

Why Echo Chambers are Useful*

Ole Jann

Nuffield College, University of Oxford

Christoph Schottmüller

University of Cologne and TILEC

October 2, 2018

Abstract

Why do people appear to forgo information by sorting into “echo chambers”? We construct a highly tractable multi-sender, multi-receiver cheap talk game in which players choose with whom to communicate. We show that segregation into small, homogeneous groups can improve everybody’s information and generate Pareto-improvements. Polarized preferences create a need for segregation; uncertainty about preferences magnifies this need. Using data from Twitter, we show several behavioral patterns that are consistent with the results of our model.

JEL: D72, D82 (Asymmetric Information), D83 (Learning, Communication), D85 (Network Formation and Analysis)

Keywords: asymmetric information, echo chambers, polarization, debate, cheap talk, information aggregation, Twitter

*Jann: Nuffield College and Department of Economics, University of Oxford; ole.jann@economics.ox.ac.uk. Schottmüller: Department of Economics, University of Cologne; c.schottmueller@uni-koeln.de. We are grateful for helpful comments by James Best, Ben Brooks, Vince Crawford, Marcelo Fernandez, Ben Golub, Sanjeev Goyal, Paul Klemperer, Meg Meyer, Bill Sandholm, Kyle Woodward and Peyton Young, as well as audiences at the universities of Konstanz, Oxford and Wisconsin-Madison and at EEA 2018 (Cologne) and TTW 2018 (Northwestern).

Large parts of society are organized around the (non-market) exchange of information and opinions: People gather around breakfast and dinner tables, in meeting rooms and committees, cafés and bars, while keeping in permanent touch with friends, co-workers and strangers through electronic messaging and social media. But while people constantly seek out others’ views and knowledge, they do not seek out a wide range of different viewpoints. Instead, they segregate into homogeneous communities and limit the number of views they are exposed to.¹

This poses a theoretical puzzle: If people put so much energy into seeking and exchanging information, why do they artificially limit both the diversity and the amount of information available to themselves? It also poses a practical problem for society: The segregation into “echo chambers” has widely been decried as being responsible for recent populist insurgencies in the Western world.²

In this paper, we develop a general model of how people with different preferences and different information rationally communicate in groups, and how they sort into groups while anticipating what communication within the group will be like. Our analysis reveals that segregation into small, homogeneous groups can be a rational choice that *maximizes* the amount of information available to an individual. In fact, homophilic segregation can be efficient and even Pareto-optimal for society.

Why is that? Our argument builds on the idea that people have not only different information, but also different preferences. These differences in preferences can prevent successful communication, because people do not want to reveal their information to those who are different, and distrust the motives of those who speak to them. It then becomes easier to exchange information in segregated, homogeneous cliques than in large crowds. Echo chambers, though they may cut off potential communication with a great number of people, make actual communication possible, and are hence useful for society.

The activity of sorting into groups and communicating within them is highly complex and does not easily lend itself to strategic analysis. Every speaker chooses his message based on how he thinks different messages will be perceived – which, in turn, depends on how the listeners expect a speaker to choose his message and on what other knowledge the listeners have, which in turn may depend on who else is speaking to them and what their messages are. And all these implications have to be considered when deciding which group to interact with (which table to join, which room to enter). In sections 1 to 5, we develop a highly tractable way to model strategic information transmission (i.e. cheap talk) among many individuals, who all have different information, who all have the ability to send and receive messages, and who all freely choose within which group of people they

¹See, for example, studies on segregation in blogs (Lawrence et al., 2010), on Facebook (Del Vicario et al., 2016; Quattrociocchi et al., 2016), on Twitter (Barberá et al., 2015) and in online and offline contexts in general (Gentzkow and Shapiro, 2011).

²See, for example, articles on the role of echo chambers in the “Brexit” referendum (Chater, 2016) or the rise of Donald Trump (Hooton, 2016).

wish to communicate.

In section 6, we use empirical evidence from the micro-blogging service Twitter to examine our predictions that the exchange of information becomes harder when people have different political views, and that people interact more with others who hold similar political views. In section 7, we discuss several applications of our model to the online and offline world. Of course, we do not claim that segregation is always beneficial or that echo chambers are necessarily a good way to organize society. Our results, however, not only point to the benefits of echo chambers; they also throw doubt on some common arguments against them. We discuss these in section 7.4.

Theoretical Results We analyze a general model in which a number of individuals face aggregate uncertainty and have different preferences. These individuals sort into groups, communicate within these groups, and finally make a choice. Every person wants to make a choice that reflects his own knowledge and preferences, and would want everybody else to take that choice as well. Consider the following example: A group of voters has to decide on a level of taxation and redistribution. There are two sources of disagreement: Knowledge and preferences. People disagree over how bad taxation is for economic growth, i.e. they have different information about the state of the world. But they also have different preferences: Even if they all agreed on the state of the world, rich people would still prefer lower taxes than poor people. Everybody has a preferred level of taxation, and would prefer everybody else to vote for that level of taxation as well – whereas others, of course, may prefer a different level and would in turn want everyone to vote for *their* preferred level.

We assume (for now) that people’s preferences are common knowledge, while their information is private and cannot verifiably be communicated. Before voting for a level of taxation, people can communicate their information about the harmfulness of taxes. But differences in preferences interfere with the exchange of information: If a rich man says that taxes are harmful, is that because he really thinks so, or because he is trying to fool people into voting for lower taxes, which benefits him personally? It depends on his audience: If speaking to a group of other rich people, he wants to give them accurate information, given that they will then vote for a tax policy that is close to what he prefers. If he speaks to a group of paupers instead (who are inclined to vote for what he views as “too high” taxation), he will try to convince them that taxes aren’t hurtful, and the paupers hence have no reason to pay any attention.

