *Each message is used by at most two types that use truly mixed strategies. If a type  $\sigma_i$  mixes between  $\mu^l$  and  $\mu^h > \mu^l$ , then  $\sigma_i^k$  is the highest (lowest) type using message  $\mu^l$  ( $\mu^h$ ).*

**Proof of lemma 10:** Suppose to the contrary that three types  $\sigma_i^k$  with  $k = 1, 2, 3$  (i) use truly mixed strategies and (ii) use a message inducing belief  $\mu$  with positive probability. As each type mixes only over two adjacent messages (see lemma 9), this would imply that at least two of the three types have the same support. Clearly, indifference condition (12) cannot be satisfied for different types and the same support  $\mu^l$  and  $\mu^h$ . Consequently, each message is used at most by two types that mix.



From the indifference condition, (12), and the expression  $\Delta U_i(\sigma_i)$  it is clear that all types below (above)  $\sigma_i$  strictly prefer  $\mu^l$  over  $\mu^h$  ( $\mu^h$  over  $\mu^l$ ).  $\square$

The order property of the previous lemma can be extended. In equilibrium, higher signal types send weakly higher messages. This does not exclude the possibility that one signal type mixes or that several signal types pool on the same message.

**Lemma 11.** *The induced belief  $\mu^k$  is weakly increasing in the received signal  $\sigma_i^k$ .*

**Proof of lemma 11:** Take two signal types  $\sigma_i^h$  and  $\sigma_i^l$  with  $\sigma_i^h > \sigma_i^l$ . Suppose contrary to the lemma that  $\mu(\sigma_i^l) > \mu(\sigma_i^h)$ . In equilibrium  $\sigma_i^l$  must prefer sending his message to sending the message that  $\sigma_i^h$  sends in equilibrium, i.e.

$$\frac{\mu(\sigma_i^l) + \mu(\sigma_i^h)}{2} + \sum_{k \in R_i, k \neq i} \left\{ \frac{b_k}{n_{R_i} - 1} \right\} - b_i - \sigma_i^l \leq 0.$$

If the previous inequality holds, then it holds strictly with  $\sigma_i^h$  in place of  $\sigma_i^l < \sigma_i^h$ . That is,  $\sigma_i^h$  strictly prefers  $\mu(\sigma_i^l)$  over  $\mu(\sigma_i^h)$  which contradicts that  $\sigma_i^h$  induces  $\mu(\sigma_i^h)$  in equilibrium.  $\square$

After these preliminaries, we turn now to a model with a continuum of signals. Let signal  $\sigma_i$ , i.e. player  $i$ 's ex post belief  $\mathbb{E}[\theta_i]$ , be distributed according to some distribution  $\Phi$  with density  $\phi > 0$  on an interval, say  $[0, 1]$  for simplicity. Lemmas 10 and 11 imply then that the equilibrium is a partition of  $[0, 1]$ . Note that truthfulness, i.e. truthfully revealing one's signal no matter what the signal is, is only feasible if  $\Delta_i \equiv b_i - \sum_{k \in R_i, k \neq i} b_k / (n_{R_i} - 1) = 0$ . That is, truthfulness is only an equilibrium if all players in a room share the same bias. Furthermore, the partition will be finite with the maximal number of partition elements being less than  $1 + 1/(2|\Delta_i|)$ . Finiteness is straightforward: If the partition was not finite, there would be types  $\sigma_i$  for which  $\mu(\sigma_i) \approx \sigma_i$  and messages arbitrarily close to  $\sigma_i$  exist. But for  $\Delta_i \neq 0$ , some of these types would clearly want to misrepresent. The upper bound on the number of partition elements follows from the following observation: Let  $\sigma_i^{t-1}$ ,  $\sigma_i^t$  and  $\sigma_i^{t+1}$  be consecutive partition boundary type in an equilibrium partition. Then  $\sigma_i^t$  has to be indifferent between the two messages  $\mu^t = \mathbb{E}[\sigma_i | \sigma_i \in [\sigma_i^{t-1}, \sigma_i^t]]$  and  $\mu^{t+1} = \mathbb{E}[\sigma_i | \sigma_i \in [\sigma_i^t, \sigma_i^{t+1}]]$  which means

$$\mu^{t+1} + \mu^t = 2\sigma_i^t + 2\Delta_i.$$

As  $\mu^{t+1} \leq \sigma_i^{t+1}$  and  $\mu^t \leq \sigma_i^t$ , this implies that  $\sigma_i^{t+1} + \sigma_i^t \geq 2\sigma_i^t + 2\Delta_i$  or equivalently  $\sigma_i^{t+1} - \sigma_i^t \geq 2\Delta_i$ . Hence, every partition element (with exception of the first) has length of at least  $2\Delta_i$  if  $\Delta_i > 0$  (for  $\Delta_i < 0$  a similar argument using lower instead of upper bounds for  $\mu$  works analogously). If a partition equilibrium with  $T$  partition elements exists, then it can be computed similarly to Crawford and Sobel (1982): Say  $\Delta_i > 0$  and denote the partition by  $(\sigma_i^0 = 0, \sigma_i^1, \dots, \sigma_i^T = 1)$ . For  $t \in \{1, \dots, T-1\}$ , the indifference

condition of  $\sigma_i^t$  determines  $\sigma_i^{t+1}$ , i.e.

$$\frac{\int_{\sigma_i^t}^{\sigma_i^{t+1}} \sigma_i d\Phi(\sigma_i)}{\Phi(\sigma_i^{t+1}) - \Phi(\sigma_i^t)} = 2\sigma_i^t + 2\Delta_i - \frac{\int_{\sigma_i^{t-1}}^{\sigma_i^t} \sigma_i d\Phi(\sigma_i)}{\Phi(\sigma_i^t) - \Phi(\sigma_i^{t-1})}. \quad (13)$$

That is, as soon as  $\sigma_i^1$  is fixed, all other values are determined inductively by this condition. Note that not any  $\sigma_i^1$  belongs to an equilibrium partition as eventually the indifference condition for  $\sigma_i^{T-1}$  has to yield  $\sigma_i^T = 1$ . If the following monotonicity condition (M) holds, then there is an essentially unique equilibrium with  $T$  partition elements for all  $T$  up to some  $\bar{T}$ .

(M): Partition cutoff types obtained from some  $\sigma_i^1$  through induction by (13) are increasing in  $\sigma_i^1$ , i.e.  $\sigma_i^t(\sigma_i^1) > \sigma_i^t(\sigma_i^{1'})$  if and only if  $\sigma_i^1 > \sigma_i^{1'}$ .

To give an example, suppose  $\Phi$  is the uniform distribution on  $[0, 1]$ . Then  $\mu^t = (\sigma_i^{t-1} + \sigma_i^t)/2$  and (13) becomes

$$\sigma_i^{t+1} = \sigma_i^t + (\sigma_i^t - \sigma_i^{t-1}) + 4\Delta_i$$

which clearly satisfies (M). For  $t \geq 2$ , this can be solved as

$$\sigma_i^t = t\sigma_i^1 + 4\Delta_i \sum_{j=1}^{t-1} j.$$

A  $T$  element partition has to satisfy  $\sigma_i^T = 1$  or  $1 = T\sigma_i^1 + 4\Delta_i \sum_{j=1}^{T-1} j$  which means that  $\sigma_i^1(T) = [1 - 4\Delta_i \sum_{j=1}^{T-1} j] / T$ . If  $1 - 4\Delta_i \sum_{j=1}^{T-1} j < 0$ , then no equilibrium with  $T$  partition elements exists. This illustrates that a higher  $\Delta_i$  leads to a less informative equilibrium in the sense that there are less partition elements. The derivation of the most informative equilibrium for the case  $\Delta_i < 0$  is analogous.