What if he speaks to a mixed audience, or one that includes members from even more groups? What if those other people also speak simultaneously, possibly submitting information to the speaker and the rest of the audience? We model such debate as a multiple-sender, multiple-receiver cheap talk game in which each player has information and simultaneously sends and receives messages. Crucially, we assume that every player’s

information is independent from that of others. We consider this a realistic model of debate: Every person knows some aspect of a problem, and a combination all knowledge gives a faithful picture of the world. But when deciding which information to reveal, players focus on the preference problem (“Do I want to reveal this information?”) and not on problems of higher-order knowledge (“Does he already know what I am telling him?”). This modeling technique allows us to derive an intuitive, geometric solution (theorem 1) to the n -person cheap talk game: Whether someone tells the truth in equilibrium depends only on the distance between the speaker’s preference parameter and the average preferences of his audience.³

With this understanding of communication within arbitrary groups, we can turn to the question of how people rationally sort into groups, and when such equilibrium sorting is optimal. We assume that before any communication takes place, people can enter one of many “rooms”. Each message is heard by everyone within the same room but cannot be heard outside the room. Entering or leaving a room can have many effects: disciplining those whose preferences are close to one’s own (making them more willing to tell the truth), destroying truth-telling between people which otherwise existed, providing information to others in the room (if the entrant tells the truth), giving more information to the entrant (if he comes from a room in which he learned less) – or any combination of these. Since every player cares about his own information (to make a precise choice) and that of others (because he cares about their choices, which they make based on their information), the analysis may at first seem to be quite complex.

We show, however, that all of these considerations simplify to one: In choosing a room, a player wants to maximize the weighted sum of pieces of information that is generated by subsequent communication in all rooms (proposition 1). A “piece of information”, in this context, is simply the fact that the information of one player is available to another player. In our set-up, we can measure this information generation in bits, the basic unit of information. The only differences in motivation between players arise because they each value their own information more than that of others.

Our analysis focuses on two closely related questions: What is the welfare-maximizing allocation of people into rooms, and which room allocations can emerge as equilibria from individual behavior? The simplest way to think about these two problems is to consider a polarized society that consists of two groups of players who differ in their preferences. In section 4.4, we characterize the optimal room allocation and, when it is not an equilibrium, the welfare-optimal equilibrium of the room-choice game.

More generally, we think of polarization as “clustering” of preferences around certain values. We parameterize this notion of polarization while keeping the differences in information between players constant. This way, we can show that if the polarization of

³In the supplementary material, we show that our main arguments are robust to using different assumptions and modeling techniques.

preferences is large compared to the differences in information, full segregation by preferences is always welfare-optimal and an equilibrium, whereas integration is optimal and an equilibrium for low polarization (theorem 2). We provide a lower bound for how polarized a society would need to be before segregation becomes both optimal and an equilibrium phenomenon, by showing that if preferences are evenly distributed, segregation is neither optimal nor an equilibrium (section 4.3).

In our example of choosing a tax policy, this may mean that society optimally splits into two political parties: One bringing together the rich, the other the poor. Within each party, members can truthfully discuss their thoughts and knowledge on how the world works – while a meaningful discussion involving members of both parties would be impossible. Depending on how preferences are distributed, other outcomes are possible. Fully integrated debate may be feasible and optimal if there is no polarization in preferences. If there is stronger polarization, society could fragment into even more parties. (Of course, the assumption that preferences are simply a reflection of wealth is highly stylized. The argument would equally apply to any other polarization in preferences, as long as people were to disagree about what was the right thing to do even if they could agree on the particular fact they are currently discussing.)

Overall, our results suggest that segregation into homogeneous “echo chambers” is a rational and often Pareto-optimal response to polarized preferences. Segregation is caused by polarization, not the other way around. However, these results do not mean that polarization is in any way good for society – in fact, we can show that polarization lowers welfare (proposition 3). Segregation, as a rational response to polarization, mitigates the corrosive effects of polarization, and can hence be seen as an indicator of polarization as well as a countermeasure of society against it.

In section 5, we consider what happens if preferences as well as information are private. This hinders communication even more, since players are more skeptical and scrutinize each message both for information about the state and the sender’s motivation. This leads to even more segregation in equilibrium and welfare optimum. Indeed, we find that if we start with any situation in which full integration is optimal and an equilibrium, and increase uncertainty in preferences (while keeping expected preferences constant), we can reach a situation in which full segregation by preference types is the only equilibrium and also welfare-optimal. We suggest that this is relevant for thinking about interactions and debates on the Internet, where the precise type of one’s conversational partner as well as audience is often unclear. Under such uncertainty, there is a stronger need for segregation than there may be with off-line interactions, and hence further reduced welfare.

Empirical Evidence In the last part of our paper, we provide evidence for some of the results and predictions of our theoretical considerations. Since we understand our results to be at a high level of abstraction, we do not think that all of them can directly

be translated into measurable behavior, or that one could even try to estimate model parameters. Instead, we consider several behavioral patterns that would be consistent with the mechanisms of our model, and show that these patterns are present in observed behavior on a large social media platform.

On the online messaging and networking platform Twitter, users can send different kinds of messages (“tweets”) which are seen by different kinds of audiences. We develop a novel way to estimate the ideological stance of Twitter users based in the United States, by measuring how similar their tweets are to current members of the U.S. Congress. With this tool, we can examine how the nature of interactions on Twitter changes with the ideological distance between participants – where we interpret a user’s ideological stance as his “bias”.

The main mechanism in our model is that when people have very different preferences, they find it hard to exchange credible information via cheap talk. In the model, this shows itself in the fact that only babbling is possible. It is not clear, however, what such babbling would mean in practice: The same message can be meaningful and informative or completely meaningless, depending on the sender’s intention and the receiver’s expectation. But while it may seem futile to try to observe babbling directly, we believe that the *consequences* of babbling are more easily spotted.

How should we expect people to behave if cheap talk is indeed impossible because of a large ideological distance? We see three possible consequences: (i) Not to send a message at all, since there is little to be gained. (ii) Sending a short, emotional (and potentially abusive) message to satisfy an emotional need, not to transmit any information. (iii) Trying to persuade anyway – not by cheap talk, but by arguments and verifiable information.

In section 6, we present evidence that is consistent with all three effects. Twitter users engage more with people who have similar ideology than with people who are different. The larger the ideological distance between two twitter users, the more emotional and negative interactions are – an effect that is much stronger for short tweets than for longer ones. And overall, as the ideological distance between twitter users grows, we see more long and complex tweets that make use of hyperlinks to outside sources.

Relation to other research Our work closely relates to four different methodological approaches, and ties into a wider-ranging literature on segregation, isolation and echo chambers.

In methodological terms, we develop a highly tractable model of many-to-many cheap-talk. Our simple geometrical solution avoids much of the exponential complexity that usually appears in models with multiple senders or receivers. As such, our model can reproduce and simplify some insights from other multi-sender or multi-receiver models. For example, similarly to the classical analysis by Farrell and Gibbons (1989), the presence

of other receivers may either discipline the sender or subvert truth-telling. In contrast to most other papers, we allow for an arbitrary number of agents who are both receivers and senders and add a first stage in which agents decide whom to communicate with.⁴ In our main analysis, we restrict ourselves to binary signals and messages, but show in the supplementary material that our main results are robust to the introduction of an arbitrary finite number of states and signals.

While the rooms of our analysis are a novel modeling device, they can in principle be thought of as fully connected, disjoint networks. A related paper by Galeotti et al. (2013) analyses communication in networks by agents who face a decision problem similar to ours, but in their setup the most informative (or welfare optimal) equilibrium can be in mixed strategies. Such mixed equilibria are a common occurrence in similar models but are generally intractable. In our model, however, the most informative equilibrium is always in pure strategies. There is, of course, a much larger literature on endogenous network formation. The principal differences to our paper are that we consider cheap talk, do not focus on directed networks, and construct a tractable model of room choice, which allows us to study (efficient) segregation.

The welfare analysis of room choice in our model can also be seen as an information design problem: How can an information designer induce information exchange between several agents, if these agents have an incentive to manipulate others through lies, and if commitment to a disclosure rule (as in the literature on Bayesian Persuasion) is not available? The right construction of mixed groups can induce truth-telling. Rooms endogenously create costs to lying (the main instrument of discipline in Kartik 2009), and they induce truth-telling despite the fact that different senders' information is orthogonal to each other and there hence exists no mechanism (as in e.g. Krishna and Morgan 2001) to elicit information by playing senders off against each other.

We also show that uncertainty about preferences has a corrosive effect on truth-telling. This is similar to Morgan and Stocken (2003), who consider financial analysts who are biased in a known direction, but whose precise bias is unknown. Such uncertainty "in one direction" leads to losses in informativeness in one direction (i.e. one of two messages becomes more common but less informative). Our analysis extends to general distributions of players' biases and hence considers uncertainty about the size and the sign of the sender's bias, which may be continuously or discretely distributed. What turns out to matter is the concentration of probability mass around certain values, and hence we can show that uncertainty about size and direction of a bias does not necessarily help with information transmission (as it does in Li and Madarász, 2008). Our results and methods generalize without loss to large groups of players and general distributions of biases. Of

⁴While our novel setup allows us to vastly simplify the analysis of many-to-many cheap talk, our main arguments are not dependent on this particular setup and can be derived in a more classical cheap-talk setting akin to Crawford and Sobel (1982), as we show in the supplementary material.

course, we are mostly interested in these results as a preliminary for room choice, as rooms are optimally and in equilibrium more segregated for higher uncertainty. To our knowledge, we are the first to generally analyze how uncertainty about bias influences whom people want to associate and communicate with, and how it increases the appeal and the usefulness of segregation.

Finally, in our empirical work, we develop a novel way to score Twitter users on a partisan left-to-right scale, based only on their tweets. The method is similar to how Gentzkow and Shapiro (2010) score newspaper editorials; we demonstrate that such a method is valid for scoring arbitrary Twitter users. The main differences from this earlier work are in the size of our partisan dictionary (which is about 18 times the size of Gentzkow and Shapiro’s dictionary) and the causal agnosticism with which it is compiled: While earlier works have focused on phrases with clear ideological content, our dictionary also contains non-obvious (but informative) entries such as hashtags, names and locations.

The debate about echo chambers has recently been given urgency by several studies and popular treatises on how the internet changes the way societies debate. Sunstein (2001, 2017) prominently makes the case that the internet has been increasing ideological segregation and that this endangers democracy. Gentzkow and Shapiro (2011) point out, however, that the segregation of “offline” interactions is larger than that of “online” interactions. But while such offline segregation can happen simply because we live close to people who are like us in many socio-economic aspects, segregation on the internet is driven more by choice. Lawrence et al. (2010), for example, show that blog readers tend to read blogs that agree with their own ideological bias. Our model allows us to analyze the informational effects of any kind of segregation or integration, as well as predicting which communication structures arise from individual optimizing behavior, and whether they are socially optimal. Most importantly, we argue that those who see in segregation the ruin of societies are focusing on a symptom, not the cause. Polarization of preferences and mutual mistrust are the real culprits; informational segregation is a rational behavior that mitigates the harm they do.

1. Model

There is an unknown state of the world $\theta = \sum_{k=1}^n \theta_k$. Each θ_i is independently drawn to be 0 or 1 with equal probabilities, so that θ is binomially distributed on $\{0, 1, \dots, n\}$. n individuals each make an observation about the state. In particular, individual i receives a private signal $\sigma_i \in \{\sigma^l, \sigma^h\}$ of accuracy p about θ_i , i.e. $Pr(\sigma_i = \sigma^h | \theta_i = 1) = Pr(\sigma_i = \sigma^l | \theta_i = 0) = p > 1/2$. Before observing his signal, a player can access one of n “rooms”. There are no costs to entering a room, and rooms have no capacity constraints – but each player can only be in exactly one room. After observing his signal, a player sends a cheap-talk message $m_i \in \{m^l, m^h\}$ that is received by all players in the same

room. Finally, each player takes an action a_i .

The payoff of player i is

$$\begin{aligned} u_i(a, b_i, \theta) &= -(a_i - b_i - \theta)^2 - \alpha \sum_{j \neq i} (a_j - b_i - \theta)^2 \\ &= - \left(a_i - b_i - \sum_{k=1}^n \theta_k \right)^2 - \alpha \sum_{j \neq i} \left(a_j - b_i - \sum_{k=1}^n \theta_k \right)^2 \end{aligned} \quad (1)$$

where a denotes the vector of actions of all players and $b_i \in \mathbb{R}$ is a commonly known “bias” of player i . That is, actions of all players affect i ’s payoff, and i would like that all players choose the action $b_i + \theta$. We can hence think of b_i as the *preferences* of the players, whereas θ_k is the *information* of player k . Note that only the relative positions of biases matters (i.e. the distances between biases), not their absolute magnitude. The parameter α measures the relative weight players assign to other players’ behavior. Players maximize their expected payoff.