Regarding room choice, we do not attempt a full characterization of the equilibrium. However, our result that sufficiently large polarization makes segregation generically optimal follows directly from the derivations above: Let the set of biases be such that no possible room has  $\Delta_i = 0$  for some player  $i$  unless the room consists only of players sharing the same bias. This is satisfied for generic bias values. Consider a  $\mathcal{B}_\eta$  scaling of the biases as in the paper. Note that  $\Delta_i$ , now denoted as  $\Delta_i(\eta) = \eta\Delta_i(1)$ , scales linearly in  $\eta$ . For  $\eta$  sufficiently high the upper bound on the number of partition elements  $1 + 1/(2|\Delta_i(\eta)|)$  will be below 2 and babbling will be the only equilibrium. This is true in all possible rooms not consisting of only players sharing the same bias. Note that the number of possible rooms is finite due to the finite number of the players and therefore there exists a  $\bar{\eta}$  such that babbling is the unique equilibrium in all rooms in which players do not share the same bias for all  $\eta \geq \bar{\eta}$ . It follows immediately that full segregation is optimal and an equilibrium for  $\eta \geq \bar{\eta}$ . Similarly, it is straightforward that equilibrium partitions

can be arbitrarily fine as  $\eta \rightarrow 0$ . Consequently, full integration is welfare optimal and an equilibrium for  $\eta$  sufficiently low.<sup>13</sup>

### 10.3. Single state

In this variation, a state of the world  $\theta \in \Theta$  is distributed according to distribution  $F$ . The state is unobserved but each player  $i$  out of  $n$  players receives a noisy signal  $\sigma_i \in \Sigma$  of the state where  $\sigma_i$  is conditional on  $\theta$  distributed according to  $G_\theta$ . The signals are private and – conditional on the state – independent across players. (The latter assumption is relaxed at the end of this subsection.) After observing his signal, a player can access one of  $K \geq 2$  “rooms” and send a message  $m_i \in \mathcal{M}$ . The message is received by all players in the same room. Afterwards each player takes an action  $a_i$ .

The payoff of player  $i$  is  $u(a, b_i, \theta) = -(a_i - b_i - \theta)^2 - \alpha \sum_{j \neq i} (a_j - b_i - \theta)^2$  where  $a$  denotes the vector of actions of all players and  $b_i \in \mathcal{B}$  is a commonly known “bias” of player  $i$ . That is, player  $i$  would like that all players choose the action  $b_i + \theta$ . The parameter  $\alpha$  measures the relative weight players assign to other players’ behavior. Players are assumed to maximize expected utility.

The solution concept used is perfect Bayesian Nash equilibrium.

For simplicity, let  $\Theta = \{\theta^h, \theta^l\}$  and  $\Sigma = \{\sigma^l, \sigma^h\}$  and the signal structure is such that  $\text{prob}(\sigma^j | \theta^j) = p > 1/2$ . We let the message space be binary as well:  $\mathcal{M} = \{h, l\}$ . Furthermore, we let  $\mathcal{B} = \{0, b\}$  and assume that there is at least one player with each of the two biases.

**Action choice** Denote the belief of player  $i$  that the state of the world is high by  $\mu_i$  (after observing his signal and listening to all the messages in his room). The expected utility of player  $i$  can then be written as

$$\begin{aligned} U(a, \mu_i) &= -a_i^2 - \mathbb{E}[(b_i + \theta)^2] + 2a_i(b_i + \mathbb{E}[\theta]) - \alpha \sum_{j \neq i} \mathbb{E}[(a_j - b_i - \theta)^2] \\ &= -a_i^2 - \mu_i(b_i + \theta^h)^2 - (1 - \mu_i)(b_i + \theta^l)^2 + 2a_i(b_i + \mu_i\theta^h + (1 - \mu_i)\theta^l) \\ &\quad - \alpha \sum_{j \neq i} [\mu_i(a_j - b_i - \theta^h)^2 + (1 - \mu_i)(a_j - b_i - \theta^l)^2] \end{aligned} \quad (14)$$

The optimal action choice of player  $i$  is then

$$a_i^* = b_i + \mathbb{E}[\theta] = b_i + \theta^l + \mu_i(\theta^h - \theta^l). \quad (15)$$

---

<sup>13</sup>To see this it is sufficient to note that (i) full integration is the unique welfare optimal allocation for  $\eta = 0$ , i.e. in a situation in which all players have the same bias, and (ii) information (in the most informative messaging equilibrium) in any given room allocation approaches full information as  $\eta \rightarrow 0$ . This implies that welfare in a given room allocation approaches welfare under full information in this room allocation as  $\eta \rightarrow 0$ .

**Cheap talk** The cheap talk game can – as usual – have several equilibria. There is always a babbling equilibrium where the message is independent of the observed signal and therefore nothing about the state of the world is learned, e.g.  $m_i(\sigma_i) = \sigma^h$  for all  $\sigma_i \in \Sigma$  and  $\mu_i = p$  ( $\mu_i = 1 - p$ ) if  $\sigma_i = \sigma^h$  ( $\sigma_i = \sigma^l$ ). We will focus on most informative equilibria, that is equilibria where  $m_i(\sigma_i) = \sigma_i$  with as high probability as possible.

Truthful communication is an equilibrium for a given room if all players in this room have the same  $b_i$ . To see this, suppose player  $i$  could maximize his expected utility (1) not only over  $a_i$  but also over the  $a_j$  of all the players in his room. Clearly, he would choose the same action for everyone namely  $b_i + \theta^l + \mu_i(\theta^h - \theta^l)$ . Deviating from the truthful strategy is not profitable because by adhering to truthfulness player  $i$  ensures that all other players in the room choose precisely the action he would have chosen for them (while deviating changes the other players' beliefs and therefore their optimal action). Note that this argument depends on all players having the same bias and truthful communication is normally not an equilibrium if players in a given room have different biases. We state this result for future reference in the following lemma.

**Lemma 12.** *If all players in a given room have the same bias, truthful communication in this room is the most informative equilibrium of the cheap talk game (taking room choice as given).*

We will now analyze the cheap talk problem in rooms in which players with both types of biases are present. In particular, we will be interested in the case of strong differences in opinion, i.e. the case where  $b$  is sufficiently large.

**Lemma 13.** *Let  $n_0 \geq 1$  players with bias  $b_i = 0$  and  $n_b \geq 1$  players with bias  $b_i = b$  be in a room. There exists a  $\bar{b}$  such that for  $b \geq \bar{b}$  babbling is the only equilibrium of the cheap talk game.*

**Proof of lemma 13:** Suppose that there is a non-babbling equilibrium, i.e. an equilibrium where belief  $\mu_j$  depends on the messages of players  $i \neq j$ . Let  $i$  be a player affecting  $j$ 's belief. Without loss of generality, say  $\mu_j$  is lower if  $i$  sends the message  $l$  and higher if  $i$  sends the message  $h$ . By Bayesian updating and independence of the signals,  $\mu_k$  will then be lower when  $i$  sends message  $l$  than when he sends message  $h$  for all  $k \neq i$ . (Moreover two players that observe the same signal themselves and are in the same room will have the same belief because of Bayesian updating and independence of signals.) Hence it is without loss of generality to assume that  $b_i \neq b_j$ . For concreteness, let  $b_i = 0$  and  $b_j = b$  (the proof for the opposite case is analogous).

Now suppose  $i$  observes signal  $\sigma^h$ . We will show that it is optimal for  $i$  to send message  $l$  if  $b$  is sufficiently high. To see this, denote the change in  $i$ 's expected utility (14) when

sending message  $l$  instead of message  $h$  as  $\Delta U_i$ <sup>14</sup>

$$\begin{aligned}
\Delta U_i &= -\alpha \sum_{j \neq i} \mathbb{E} [a_j(l)^2 - a_j(h)^2 - 2\theta(a_j(l) - a_j(h))] \\
&= -\alpha \sum_{j \neq i} \mathbb{E} \left[ (\mu_j(l)^2 - \mu_j(h)^2) (\theta^h - \theta^l)^2 + 2(\mu_j(l) - \mu_j(h))(\theta^h - \theta^l)(b_j + \theta^l - \theta) \right] \\
&= \alpha \sum_{j \neq i} \mathbb{E} [(\mu_j(h) - \mu_j(l)) (\theta^h - \theta^l) * ((\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2(b_j + \theta^l - \theta))] \\
&= \alpha (\theta^h - \theta^l) n_b \mathbb{E} [(\mu_j(h) - \mu_j(l)) * ((\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2(b + \theta^l - \theta))] \\
&\quad + \alpha (\theta^h - \theta^l) (n_0 - 1) \mathbb{E} [(\mu_j(h) - \mu_j(l)) * ((\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2(\theta^l - \theta))] \\
&= \alpha (\theta^h - \theta^l) (n_b + n_0 - 1) \\
&\quad \mathbb{E} \left[ (\mu_j(h) - \mu_j(l)) * \left( (\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2 \left( \frac{n_b b}{n_b + n_0 - 1} \theta^l - \theta \right) \right) \right].
\end{aligned}$$

If  $b \geq \theta^h(n_b + n_0 - 1)/(\theta^l n_b)$ , the term inside the expectation is positive for any  $\theta$  and therefore  $\Delta U_i$  is definitely strictly positive. Hence,  $i$  strictly prefers sending message  $l$  to message  $h$  and  $i$  receives signal  $\sigma^h$ . This would imply that  $i$  sends message  $l$  with probability 1 if the signal is  $\sigma^h$  in this equilibrium. But this contradicts that  $\mu_j(l) > \mu_j(h)$ . Hence, choosing  $\bar{b} = \theta^h N / \theta^l$  where  $N$  is the total number of agents implies  $\bar{b} \geq \theta^h(n_b + n_0 - 1)/(\theta^l n_b)$  and gives the result.  $\square$

Lemma 13 implies that – given a finite number of players – the only way allowing meaningful communication if differences in opinion is high is to have only players with the same bias in a room.