The timing of the game is:

1. Players simultaneously decide which room to enter.
2. Players privately observe their signals σ_i , and room choices become common knowledge. Players simultaneously send messages m_i that are observable by everyone in the same room R_i .
3. Players simultaneously take actions a_i ; payoffs are realized.

We analyze the model by backwards induction: First we characterize optimal choice of action given messages, then the optimal choice of message given a room allocation, and then we analyze the game in which players choose which room to enter. The solution concept used throughout is Perfect Bayesian Equilibrium.⁵

2. Equilibrium Behavior Within a Room

2.1. Choice of Action

We can immediately see that only the first part of expression 1 matters for determining i ’s optimal action a_i^* . The first-order condition yields

$$a_i^* = b_i + \mathbb{E}[\theta] = b_i + \sum_{j=1}^n \mathbb{E}[\theta_j], \quad (2)$$

i.e. the optimal action is simply i ’s bias plus his expectation of the state, conditional on his own signal and on the messages he has received.

⁵All messages occur in equilibrium and there is no hidden information at the time that people choose rooms, so that our results are insensitive to assumptions about off-path beliefs.

In the following, we will denote by $\mu_{ij} = \mathbb{E}_i[\theta_j]$ i 's belief about θ_j , so that expression (2) becomes $a_i^* = b_i + \sum_{j=1}^n \mu_{ij}$.

2.2. Choice of Message

Now that we have established each agent's optimal action choice given beliefs, we can consider the optimal choice of message. For this, we focus on a single room, and consider the equilibria of the cheap talk game in this room. This means that when we speak of "equilibrium" in this section, we mean the equilibrium in a specific room (with a given set of members with given biases), and not the overall equilibrium of the game. We can do this because once players have sorted into rooms, the messages in other rooms are unobservable and the actions of players in other rooms are irrelevant to a player's optimization problem. Hence, an equilibrium of the subgame after room choice can be disassembled into one equilibrium of the cheap talk game for each room.

Definition 1. *We call a messaging strategy m_i ...*

- babbling if m_i is independent of i 's observed signal σ_i and therefore nobody learns anything from m_i .
- truthful if $m_i(\sigma^l) = m^l$ and $m_i(\sigma^h) = m^h$.
- lying if $m_i(\sigma^h) = m^l$ or $m_i(\sigma^l) = m^h$.
- pure if m_i is either babbling or truthful.
- mixed if for some signal σ^k , $k \in \{l, h\}$, both messages are sent in equilibrium and the strategy is not babbling.

The cheap talk game within a room can – as usual – have several equilibria. For each player i , there always exists an equilibrium in which i babbles. (Consequently, there also always exists an equilibrium in which all players babble.) In line with the cheap talk literature, we will focus on the most informative equilibrium.⁶ The following lemma implies that the most informative equilibrium is in pure strategies.

Lemma 1. *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 on page 36.)*

What is the intuition for this result? Imagine an equilibrium in which player i mixes between messages after observing signal σ^h . That is, i is indifferent between sending a high message that induces high actions by the other players in his room and a low message

⁶The concept of "most informative" equilibrium is not necessarily well defined in multi-sender cheap talk games. However, the following paragraphs will make clear that this concept is straightforward in our model.

that induces lower actions by the players in his room. This means that the low actions induced by m^l are somewhat too low from i 's point of view and the high actions induced by m^h are somewhat too high from i 's point of view. Note that i will always send the low message in case he observes a low signal in such an equilibrium because the actions i would like the other players to take are increasing in his signal. Consequently, a high message perfectly reveals i 's high signal. Now consider switching to an equilibrium in which i uses the truthful strategy. When i now observes a high signal, sending the high message will lead to exactly the same actions by the other players as in the original equilibrium. However, sending a low message will lead to a lower belief than in the original equilibrium and therefore to lower actions by the other players. Player i will then strictly prefer the high message as these lower actions are too low (given that i was indifferent in the original equilibrium).

The main implication of lemma 1 is that the most informative equilibrium is always in pure strategies: Starting from any mixed equilibrium we can switch the mixing players one by one to truthful – and therefore more informative – strategies and the resulting strategy profile remains an equilibrium.

Corollary 1. *The most informative equilibrium in a room is always in pure strategies.*

We can now characterize the most informative equilibrium. Intuitively, we might expect that the distance of b_i to the biases of the other players is crucial for i 's incentive to tell the truth, since i becomes more interested in misleading the other players if their biases differ by a lot. We formalize this intuition and specify the most informative equilibrium in the following result, which is illustrated by figure 1:

Theorem 1. *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*

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

and babbles otherwise. (Proof on page 37.)

The size of the truth-telling interval increases in both n_R , the number of people in the room, and p , the precision of individual signals. The increase in n_R can be seen as a correction term: What really matters for the motivation of a player is his distance from the average bias of the *other* players in the room. Hence, if we write a symmetric interval around \bar{b} (which includes b_i), we have to add this correction.⁷ When p , the precision of signals, is higher, each truthful signal causes a greater change in the actions of others. People communicate truthfully if they are disciplined by the danger of influencing others'

⁷Intuitively, one could also think that the average room bias “stabilizes” for larger n_R , so that a player can be further away from the average room bias and have the same distance from the average bias of other players in the room.

actions too much by lying. Hence, if p is higher, this disciplining force is stronger and a player can be further away from the average bias of others and still tell the truth.

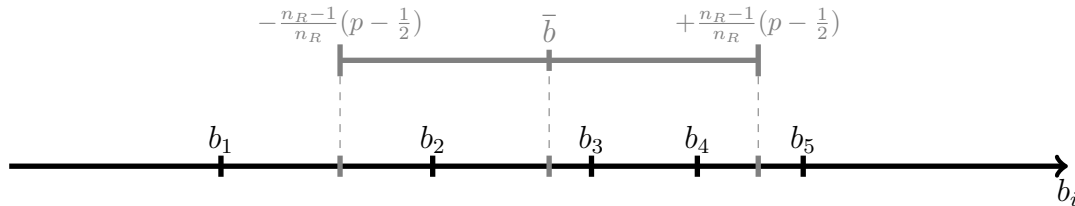


Figure 1: Finding the most informative equilibrium in a room consisting of players 1 to 5. We find the average bias and construct a symmetric interval around it. Players 1 and 5 babble in the most informative equilibrium, since their biases are too far from \bar{b} . Players 2, 3 and 4 tell the truth.

3. Room Choice

We can now analyze room choice, under the assumption that the most informative equilibrium will be played in any room (i.e. in each subgame). We will first derive some results about the welfare-optimal room allocation, and then analyze under which conditions this optimal room allocation is in fact an equilibrium.

3.1. Welfare-Optimal Room Allocation

Given the expression for individual payoff (1), overall welfare in the model is given by

$$W(a, b, \theta) = \sum_{i=1}^n u_i(a, b_i, \theta) = - \sum_{i=1}^n \left(a_i - b_i - \sum_{k=1}^n \theta_k \right)^2 - \alpha \sum_{i=1}^n \sum_{j \neq i} \left(a_j - b_i - \sum_{k=1}^n \theta_k \right)^2.$$

This expression, of course, is not yet very helpful in trying to compare different room allocations. However, we can show that in our model, welfare can simply be expressed in terms of the aggregate amount of information that is held by all players after communication has taken place.

Consider the information that is available to a single player. A player always receives his own signal σ_i . We can call this *one piece of information*. Assume that i also receives truthful signals from two other players; then we can say that i has three pieces of information about θ . Let $\zeta_i \in \{1, 2, \dots, n\}$ be the number of pieces of information available to player i which are either his own signal or truthful messages from other players. Given that each σ_j has two possible values (high or low), ζ_i in fact measures player i 's information in *bits*, the unit of information. The following result shows that all welfare comparisons reduce to informational accounting in bits:

Proposition 1. *Welfare depends only on $\sum_i \zeta_i$ and model parameters and is linearly increasing in $\sum_i \zeta_i$. (Proof on page 38.)*

Because payoffs are quadratic, we can additively separate a player's payoff into (i) losses through preference differences and (ii) losses from variance due to lack of information. In an equilibrium of the messaging game, the former losses are unavoidable, but the latter can be mitigated by increasing the flow of information between players. We can measure this flow of information simply by counting the pieces of information that each player has when making their decision. Since every player i has exclusive knowledge about θ_i , there are no decreasing marginal returns to information, and the sum of all ζ_i is indeed a sufficient statistic for welfare.

This result means that we can quickly compare any two room allocations. Consider, for example, the room allocation in figure 1. Having everybody in the same room generates 17 pieces of information: 3 players have 3 pieces of information each, while two players (those who babble) have 4 pieces each. Would it be possible to improve on this allocation? We can immediately see that this cannot be achieved by splitting players up into two rooms with 3 and 2 players, respectively: Even if everybody in these rooms was telling the truth, only $3^2 + 2^2 = 13$ pieces of information would be produced. The same is true for splitting them into a higher number of even smaller rooms. But even if we somehow could get 4 people in one room to tell the truth by putting one of the players into a separate room, the total number of pieces of information would be $4^2 + 1 = 17$ – the same as with full integration. Hence the room allocation shown in the figure is welfare-optimal.

Of course, we may often not be able to make such quick deductions in general cases and might have to consider many possible room allocations before concluding what the optimal one is. This problem gets more complex as n grows, since the number of possible partitions of a set (given by the Bell sequence) grows quite rapidly in the size of the set. We will derive general results on optimal allocations in section 4 below.

3.2. When is the Welfare-Optimal Allocation an Equilibrium?

Finding out which room allocation is welfare-optimal is, of course, only half the story. Since people are allowed to freely choose which room they want to be in, we now need to consider which room allocations can arise in equilibrium.

We begin by rewriting player i 's payoff analogously to our result on welfare (proposition 1; the derivation is in the proof of that proposition):

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

This redefines i 's choice of room in purely informational terms: When choosing a room, i wishes to maximize a weighted sum of his own information (after communication) and that of other players. When he considers switching from, say, room R_A to R_B , i will consider how much more he can learn in room R_B , as well as how much more or less the other people in both rooms will learn after his switch. How exactly i is willing to trade

off these informational effects against each other depends on α . For $\alpha = 1$, each agent simply maximizes welfare, and we can hence derive the following result:

Proposition 2. *For any configuration of biases, there exist $\underline{\alpha}$ and $\bar{\alpha}$ such that $\underline{\alpha} \leq 1 \leq \bar{\alpha}$ and a welfare-optimal room allocation is also an equilibrium of the room choice game if $\alpha \in [\underline{\alpha}, \bar{\alpha}]$. (Proof on page 40.)*

Depending on the deviations that are possible in the welfare-optimal allocation, $\underline{\alpha}$ and $\bar{\alpha}$ will often be strictly below and above 1. We can intuitively analyze the situation by considering figure 2, which shows all possible deviations for all players in the welfare-optimal room allocation of an exemplified game. Deviations in the shaded area cannot exist if players are allocated to rooms in a welfare-optimal way. The condition $\Delta\zeta_i + \alpha\Delta\sum_{j\neq i}\zeta_j \leq 0$ translates to $\Delta\sum_{j\neq i}\zeta_j \leq -\frac{1}{\alpha}\Delta\zeta_i$: Any line through the origin that has no points (i.e. deviations) above it corresponds to an α for which the welfare-optimum is also an equilibrium. In the given example, this is true for all $\alpha \in [0.75, 1]$.