If the differences in opinion are minimal, i.e.  $b$  is very low, truthful communication is an equilibrium for any room composition. The reason is the coarseness of the signal structure: Lying in the message game leads – in a truthful equilibrium – to a discrete reaction of all other players in the room. If the difference in bias is very small, this discrete reaction is “too high”, i.e. even those players with a (slightly) different bias react more than the deviating player would wish for. The following lemma formalizes this generalization of lemma 12.

**Lemma 14.** *Let there be  $n_0 \leq n$  players with  $b_i = 0$  and  $n_b \leq n - n_0$  players with  $b_i = b$  in a room. There exists a  $\underline{b} > 0$  such that for  $b \leq \underline{b}$  truthful communication is an equilibrium.*

**Proof of lemma 14:** For  $b = 0$ , truth-telling is strictly better than lying (given that all other players tell the truth). Note that  $i$ 's expected utility is continuous in  $a_j$  and  $a_j^*$  is continuous in  $b_j$ , see (15). Hence,  $U_i$  is continuous in  $b_j$ . However,  $\mu_j$  and therefore  $a_j^*$  reacts discretely to lying. Consequently, truth-telling is still a best response to truth-telling for  $b_j > 0$  sufficiently small.  $\square$

<sup>14</sup>For a more general proof, one could already go from the first line to  $\alpha \sum_{j \neq i} \mathbb{E} [(a_j(h) - a_j(l)) * (a_j(h) + a_j(l) - 2\theta)]$  and then note that for  $b$  high enough even  $a_j(l) > \theta^h$ .

From lemma 13 and lemma 14 we know that for  $b$  low the most informative equilibrium in a room with a given configuration is truth-telling and for  $b$  sufficiently high the “most informative” equilibrium is babbling if players with different biases are present. It seems most likely that  $\bar{b} > \underline{b}$ . In this case, there are mixed strategy equilibria for  $b \in (\underline{b}, \bar{b})$ .

**Room choice equilibria** We claim that separation is an equilibrium if differences in opinion, i.e. the parameter  $b$ , are sufficiently high.

**Proposition 24.** *If  $b \geq \bar{b}$ , the following strategies constitute an equilibrium:*

1. *Players with bias 0 ( $b$ ) go to room 0 (1).*
2. *A player sends truthful messages if only players of the same type are in his room and babbles otherwise.*
3. *Actions are taken according to (15) and beliefs  $\mu_i$  are formed using Bayes’ rule (given the equilibrium strategies in 1 and 2).*

*This equilibrium is the most informative equilibrium in the sense that no player has more precise information about the state  $\theta$  in any other equilibrium.*

**Proof of proposition 24:** Given lemma 13, unilateral deviations to other rooms are not profitable: Any such deviation would either lead to being alone in a room or babbling. In either case, the deviating player does not have any information beyond his own signal about the state of the world. This reduces his expected utility directly. Furthermore, deviations lead to less information for other players which again lowers the deviating player’s payoff: Less information for players with the same bias as player  $i$  implies that their actions are further away from  $b_i + \theta$  in expectation. Furthermore, the players with  $b_j \neq b_i$  choose actions further away to  $b_j + \theta$  if they have less information, i.e. variance of their choice is increased while the expected value stays the same. Given the strictly concave loss function, player  $i$  loses from this as well.

Lemmas 12 and 13 imply that no profitable deviation in the cheap talk stage exists. As (15) gives the optimal action (given one’s beliefs), no deviation in choosing one’s action is profitable either.

By lemma 13, a given player  $i$  cannot observe more “non-babbling” messages than in the suggested equilibrium in any other equilibrium. Given that communication is truthful in the suggested equilibrium, player  $i$  can therefore not have more precise information about  $\theta$  in any other equilibrium.  $\square$

For  $b \leq \underline{b}$ , the most informative equilibrium is clearly that every player goes to the same room and truthfully reports his signal.

**Welfare optimal room allocation** Suppose a social planner could allocate players to rooms. After being assigned a room, players play the same game as above; that is, the planner has no influence on messages or actions. We claim that for  $b \geq \bar{b}$  the welfare optimal allocation is to assign everyone with bias 0 in one room and everyone with bias  $b$  in another room, i.e. the equilibrium described in proposition 24 is welfare optimal. The idea is the following: For  $b \geq \bar{b}$ , the cheap talk game in a room where players with both bias types are present will only have a babbling equilibrium by lemma 13. Consequently, any room allocation that assigns players with different biases to the same room will lead to completely uninformative messages and is therefore equivalent to putting every player to a separate room. By assigning players with the same bias to the same room, the planner achieves the most informative equilibrium. That is, truthful communication is possible in each room. The additional information ensures that player with the same bias as player  $i$  choose actions closer to  $b_i + \theta$ . Furthermore, the players with  $b_j \neq b_i$  choose actions closer to  $b_j + \theta$ , i.e. the variance is reduced while the expected value stays the same. Given the strictly concave loss function, player  $i$  gains from this as well. Note that the welfare notion can be chosen quite strict in the sense that the described allocation maximizes the welfare of every agent. That is, if agent  $i$  could dictatorially decide the room allocation (without having any influence on the messages or actions taken by other players), the same allocation would result.

Similarly, the most informative equilibrium is welfare optimal in the strong sense established above if  $b \leq \underline{b}$ .

**Correlated signals** Finally, we want to discuss an extension to this model: People with similar biases might be similar in other respects and therefore have similar information. More precisely, one could imagine that the signals of people with the same bias are positively correlated conditional on the state. The following paragraphs shows that similar results as before hold when signals are correlated.

The main difficulty is to show a result similar to lemma 13 all other results go through without change. The following lemma states that in the limit as  $b$  grows large no information can be transmitted in equilibrium. The result is somewhat weaker than lemma 13 but similar in nature.

**Lemma 15.** *Let  $n_0 \geq 1$  players with bias  $b_i = 0$  and  $n_b \geq 1$  players with bias  $b_i = b$  be in a room. Let the signal technology be such that signals are not perfectly correlated and such that all signal vectors have strictly positive probability. For every  $\varepsilon > 0$ , there exists a  $b_\varepsilon$  such that  $\mathbb{E}_{m_{-i}, \sigma_j} [\mu_j(m_i = \sigma^h) - \mu_j(m_i = \sigma^l) | \sigma_i] < \varepsilon$  in every equilibrium.*

**Proof of lemma 15:** Suppose that there is a non-babbling equilibrium, i.e. an equilibrium where belief  $\mu_j$  depends on the messages of players  $i \neq j$ . Let  $i$  be a player affecting  $j$ 's belief. Without loss of generality, say  $\mu_j$  is lower if  $i$  sends the message  $l$  and higher if

$i$  sends the message  $h$ . By Bayesian updating,  $\mu_k$  will then be lower when  $i$  sends message  $l$  than when he sends message  $h$  for all  $k \neq i$ . First, let  $b_i \neq b_j$ . For concreteness, let  $b_i = 0$  and  $b_j = b$  (the proof for the opposite case is analogous).

Now suppose  $i$  observes signal  $\sigma^h$ . To make an informative equilibrium possible, the change in  $i$ 's expected utility (1) when sending message  $l$  instead of message  $h$ ,  $\Delta U_i$ , must not be strictly positive:

$$\begin{aligned}
\Delta U_i &= -\alpha \sum_{j \neq i} \mathbb{E} [a_j(l)^2 - a_j(h)^2 - 2\theta(a_j(l) - a_j(h))] \\
&= -\alpha \sum_{j \neq i} \mathbb{E} [(\mu_j(l)^2 - \mu_j(h)^2) (\theta^h - \theta^l)^2 + 2(\mu_j(l) - \mu_j(h))(\theta^h - \theta^l)(b_j + \theta^l - \theta)] \\
&= \alpha \sum_{j \neq i} \mathbb{E} [(\mu_j(h) - \mu_j(l)) (\theta^h - \theta^l) * ((\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2(b_j + \theta^l - \theta))] \\
&> \alpha \sum_{j \neq i} \mathbb{E} [(\mu_j(h) - \mu_j(l)) (\theta^h - \theta^l) * ((\mu_j(h) + \mu_j(l))(\theta^h - \theta^l) + 2(b_j + \theta^l - \theta^h))] \\
&= \alpha (\theta^h - \theta^l) (n_b b \mathbb{E} [\mu_j(h) - \mu_j(l) | b_j = b, \sigma_i = \sigma^h] \\
&\quad + \sum_{j \neq i} \mathbb{E} [(\mu_j(h) - \mu_j(l))(\mu_j(h) + \mu_j(l) - 2)(\theta^h - \theta^l)]) \\
&> \alpha (\theta^h - \theta^l) (n_b b \mathbb{E} [\mu_j(h) - \mu_j(l) | b_j = b, \sigma_i = \sigma^h] - 2N(\theta^h - \theta^l))
\end{aligned}$$

where  $N = n_b + n_0$ . Clearly, the last expression is greater than zero if  $b \mathbb{E} [\mu_j(h) - \mu_j(l) | b_j = b, \sigma_i = \sigma^h] > 2N(\theta^h - \theta^l)/n_b$ . Hence,  $b_\varepsilon = 2N(\theta^h - \theta^l)/(n_b \varepsilon)$  gives the result in the lemma.

Second, let  $b_j = b_i$ . Note that the result above says that  $i$ 's message contains no information in the limit as  $b \rightarrow \infty$ . It follows that given that the signal technology is (i) not perfectly correlated and (ii) puts strictly positive probability on all signal vectors, the result has to hold also for  $j$  with  $b_j = b_i$ .<sup>15</sup>  $\square$

## 11. Follower model

This section replicates our results for a slightly different model in which instead of choosing “rooms” in the first stage, players choose which other players to “follow”. Players are unrestricted regarding the size and composition of the set of players they follow. In the second stage every player sends one cheap talk message to his “followers”. We will adopt the convention that each player follows himself (which is immaterial for the results but allows us to proceed with some of the derivations analogously to the paper.) Signal technology and preferences are the same as in the paper.

<sup>15</sup>The two assumptions avoid that lying leads to a zero probability event where beliefs cannot be determined by Bayes' rule.



Clearly, the optimal action is still

$$a_i^* = b_i + \sum_{j=1}^n \mathbb{E}[\theta_j].$$

More importantly, lemma 1 still applies and we can concentrate on pure strategy equilibria.

**Lemma 16.** *Let  $(m_1, \dots, m_n)$  be equilibrium strategies. If  $m_i$  is a mixed strategy, then there also exists an equilibrium with strategies  $(m_i^t, m_{-i})$ , where  $m_i^t$  is the truthful strategy.*

**Proof.** Denoting  $i$ 's followers by  $F_i$ , the set of player  $i$  is following by  $f_i$  and fixing some equilibrium  $(m_1, \dots, m_n)$ , player  $i$ 's expected payoff when sending message  $m_i$  to  $F_i$  can be written as

$$U_i(m_i|\sigma_i) = \mathbb{E} \left[ - \left( a_i(m_{-i,R_i}, \sigma_i) - b_i - \sum_{k=1}^n \theta_k \right)^2 - \alpha \sum_{j \notin F_i} \left\{ \left( a_j(m_{-i,f_j}, \sigma_j) - b_i - \sum_{k=1}^n \theta_k \right)^2 \right\} - \alpha \sum_{j \in F_i, j \neq i} \left\{ \left( a_j(m_i, m_{-i,f_j}, \sigma_j) - b_i - \sum_{k=1}^n \theta_k \right)^2 \right\} \middle| \sigma_i \right].$$

which can be split in a part that is independent of  $i$ 's message  $m_i$  and a part that depends on  $m_i$ :

$$U_i(m_i) = \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in F_i, j \neq i} \left( a_j(m_i, m_{-i,f_j}, \sigma_j) - b_i - \sum_{k=1}^n \theta_k \right)^2 \middle| \sigma_i \right].$$

Specifically, sending message  $m^h$  gives expected payoff

$$U_i(m^h) = \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in F_i, j \neq i} \left( b_j - b_i + \mu_{ji}^h + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right]$$

where  $\mu_{ji}^h = \mathbb{E}[\theta_i | m_i = m^h]$ , i.e.  $\mu_{ji}^h$  is the belief of a player  $j$  (following  $i$ ) concerning  $\theta_i$  if player  $i$  sends message  $m^h$ . Note that this belief is the same for all players  $j \neq i$  following  $i$ . Sending message  $m^l$  gives

$$U_i(m^l) = \mathbb{E} \left[ \text{const} - \alpha \sum_{j \in F_i, j \neq i} \left( b_j - b_i + \mu_{ji}^l + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right]$$

where  $\mu_{ji}^l = \mathbb{E}[\theta_i | m_i = m^l]$ . The difference in expected payoff is then

$$\begin{aligned}
\Delta U_i(\sigma_i) &= (U_i(m^h) - U_i(m^l))/\alpha \\
&= - \sum_{j \in F_i, j \neq i} \mathbb{E} \left[ \mu_{ji}^h{}^2 - \mu_{ji}^l{}^2 + 2(\mu_{ji}^h - \mu_{ji}^l) \left( b_j - b_i + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right) \middle| \sigma_i \right] \\
&= -2(\mu_{ji}^h - \mu_{ji}^l) \sum_{j \in F_i, j \neq i} \left[ \frac{\mu_{ji}^h + \mu_{ji}^l}{2} + b_j - b_i - \mathbb{E}[\theta_i | \sigma_i] \right] \\
&= 2(\mu_{ji}^h - \mu_{ji}^l)(n_{F_i} - 1) \left[ -\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \frac{\sum_{j \in F_i, j \neq i} b_j}{n_{F_i} - 1} + b_i + \mathbb{E}[\theta_i | \sigma_i] \right] \tag{16}
\end{aligned}$$

where  $n_{F_i}$  denotes the number of elements in  $F_i$ . (For the transformation to line 3, we make use of the fact that  $\mu_{ji}$  is the same for all  $j \in F_i \setminus \{i\}$ .)

Player  $i$  is only willing to choose a mixed strategy after receiving signal  $\sigma_i$  if  $\Delta U_i(\sigma_i) = 0$ . From expression (16) it is clear that this can only be true for at most one signal as  $\mathbb{E}[\theta_i | \sigma_i]$  varies in  $\sigma_i$ . Furthermore,  $U_i(\sigma^h) = 0$  implies  $U_i(\sigma^l) < 0$  and similarly  $U_i(\sigma^l) = 0$  implies  $U_i(\sigma^h) > 0$ .