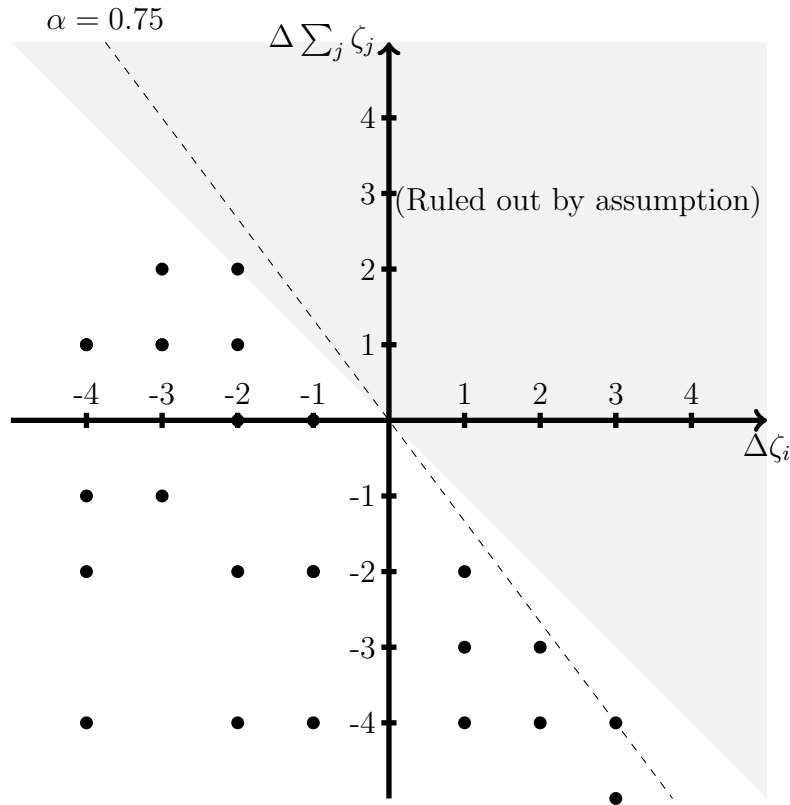


Figure 2: Possible deviations for all players in the welfare optimal room allocation. Deviations in the shaded area cannot exist in the welfare optimum. Any α for which $\Delta\sum_{j\neq i}\zeta_j \leq -\frac{1}{\alpha}\Delta\zeta_i$ means that the welfare-optimal allocation is also an equilibrium of the room choice game.

4. Polarization and Segregation

We have now shown that the messaging problem inside each room has a simple geometrical interpretation, and that the room choice game reduces to a problem in which all players wish to reduce a weighted sum of their own uncertainty and that of the other players. Using these results, we can analyze how the composition of the set of biases influences which room allocations are optimal, and which allocations can be achieved in equilibrium.

Despite the complexity of the model and the discrete nature of optimal communication and room choice, the tools and simplifications we have derived in the preceding sections allow us to gain general insights. The main insight is that segregation is welfare-optimal and an equilibrium if players are sufficiently polarized. By polarization, we mean that biases are clustered in two or more groups, instead of being clustered in one group or evenly spread out.

We approach polarization in three ways. First, we will consider arbitrary bias configurations and either increase or decrease the polarization of preferences compared to the differences in information between players. We show that total segregation is the welfare-optimal room allocation and an equilibrium if the polarization of preferences is large compared to the differences in information between players. On the opposite end of the spectrum, full integration is welfare-optimal and the unique equilibrium of the room choice game if the differences in information are large compared to the polarization of preferences.

Second, we establish a lower bound for how polarized biases need to be for segregation to be optimal. We show that if biases are evenly spread on some interval of the real line, then full integration (or a very similar room allocation) is the welfare-optimal room allocation and an equilibrium.

Third, we narrow our analysis to the case in which there are only two bias types, and fully characterize the set of all optimal allocations for different distances between the two biases as well as different proportions between the two groups, and whether these allocations are equilibria. In this case, we can generally show that if the welfare-optimal room allocation is not an equilibrium, then the welfare-optimal equilibrium allocation will have *too little* segregation compared to the welfare-optimum.

We will begin with a simple graphical example of how segregation can be a welfare-optimal equilibrium if players are polarized, and then introduce our general results.

4.1. A Simple Example

Consider a set of biases as in panel (i) of figure 3: A group of 6 players, 3 of whom have relatively small biases, while the other 3 have relatively large biases. If all players are within the same room (panel i), the truth-telling interval within this fully integrated room does not cover any of the players' biases, which means that in the most informative

equilibrium none of them reveals any information. The number of pieces of information generated is hence 6.

Suppose the players segregate by bias type into two separate rooms – see panel (ii). The truth-telling interval in both rooms covers all the players in the respective rooms, which means that all players reveal their information truthfully. In each room, 9 pieces of information are generated, which means that overall this allocation generates 18 pieces of information.

Is this segregation an equilibrium? We can consider the most profitable deviation of player 3 (which is symmetric to the most profitable deviation of player 4 and better than the best deviations of any other players) – see panel (iii). If player 3 moves into the other room, he will move the average in this room so that players 5 and 6 no longer tell the truth in any equilibrium. He himself also does not tell the truth anymore, so that his move completely deprives society of the information of players 3, 5 and 6. (The lengthening of the truth-telling interval that results from 3’s move is not enough to compensate for the change in average bias.) The resulting room allocation generates $2^2 + 4 + 3 = 11$ pieces of information, which clearly leads to lower welfare. But it is also inferior for player 3, since he now has 2 pieces of information (his own and the message from player 4) instead of 3, so that his payoff definitely decreases. Hence this deviation is not optimal for player 3, and no player has a profitable deviation from two segregated rooms – which means that this allocation is not only welfare-optimal, but also an equilibrium.

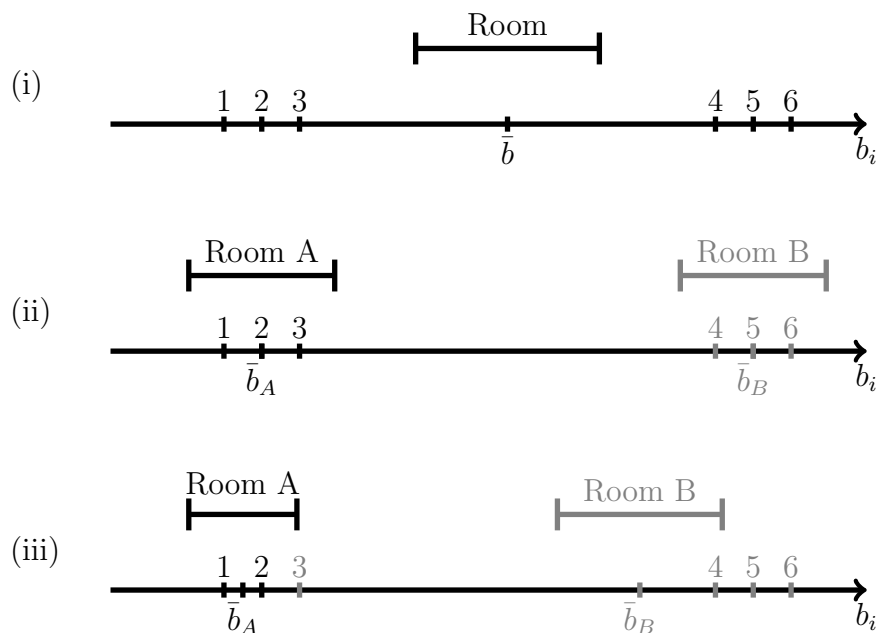


Figure 3: Truth-telling intervals for (i) the fully integrated room, (ii) two segregated rooms, (iii) player 3’s best deviation from the segregated room.



Figure 4: Welfare-optimal allocations that are also equilibria for large and small η .

4.2. When is Segregation optimal?

Consider any bias configuration \mathcal{B} , where an arbitrary number of players share each bias type. Let \mathcal{B}_η be the bias configuration that contains ηb_i for each b_i in \mathcal{B} . Then the following is true:

Theorem 2. (i) *If η is sufficiently close to 0, full integration is welfare-optimal and a room-choice equilibrium.*

(ii) *If η is sufficiently large, full segregation by bias types is generically welfare-optimal and a room-choice equilibrium. (Proof on page 40.)*

Figure 4 summarizes the result. We can intuitively explain it in the following way: If biases are clustered very closely compared to how different the players' information is, having all players in one room would result in universal truth-telling. This cannot be improved upon in welfare terms, and it is also an equilibrium since any player would lose by leaving the fully integrated room.

On the opposite end of the spectrum, we consider the case where biases are clustered very widely compared to differences in information, and we do not assume special, non-generic properties such as that one bias is the exact average of two other biases. Then truth-telling will be impossible in any room that contains two or more players with different biases. Hence there exists no room allocation that can improve welfare compared to full segregation by bias types. Similarly, no player has an incentive to deviate from full segregation, since such a deviation cannot provide more information to the player himself or any other player.

4.3. A Lower Bound for Polarization

To find out how polarized biases need to be so that segregation is optimal and an equilibrium, we can consider the stylized case of biases that are evenly 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.⁸ What is the welfare-optimal allocation in this case? Given that biases could be evenly distributed on a very large interval, we might expect that the optimal allocation involves separating

⁸If biases are tightly clustered around a central value, we could think of this as “negative polarization” – we consider this case in the supplementary material.

the players into several rooms. But in fact, the opposite is true: If biases are evenly distributed, the welfare-optimal allocation often involves a single integrated room that is also an equilibrium. In some special cases, an allocation that places one player outside the room is welfare-optimal, and this allocation is an equilibrium if α is large enough. We state this here as an informal result and refer the reader to the supplementary material for a more formal statement and derivation.

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.*

Why is this? 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.

4.4. Bipolar Polarization

We now focus on the case where there are two bias groups, i.e. $b_i \in \{0, b\}$ for some $b > 0$. This “bipolar polarization” is often used synonymously with the word polarization. Our results allow us to generally solve this setting for all possible parameter values. We will first heuristically derive solutions for the case where both groups have equal size and then comment on the case with unequal sizes. Detailed derivations are presented in the supplementary material.

If all players are in one room, the average bias will be $b/2$ and all players send truthful messages if and only if $b/(p - 1/2) \leq 2(n - 1)/n$, see theorem 1. Clearly, if this inequality holds, such a fully integrated room will then be both welfare optimal and an equilibrium. At the other extreme, consider the case where the presence of one player of bias b in a room containing all players with bias 0 will lead to babbling by all players. The average bias in such a room is $b/(n/2 + 1)$ and by theorem 1 babbling even by the players with bias 0 is inevitable if and only if $b/(p - 1/2) > n/2$. In this case, any room containing players of both bias types will lead to babbling. Segregating the two groups is consequently both welfare optimal and an equilibrium.

This illustrates that segregation is optimal and an equilibrium if polarization is high (i.e. if b is large), and full integration is optimal and an equilibrium if polarization is low (if b is sufficiently low). For intermediate levels of polarization, the welfare optimal room allocation need not be an equilibrium. More precisely, the two groups may not be segregated enough in any equilibrium. Intuitively, if segregation is welfare optimal, players might have an incentive to switch to the room where the players with the opposite bias are because this allows them to receive more messages. They neglect the negative externality of this deviation, namely the loss of their own truthful message for players of their own bias. These results are depicted in figure 4.4.

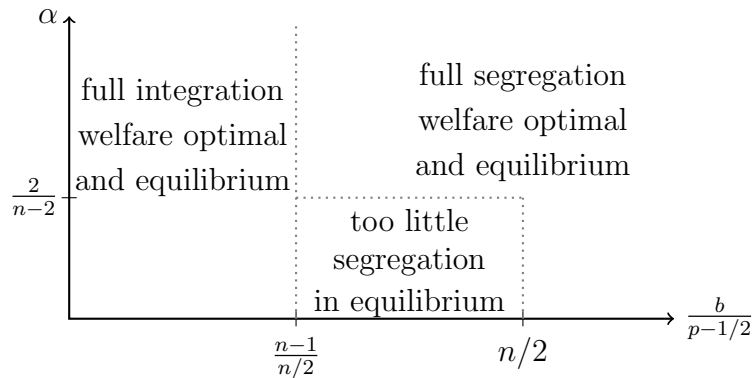


Figure 5: Welfare and equilibria for equally sized bias groups.