Now suppose  $i$ 's equilibrium strategy  $m_i$  is mixed after signal  $\sigma^h$ . Then,  $\Delta U_i(\sigma^h) = 0$  implies  $\Delta U_i(\sigma^l) = 2(\mu_{ji}^h - \mu_{ji}^l)(n_{F_i} - 1)(1 - 2p) < 0$  and therefore  $m_i(\sigma^l) = m^l$  which implies  $\mu_{ji}^h = p$  as a  $m^h$  is only sent by  $i$  after receiving signal  $\sigma^h$ . This implies  $(\mu_{ji}^h + \mu_{ji}^l)/2 \geq 1/2$  as  $\mu_{ji}^l \geq 1 - p$ . Now consider the equilibrium candidate  $(m_i^t, m_{-i})$ . With the truthful strategy  $m_i^t$ ,  $\mu_{ji}^{th} = p$  and  $\mu_{ji}^{tl} = 1 - p$  and therefore  $(\mu_{ji}^{th} + \mu_{ji}^{tl})/2 = 1/2$ . This implies that  $\Delta U_i(\sigma^h) > 0$  in the equilibrium candidate  $(m_i^t, m_{-i})$ , i.e. truthful reporting is optimal for  $i$  after receiving signal  $\sigma^h$ . In the equilibrium candidate  $(m_i^t, m_{-i})$ , truthful messaging is still optimal after signal  $\sigma^l$  as well: From  $p > 1/2$ ,  $\mu_{ji}^h \leq p$  and  $\mu_{ji}^l \leq 1/2$  it follows that  $-1/2 + (1 - p) < -(\mu_{ji}^h + \mu_{ji}^l)/2 + p$ . As in the original equilibrium  $(m_i, m_{-i})$  we had  $\Delta U_i(\sigma^h) = 0$  and therefore  $-(\mu_{ji}^h + \mu_{ji}^l)/2 + p = \sum_{j \in R_i, j \neq i} b_j / (n_{F_i} - 1) + b_i$ , we get that  $-1/2 + 1 - p < \sum_{j \in F_i, j \neq i} b_j / (n_{F_i} - 1) + b_i$  and therefore  $U_i(\sigma^l) < 0$  in the truthful equilibrium candidate  $(m_i^t, m_{-i})$ . Hence, truthful messaging is  $i$ 's best response in the equilibrium candidate  $(m_i^t, m_{-i})$ . Finally, note that the  $\Delta U_j(\sigma_j)$  for  $j \neq i$  is not affected by changing  $i$ 's strategy from  $m_i$  to  $m_i^t$ . Hence,  $(m_i^t, m_{-i})$  is an equilibrium.

The argument in case  $i$ 's strategy is mixed after signal  $\sigma^l$  is analogous.  $\square$

The previous lemma (and its proof) allow a characterization of the equilibrium messaging strategy in the most informative equilibrium and an analogue to theorem 1.

**Theorem 5.** Let  $\bar{b} = \frac{\sum_{k \in F_i} b_k}{n_{F_i}}$  be the mean bias of  $i$ 's followers. In the most informative equilibrium in this room, a player  $i$  tells the truth to his followers if

$$b_i \in \left[ \bar{b} - \frac{n_{F_i} - 1}{n_{F_i}} \left( p - \frac{1}{2} \right), \bar{b} + \frac{n_{F_i} - 1}{n_{F_i}} \left( p - \frac{1}{2} \right) \right]$$

and babbles otherwise.

**Proof.** As in theorem 1 in the paper.  $\square$

We proceed by turning to stage 1. Take some follower allocation as fixed, then the expected payoff of player  $i$  following players in  $f_i$  while having followers  $F_i$  equals

$$U_i = -\mathbb{E} \left[ \left( \sum_{j \in f_i^{truth} \cup \{i\}} (\mu_{ij} - \theta_j) + \sum_{j \notin f_i^{truth} \cup \{i\}} \left(\frac{1}{2} - \theta_j\right) \right)^2 \right. \\ \left. + \alpha \sum_{j \neq i} \left( b_j - b_i + \sum_{k \in f_j^{truth} \cup \{j\}} (\mu_{jk} - \theta_k) + \sum_{k \notin f_j^{truth} \cup \{j\}} \left(\frac{1}{2} - \theta_k\right) \right)^2 \right]$$

where  $f_i^{truth}$  are the players in  $f_i$  that send truthful/informative messages in equilibrium and  $f_i \setminus f_i^{truth}$  are those players in  $f_i$  that are babbling.

For any  $i \neq j$ , the two values of  $\theta_i$  and  $\theta_j$  are independent; the same is true for  $\mu_{ij}$  and  $\mu_{ik}$ . Hence  $\mathbb{E}[\mu_{ij} - \theta_j] = 0$  and  $\mathbb{E}[(\mu_{ij} - \theta_j)(\mu_{ik} - \theta_k)] = 0$ , which means that the above expression can be rewritten as

$$U_i = - \sum_{j \in f_i^{truth} \cup \{i\}} \mathbb{E}[(\mu_{ij} - \theta_j)^2] - \sum_{j \notin f_i^{truth} \cup \{i\}} \mathbb{E}\left[\left(\frac{1}{2} - \theta_j\right)^2\right] \\ - \alpha \sum_{j \neq i} (b_j - b_i)^2 - \alpha \sum_{j \neq i} \sum_{k \in f_j^{truth} \cup \{j\}} \mathbb{E}[(\mu_{jk} - \theta_k)^2] - \alpha \sum_{j \neq i} \sum_{k \notin f_j^{truth} \cup \{j\}} \mathbb{E}\left[\left(\frac{1}{2} - \theta_k\right)^2\right].$$

Now note that  $\mathbb{E}[(\mu_{jk} - \theta_k)^2]$  can have two possible values in the most informative equilibrium: If  $k \in f_j^{truth} \cup \{j\}$ , i.e. if  $j$  has received information about  $\theta_k$ , then  $\mathbb{E}[(\mu_{jk} - \theta_k)^2] = p(1-p)$ . If  $j$  has not received information about  $\theta_k$ , then  $\mathbb{E}[(\mu_{jk} - \theta_k)^2] = \frac{1}{4}$ . (We can check that information always reduces variance and increases welfare since  $p > \frac{1}{2}$  and hence  $p(1-p) < \frac{1}{4}$ .) This allows to denote utility in the notation of the paper using pieces of information

$$U_i = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n-1)n] + (1/4 - p(1-p)) \left[ \zeta_i + \alpha \sum_{j \neq i} \zeta_j \right]$$

and express welfare as

$$W = \sum_i U_i = \sum_i \left[ -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n-1)n] + (1/4 - p(1-p)) \left[ \zeta_i + \alpha \sum_{j \neq i} \zeta_j \right] \right] \\ = -\alpha \sum_{i=1}^n \sum_{j \neq i} \{(b_j - b_i)^2\} - \frac{1}{4} n^2 [1 + \alpha(n-1)] + (p - \frac{1}{2})^2 (1 + \alpha(n-1)) \sum_i \zeta_i.$$

In this expression, all terms are model parameters except for the sum over all  $\zeta_i$ , which

shows that welfare is linearly increasing in  $\sum_i \zeta_i$ .

We are now ready to analyze equilibrium follow decision. The following describes player  $j$ 's best response: Hold arbitrary stage 1 decisions of players other than  $j$  fixed. From the expression for  $U_i$  in terms of pieces of information, it is clear that it is uniquely optimal for  $j$  to follow  $i$  if  $i$  tells the truth given the stage 1 decisions of the other players and  $j$ 's decision to follow. Furthermore,  $j$  is indifferent between following  $i$  and not following  $i$  if  $i$  babbles regardless of  $j$ 's choice. Last but not least,  $j$  optimally does not follow  $i$  if following leads to babbling by  $i$  while not following allows informative messages by  $i$ . This leads to the following result which is in line with the empirically found homophily.

**Proposition 25.** *It is weakly dominant for  $j$  to follow  $i$  if  $|b_j - b_i| \leq (p - 1/2)/2$ .*

**Proof.** From the reasoning of the previous paragraph, it is sufficient to show that  $j$  following  $i$  will not cause  $i$  to babble (given some arbitrary first stage choices of the other players) if  $|b_j - b_i| \leq (p - 1/2)/2$ . If  $j$  follows  $i$ , then  $n_{F_i} \geq 2$  (recall that by convention  $i \in F_i$ ) which implies  $(n_{F_i} - 1)/n_{F_i} \geq 1/2$ . Consequently,  $i$  will tell the truth if  $j$  is the only player following  $i$ . Now consider the case where some players (other than  $j$ ) are following  $i$ . If  $i$  is truthtelling without  $j$  following him, then, by theorem 5,  $b_i \geq \bar{b} - (p - 1/2)(n_{F_i} - 1)/n_{F_i}$  where  $\bar{b}$  is the average bias of players other than  $j$  following  $i$ . By the condition of the proposition  $b_i \geq b_j + (p - 1/2)/2$  and bringing the previous two inequalities together yields

$$b_i \geq \frac{(n_{F_i} - 1)\bar{b} + b_j}{n_{F_i}} - (p - 1/2) \frac{n_{F_i}^2 - 3n_{F_i}/2 + 1}{n_{F_i}^2}$$

which implies

$$b_i \geq \frac{(n_{F_i} - 1)\bar{b} + b_j}{n_{F_i}} - (p - 1/2) \frac{n_{F_i}}{n_{F_i} + 1}.$$

Similarly,  $b_i \leq \bar{b} + (p - 1/2)(n_{F_i} - 1)/n_{F_i}$  and  $b_i \leq b_j + (p - 1/2)/2$  imply

$$b_i \leq \frac{(n_{F_i} - 1)\bar{b} + b_j}{n_{F_i}} + (p - 1/2) \frac{n_{F_i}^2 - 3n_{F_i}/2 + 1}{n_{F_i}^2}.$$

Consequently,  $i$  will be truthtelling when  $j$  follows him if  $i$  is truthtelling when  $j$  does not follow him.  $\square$

The simple characterization of equilibria above makes it straightforward to characterize the structure of the most informative and therefore welfare maximizing equilibrium in stage 1. The welfare optimal set of followers for  $i$  can be determined independently of the welfare optimal set of followers of other players. In fact, it is given by simple maximization problem:

**Proposition 26.** *The welfare optimal set of  $i$ 's followers,  $F_i^*$ , is given by the maximization problem maximizing the number of elements of  $F_i$  subject to the truthtelling constraint*

$$b_i \in \left[ \frac{\sum_{j \in F_i} b_j}{n_{F_i}} - \frac{n_{F_i} - 1}{n_{F_i}}(p - 1/2), \frac{\sum_{j \in F_i} b_j}{n_{F_i}} + \frac{n_{F_i} - 1}{n_{F_i}}(p - 1/2) \right] \quad (17)$$

where  $n_{F_i} = \sum_{j \in F_i} \mathbb{1}_{j \in F_i}$ . The welfare optimal follower allocation  $(F_1^*, \dots, F_n^*)$  is an equilibrium.

**Proof.** Welfare is increasing in the pieces of information provided in equilibrium. The maximal number of pieces of information provided by  $i$  is given by the results of the maximization problems in the proposition. As there are no constraints on how many players to follow,  $(F_1^*, \dots, F_n^*)$  is feasible. It is also an equilibrium: No player  $i$  wants to follow an additional player  $j$  as – by the definition of  $(F_1^*, \dots, F_n^*)$  – this would lead to babbling by  $j$ . As each player is only following players that are truthtelling, no player  $i \in F_j^*$  benefits from not following  $j$ .  $\square$

Note two implications of the previous proposition. First, a pure strategy equilibrium exists. Second, the welfare optimal follower allocation always coincides with the follower allocation in the welfare optimal equilibrium.

Let  $\mathcal{B} = \{b_1, b_2, \dots, b_n\}$  be a bias configuration. (Note that this is not a set, as several people can have the same bias.) Assume that  $\mathcal{B}$  is generic in the sense that no bias is the average of any set of other biases (except in cases where several people have the same bias). Now we can consider an alternative bias configuration  $\mathcal{B}_\eta$ , with  $\eta \in (0, \infty)$ , which for every  $b_i$  in  $\mathcal{B}$  contains  $\eta b_i$ . Then the following is true:

**Theorem 6.** (i) *If  $\eta$  is sufficiently close to 0, full integration, i.e.  $F_i = \{1, \dots, n\}$  for all  $i = 1, \dots, n$ , is welfare-optimal for bias configuration  $\mathcal{B}_\eta$ .*

(ii) *If  $\eta$  is sufficiently large, full segregation by bias types is generically welfare-optimal for bias configuration  $\mathcal{B}_\eta$ .*

**Proof.** Note that the truthtelling constraint (17) for set of biases  $\mathcal{B}_\eta$  can be written as

$$b_i \in \left[ \frac{\sum_{j \in F_i} b_j}{n_{F_i}} - \frac{1}{\eta} \frac{n_{F_i} - 1}{n_{F_i}}(p - 1/2), \frac{\sum_{j \in F_i} b_j}{n_{F_i}} + \frac{1}{\eta} \frac{n_{F_i} - 1}{n_{F_i}}(p - 1/2) \right].$$

For  $\eta \rightarrow 0$ , this constraint is arbitrary slack while for  $\eta \rightarrow \infty$  it is arbitrarily strict. The latter implies that for  $\mathcal{B}_\eta$  such that no element is a convex combination of other elements (not all of which equal to the initial element) no  $F_i$  apart from full segregation can satisfy the constraint.  $\square$

**Example 1.** *As a straightforward example consider the binary case where  $b_i \in \{0, b\}$  for all players. Let  $n_b$  ( $n_0$ ) be the number of players with  $b_i = b$  ( $b_i = 0$ ). The welfare optimal follower allocation is then as follows:  $F_i$  consists of all players  $j$  with  $b_j = b_i$  and  $k$  players*

with  $b_j \neq b_i$  where  $k$  is the highest integer such that  $i$ 's truthtelling constraint still holds. This implies the following: A majority player has more followers than a minority player. A majority player has (weakly) more followers of a different bias type than a minority player. As  $b$  grows larger, players have less and less followers of the other bias type. For  $b$  above some critical  $\underline{b}$  each minority player is only followed by the other members of the minority. For  $b$  above some critical  $\tilde{b} \geq \underline{b}$  each majority player is only followed by the other members of the majority.

Moving away from the welfare optimal equilibrium note that other equilibria exist in stage 1. In particular there are equilibria in which players babble. From the best response structure we immediately get the following result.

**Lemma 17.** *If player  $i$  babbles given followers  $F_i$ , then  $i$  would still babble if his set of followers was  $F_i \setminus \{j\}$  for  $j \in F_i$ .*

**Proof.** Suppose there existed a  $j \in F_i$  such that  $i$  would not babble with set of followers  $F_i \setminus \{j\}$ . In this case,  $j$  has a profitable deviation: Not following  $i$  will increase the number of pieces of information of some other players while it will not reduce the number of pieces he has himself. By  $\alpha > 0$  the deviation is profitable.  $\square$

The lemma indicates that in equilibrium there can be players that babble because they are followed by too many other players. These player are however so much over-subscribed by players with very different biases that they could still not tell the truth if an arbitrary single player decided not to follow them anymore. That is, they are, so to speak, far away from being tempted to tell what they know.

Some interesting comparison between extremists and centrists can be made based on the best response structure of the cheap talk stage. Consider for instance “extremists”, i.e. players with an unusual high or low bias. These players can send truthful messages if they are followed by similarly extreme players. These will typically be only a few people given that only a minority can have “extreme”, i.e. unusually high or low, biases. Now consider a centrist, i.e. someone whose bias is close to the average of the population. He can be followed by (nearly) everyone and he can still be truthtelling. In the extreme case where his bias equals the average bias in the population, indeed everyone will follow him in the welfare optimal equilibrium and he will be truthtelling. Compare this to the extremist: If (sufficiently) many people follow an extremist, he will be babbling. The following proposition uses the same intuition to show that in the welfare optimal equilibrium centrists have more followers than extremists if the distribution of biases is single peaked and symmetric.

**Proposition 27.** *Let biases be distributed on an equally spaced finite grid and let the distribution of biases be single-peaked and symmetric around the mean. Then the number of followers in the welfare maximal equilibrium is lower, the farther a player's bias is away from the mean bias.*

**Proof of proposition 27:** Denote the mean bias in the population by  $\mu_b$  and order – without loss of generality – players according to their biases, i.e.  $b_1 \leq b_2 \leq \dots \leq b_n$ . By proposition 26,  $F_i^*$  is given by  $\max n_{F_i}$  subject to  $|b_i - \sum_{j \in F_i, j \neq i} b_j / (n_{F_i} - 1)| \leq (p - 1/2)$ . Given the assumptions in the proposition, the solution to this maximization problem is straightforward: If  $b_i > \mu_b$ , then  $F_i$  is the set of players  $\{\underline{i}, \underline{i} + 1, \dots, n\}$  where  $\underline{i} \leq i$  is determined such that  $b_i - \sum_{j \geq \underline{i}, j \neq i} b_j / (n_{F_i} - 1) \leq (p - 1/2)$  and  $b_i - \sum_{j \geq \underline{i} - 1, j \neq i} b_j / (n_{F_i} - 1) > (p - 1/2)$ . Similarly, if  $b_i < \mu_b$ , then  $F_i$  is the set of players  $\{1, 2, \dots, \bar{i}\}$  where  $\bar{i} \geq i$  is determined such that  $-b_i + \sum_{j \leq \bar{i}, j \neq i} b_j / (n_{F_i} - 1) \leq (p - 1/2)$  and  $-b_i + \sum_{j \leq \bar{i} + 1, j \neq i} b_j / (n_{F_i} - 1) > (p - 1/2)$ . For  $b_i = \mu_b$ , clearly  $F_i^* = \{1, \dots, n\}$ . From the definitions of  $\underline{i}$ ,  $\bar{i}$  and the single peakedness of the bias distribution, the result follows directly.  $\square$

## 12. Mediated talk

In this section we analyze to what extent an impartial mediators can improve communication in a given room. Instead of sending a message to all players in the room, player  $i$  sends a private message to mediator  $i$ . The mediator then makes an announcement that is heard by all players in the room. Ex ante the mediator commits to a strategy which maps from the messages he receives into the announcements he makes. This commitment is the key that allows us to improve communication.

Before going into the details, we want to make a few remarks about the setup. First, we start with a setting where each player has a separate mediator and each mediator receives only the message of one player. This allows us to use commitment to improve cheap talk. There is another version in which all players send messages to the same mediator and we will come back to this below. Second, our mediators send the same message to each player in the room. If we allowed the mediator to send private messages to each player in the room we would effectively destroy the room structure as mediators could effectively create subrooms and all kind of other network structures. (This is particularly true if we move to a setting with only one mediator.) To stay in line with our model we therefore assume that a mediator sends one message received by all players in the room. Furthermore, our mediators have no information apart from the message that they receives from the players. In particular, a mediator does not receive messages from players in other rooms. Again assuming anything else would effectively destroy the room structure.

By an argument akin to the revelation principle, we can focus on mediator strategies that induce the agents to truthfully reveal their signal. As the only thing the players are interested in (for choosing their actions) is the expected state  $\theta$ , it is also without loss to let mediator  $i$  announce a forecast for  $\theta_i$ . It is without loss of generality to restrict the mediator's strategy such that  $i$  believes the forecast for each  $\theta_j$  where  $j \neq i$ . By Bayesian rationality, the mediator's strategy has then to be such that the expected forecast has to equal the ex ante expected value of  $\theta$ . The question is whether mediation can lead to

more truthtelling. Note that for those player in a given room that are truthtelling without mediation their mediator can commit to truthtelling. (Recall that other players' strategies – and therefore also their mediators' strategies, are irrelevant for player  $i$ 's incentives to tell the truth.) We can consequently focus on those players who are babbling without mediation and start by giving an example where mediation improves truthtelling.

Let there be three players in the room with  $b_1 = -b$ ,  $b_2 = 0$  and  $b_3 = b$  for some  $b > 0$ . The truthtelling interval without mediation is then  $[-2(p - 1/2)/3, 2(p - 1/2)/3]$  and only includes player 2 if  $b > 2(p - 1/2)/3$  which will be assumed here. Now suppose mediator 1 commits to the following strategy: Whenever player 1 sends message  $h$ , mediator 1 will send message  $h$  but if player 1 sends message  $l$  the mediator will mix between  $l$  and  $h$  with probabilities  $\lambda$  and  $1 - \lambda$ . This implies that players  $-i$  know that  $i$ 's signal was  $l$  whenever the mediator sends message  $l$ , and therefore  $\mu_{ji}^l = 1 - p$ , but adopt belief  $\mu_{ji}^h = (1 - \lambda(1 - p))/(2 - \lambda)$  if they receive message  $h$  from the mediator. It is straightforward that  $i$  has an incentive to tell the truth to the mediator if his signal is  $l$ . Therefore, consider  $i$ 's incentives when his signal is  $h$ . With probability  $\lambda$  the mediator will send message  $l$  regardless of  $i$ 's message and this case is therefore irrelevant for comparing  $i$ 's truthtelling incentives. With probability  $1 - \lambda$   $i$ 's message decides whether  $\mu_{ji}$  is either  $\mu_{ji}^l$  or  $\mu_{ji}^h$ . Following 16 in the main text (proof of lemma 1),  $i$  will therefore have incentives to tell the truth to the mediator if and only if

$$-\frac{3/2 - \lambda + \lambda p - p}{2 - \lambda} - \frac{b}{2} - b + p \geq 0$$

which is equivalent to

$$\lambda \left( \frac{3}{2}b - (2p - 1) \right) \geq 3(b - (p - 1/2)).$$

If  $b \in (2(p - 1/2)/3, p - 1/2)$ , then there exist  $\lambda \in (0, 1)$  such that this inequality holds. This implies that the suggested mediation scheme can improve communication for players who are just outside the no-mediation truthtelling interval but not for players who are very far outside this interval. (Of course, the analysis for player 3 is analogous to player 1 in this example).

The strategy of the mediator in the example above is in fact the best the mediator can possibly do. Note that the problem of player 1 is not to truthfully report message  $l$  but to truthfully report message  $h$ . By mixing after message  $l$ , the mediator relaxes this truthtelling constraint in two ways: First, with some probability  $1 - \lambda$  the mediator sends message  $h$  regardless of  $i$ 's message. Second, the effect of the mediator sending message  $h$  is less problematic for player  $i$  as it leads to a belief  $\mu_{ji}^h$  below  $p$ . (Note that the belief  $\mu_{ji}^l$ , however is kept at  $1 - p$  and therefore as low as possible in order to make sending the low message after a high signal as unattractive as possible for player  $i$ .) However,



the mediator cannot arbitrarily lower  $\mu_{ji}^h$ : As the signal structure has to be consistent with Bayes' rule, the lowest possible belief  $\mu_{ji}^h$  is the prior  $1/2$ . If  $b$  is so high that  $i$  is not truthtelling for  $\mu_{ji}^l = 1 - p$  and  $\mu_{ji}^h = 1/2$ , then the mediator cannot improve the outcome. While this result was shown through an example, it should be clear that this holds more general. The bounds  $\mu_{ji}^h \geq 1/2$  and  $\mu_{ji}^l \geq 1 - p$  imply through equation 16 in the main text (proof of lemma 1) that truthtelling is impossible (unless  $\mu_{ji}^h = \mu_{ji}^l$  which is equivalent to babbling) after a high signal if

$$b_i < \frac{1/2 + 1 - p}{2} + \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} - p$$

and similarly the bounds  $\mu_{ji}^h \leq p$  and  $\mu_{ji}^l \leq 1/2$  imply that truthtelling is impossible after a low signal if

$$b_i > \frac{1/2 + p}{2} + \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} - (1 - p).$$

This result can be extended to the case where one mediator receives signals by all players and the publicly announces a forecast: For  $b_i - \left(\sum_{j \in R_i, j \neq i} b_j\right) / (n_{R_i} - 1)$  sufficiently high, the term in brackets in 16 in the main text (proof of lemma 1) will be strictly positive for all feasible values of  $\mu_{ji}^h$  and  $\mu_{ji}^l$  and therefore truthtelling after a low signal is infeasible unless  $i$ 's message does not affect the mediator's forecast. Similarly, truthtelling is impossible if  $b_i - \left(\sum_{j \in R_i, j \neq i} b_j\right) / (n_{R_i} - 1)$  is too low, i.e. too negative, as truthtelling after a high signal is impossible.

As  $b_i - \left(\sum_{j \in R_i, j \neq i} b_j\right) / (n_{R_i} - 1)$  scales in  $\eta$  if the set of biases is  $\mathcal{B}_\eta$ , theorem 2 and proposition 2 in the paper still hold if we consider mediated talk.

### 13. Alternative empirical approach

This section describes how one could use an approach similar to Barberá et al. (2015) to analyze data from Twitter in light of our model. This approach is not based on text analysis but on follower networks. Different from our approach in the main text, we will not exploit differences in the audience of reply tweets. The variation we use instead is that for some accounts ideology of the followers is closer to the ideology of the account owner than for others. Our model predicts that meaningful cheap talk communication is not possible on those accounts where audience ideology and sender ideology differ more. As a consequence, we expect that users whose audience is ideologically different from themselves are more likely to avoid cheap talk by verifying their messages. The empirical strategy is to search for hints of this verification in the form of hyperlinks. Put differently, our hypothesis is that senders whose audience differs ideologically from themselves are more likely to use hyperlinks in order to factually back up their messages.

As units of observation we chose Twitter accounts of members of Congress.<sup>16</sup> Similar to Barberá et al. (2015), we score them and their followers according to their follower networks. The method is described step by step here:

1. We downloaded the list of all followers of each member of congress. (Data collection took place between April 26 and May 21, 2019.)
2. For every account following at least three accounts of our members of Congress, we determine a score which equals the fraction of Republican accounts followed among all followed Congress accounts.
3. The *followerscore* of a member of Congress is the average score of its followers.
4. The *ownscore* of a member of Congress is determined as the fraction of Republican accounts followed among all Congress accounts followed (i.e. this score is calculated in the same way as the score for each following account).
5. We determined the absolute value of the difference between ownscore and follower-score (*absDiff*).
6. We downloaded the last 3000 Tweets of every member of Congress and determined the frequency with which hyperlinks were used.
7. We regressed the hyperlink frequency on *absDiff*.

The results of this regression are given in table 1. Clearly, a higher difference in ideological position is linked to a higher frequency of links used (significant at 1% level). This supports our hypothesis and the interpretation that verifiable information instead of cheap talk communication is used when ideological distance is large.

	Estimate	Std.Error	t value	Pr(> t )
(Intercept)	0.171569	0.0118451	14.4844	<1e-39
absDiff	0.17546	0.0589441	2.97672	0.0031

Table 1: Regression results: *absDiff* on hyperlink frequency

---

<sup>16</sup>We use the lists of twitter accounts that are provided on the twitter accounts of GOP and Democrats. Namely the list "SenateDemocrats" by "SenateDems", the list "House-Democrats" by "HouseDemocrats", the list "SenateRepublicans" by "SenateGOP" and the list "House-Republicans" by "HouseGOP". These contain the twitter handles of all current members of congress of the respective parties and a few other accounts clearly linked to the respective party (like, for example "BankingGOP" which is an account describing itself as "Senate Banking, Housing & Urban Affairs Republicans"). The important feature is that each account is clearly connected to one of the two parties.

## 14. Documentation of data collection

### 14.1. Tweets of members of Congress

#### 14.1.1. Collection

We use the lists of twitter handles that are provided on official twitter accounts of the Republican and Democratic delegations in the US Congress. Namely, these are the list "SenateDemocrats" by "SenateDems", the list "House-Democrats" by "HouseDemocrats", the list "SenateRepublicans" by "SenateGOP" and the list "House-Republicans" by "House-GOP". These contain the twitter handles of all current members of Congress that caucus with the respective party (insofar as they have a twitter account). We downloaded the lists for the 116th Congress on May 22nd, 2019 and the lists for the 117th Congress on February 26, 2021 using the static Twitter API. We then downloaded all tweets published by those twitter handles that were made on or after January 1, 2019, once again using the static Twitter API. The collection took place between February 26 and March 3, 2021.<sup>17</sup>

#### 14.1.2. Generating relative monogram/bigram frequencies

For the construction of relative word frequencies we discarded retweets and reply tweets and focused on original tweets published in the period January 1st 2019 to February 26th, 2021. (In the following, we will refer to tweets that are written by the user and are neither retweets nor reply tweets as "original tweets"). We pre-processed the tweets by replacing all links by the generic word "httplink" and then using the Julia package TextAnalysis.jl to strip the tweets of punctuation, numbers, cases, articles, prepositions, pronouns and additional whitespace. Following that we used the "Snowball" stemmer provided by the Julia package TextAnalysis.jl to stem all tweets. The same package generated a list of all monograms and all bigrams used.

We then compiled a dictionary of relative bigram frequencies and one dictionary of relative monogram frequency. For this dictionary, we only used bigrams/monograms that are (i) used at least 50 times and (ii) used at least 2 times as often by members of one party as by members of the other party. This produced to a dictionary containing 10818 entries and their respective usage numbers in the case of bigrams and 4255 entries in case of monograms.

### 14.2. Main sample and reply sample

In the period February 12 to February 19, 2021 (a period that included the end of the second impeachment of Donald Trump) we used the Twitter Streaming API to stream tweets published by users that had activated geolocation and were within the GPS-coordinates between 65°W and 125°W as well as between 26°N and 49°N (which is, roughly speaking,

---

<sup>17</sup>At the time of collection, the accounts of 11 Democratic and 15 Republican members of the 116th Congress were no longer available.

the continental US). From the obtained 789 688 tweets, we discarded all retweets and reply tweets as well as those identified to be coming from users in Canada or Mexico. We then selected all remaining original tweets that included the words “Trump”, “Biden” or “Congress”. 12 467 tweets remained. We used the accounts that published those tweets and downloaded for each of those accounts all their tweets published after January 1st, 2019, using the static Twitter API. We restricted this sample further by excluding all accounts that had fewer than 500 original tweets or fewer than 20 replies.

Of the remaining accounts in our sample, we know that they (i) tweet in the continental US (at least some of the time), (ii) have at least a passing interest in tweeting about political topics, (iii) have written a sufficiently high number of recent original tweets to allow us to estimate their political position, (iv) have engaged in a sufficiently high number of interactions (i.e. reply tweets) for us to examine their interactions. We call this data set our “main sample”. The accounts to which the accounts in our main sample have replied is the set of “reply accounts”.

We then downloaded (between March 19, 2021, and April 5, 2021) all original tweets published on or after January 1, 2019, by all reply-accounts. Again we restricted our data set to those reply-accounts with at least 500 original tweets and will call this data set the “reply sample” in the following.

### 14.3. Scoring

We scored the political stance of the accounts in the main sample first by using using the bigram dictionary described in 14.1.2: All original tweets of a user transformed to bigrams and those bigrams present in our frequency dictionary are used. The score of a user is then the average frequency among those bigrams, weighted by usage. More precisely,

$$score(text) = \frac{\sum_{bg \in text \cap dict} freq(bg)}{\sum_{bg \in text \cap dict} 1}$$

where  $text$  is a set of all bigrams constructed from the original tweets of the user,  $dict$  is the set of bigrams in the bigram dictionary constructed from the Members of Congress tweets,  $freq(bg)$  gives the frequency of Republican usage of  $bg$  in this dictionary.<sup>18</sup> We then use the monogram dictionary to produce a second score in the same manner and finally average the two scores to obtain our final score of a user’s ideological position.

In the same manner each reply tweet present in the main sample is scored using first the bigram frequency dictionary and then the monogram dictionary, i.e.  $text$  in the formula above is now a set of bigrams (respectively monograms) constructed from a single reply tweet. Here we removed each reply tweet from the data set that could not be scored, i.e. each tweet that either did not contain any bigram in contained in the bigram dictionary

---

<sup>18</sup>In case  $text \cap dict = \emptyset$ , the assigned score would have been 0.5. However, this did not occur for any user in our main sample.

or did not contain any monogram contained in the monogram dictionary.

Furthermore, each account to which one of our users from the main sample replied to is scored in the same way using its original tweets, i.e. here again *text* is the set of bigrams/monograms constructed from all original tweets of a user in the reply sample.

We then restricted our main sample slightly by (i) all reply tweets for which the user replied to could not be scored (i.e. because his original tweets did not contain a monogram and a bigram from our respective dictionaries) and (ii) eliminating all reply tweets by those senders for which all users to which they replied had the same score (which in practice means that these users only replied to one single account). The reason for the latter is that in these cases there is no within user variation that we could exploit. This leaves us with 149161 reply tweets sent from 2418 senders to 30,075 receivers.

### 14.3.1. Sentiment scores

Sentiment scores for each reply tweet were generated using the QDAP score of Hu and Liu (2004), implemented in the R-package “SentimentAnalysis”. We also created another sentiment score using the Python package “vaderSentiment” that implements the VADER score of Hutto and Gilbert (2014) in order to check all sentiment results for robustness.

### 14.3.2. Other variables of interest

variable	explanation
tweet length	length of unstemmed tweet after removing links (i.e. words starting with “http://”, “https://” or “www.”), hashtags (i.e. words starting with “#”) and directs (i.e. words starting with “@”), punctuation and additional whitespace
word length	number of characters in a tweet divided by the number of words; tweet is not stemmed but links (i.e. words starting with “http://”, “https://” or “www.”), hashtags (i.e. words starting with “#”) and directs (i.e. words starting with “@”), punctuation and additional whitespace have been removed
number of Hashtags	number of words starting with “#”
link frequency	share of words starting with “http://”, “https://” or “www.”
profanityCheck	probability that a reply tweet contains profanity as judged by the Python package profanity-check; see <a href="https://pypi.org/project/alt-profanity-check/">https://pypi.org/project/alt-profanity-check/</a> for details on profanity-check ( <i>Accessed: May 27, 2021</i> ).

## References

- Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26(10), 1531–1542.
- Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica* 50(6), 1431–1451.
- Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, Volume 4, pp. 755–760.
- Hutto, C. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 8, pp. 216–225.