When the welfare-optimal room allocation is not an equilibrium, the welfare-maximizing equilibrium is straightforward: All players of one type, say bias 0, are in one room and are joined by m players of bias b . The players with bias 0 tell the truth, while the m players with bias b babble. All other bias b players are in a separate room, where they tell the truth. The number of babbling players, m , is such that one additional bias b player in the mixed room would lead to babbling of the players with bias 0.⁹ Hence m decreases in b , until it falls to zero and full segregation is welfare-optimal and an equilibrium. From a welfare perspective, there is too little segregation in any equilibrium with a positive number m of players who babble, and the resulting babbling constitutes a socially undesirable information loss.

When the two bias groups are not of equal size, say $n_0 > n_b$ for concreteness, results are similar to above but there is now the possibility that two not fully segregated rooms are welfare optimal. To see this, consider $b/(p - 1/2)$ just high enough such that players of bias b (the minority) would no longer be truth-telling in a fully integrated room. It can then be optimal to put one (or a few) players with bias 0 in a separate room if this restores truth-telling incentives for players with bias b . Note that this may not be an equilibrium if α is small: The bias 0 players that are isolated might find it beneficial to

⁹That is, m is the integer such that $bm/(n/2 + m) - (p - 1/2)(n/2 + m - 1)/(n/2 + m) \leq 0 < b(m + 1)/(n/2 + m + 1) - (p - 1/2)(n/2 + m)/(n/2 + m + 1)$ by theorem 1.

deviate to the big room as they can get more information there. The optimal equilibrium is in this case the fully integrated room (in which bias b players babble). Hence, we obtain too little segregation in equilibrium. For slightly higher $b/(p - 1/2)$ the just described room allocation may no longer be feasible as truth-telling is no longer a best response in a room with n_b players of each bias. It then becomes optimal to have one room for all players in which the majority is truthtelling while the minority listens to the majority and babbles itself. Clearly, this is also an equilibrium. Figure 6 schematically illustrates welfare optimal and equilibrium room allocation. We refer the reader to the supplementary material for a full analysis.¹⁰

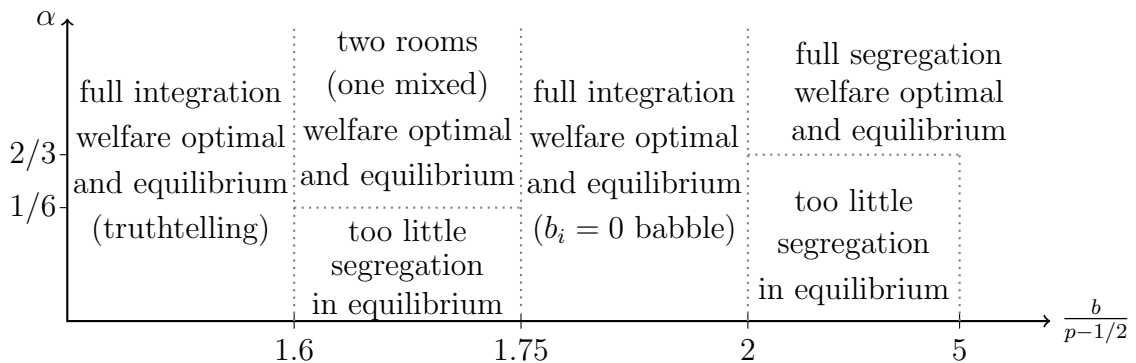


Figure 6: Welfare and equilibria when $n_0 = 5 > 4 = n_b$. (not to scale)

4.5. Polarization Destroys Welfare

We have argued that segregation is a rational and Pareto-optimal response to polarization. This does not mean that polarization in itself increases welfare – quite the opposite. If we return to the η -parametrization under which we derived our general results on integration and segregation in section 4.2, we can show that welfare is weakly decreasing in η , i.e. our measure of polarization.

Proposition 3. *Denote expected welfare in the welfare optimal room assignment with bias configuration \mathcal{B}_η by $W(\eta)$. $W(\eta)$ is decreasing in η . (Proof on page 41.)*

To illustrate this result, consider the following thought experiment: Starting with any bias configuration and any room allocation, we increase η . This will weakly decrease communication in any room, which harms welfare. Allowing for further segregation may restore some communication, which reduces the harm – but not completely.

We should hence be very precise about the mechanism by which higher polarization decreases welfare. It is not through segregation, even though higher polarization causes more

¹⁰There is one further scenario that does not show up in the example with $n_b = 4$ and $n_0 = 5$: For $b/(p - 1/2)$ slightly above $(n - 1)/n_b$, it can be welfare optimal to isolate one (or a few) players with the minority bias while keeping all other players in one room. This allows the majority in this room to be truth-telling while babbling would ensue in a fully integrated room. This scenario occurs when group sizes differ a lot.

segregation, which ultimately causes less information to be exchanged. Saying “segregation lowers welfare” would ignore the crucial intermediate step, which is that polarization in itself causes an informational breakdown. In fact, segregation *mitigates* this breakdown, without of course being able to restore communication between people that are now in separate rooms.

One could think of echo chambers as society’s (decentralized) defense mechanism against polarization. Like fever in a human body, segregation occurs as the effect of an underlying problem, and its presence hence indicates that polarization is at problematic levels. Echo chambers, and segregation more generally, are hence a symptom of polarization. And just like artificially lowering fever, treating the symptom without addressing the cause can in fact exacerbate the situation. Reducing polarization will weakly improve welfare; reducing segregation may not.

5. Uncertainty

So far, we have assumed that all biases b_i are common knowledge. In real-life situations, the type of a sender is often not known, so that the receiver is drawing inferences about the state of the world and the type of the sender at the same time. This can make informative communication much harder. In this section, we consider the effect of uncertainty about biases on the existence of within-room equilibria, and on the room choice game.

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 previous sections, in which all biases were always identical to their expected value. When we talk about “introducing” or “adding” uncertainty in this section, 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 part B of the appendix.

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 σ_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 4. *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 44.)*

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 5. *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 45.)*

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, μ_{ji}^l is j 's belief about θ_i , given that i has sent the signal m^l ; the other expressions involving μ are defined analogously.

Proposition 6. *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 45.)*

Proposition 7. *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 $\bar{\sigma}_{F_i}$ such that for all such F_i with $\text{Var}(b_i) \geq \bar{\sigma}_{F_i}^2$, $\mu_{ji}^h - \mu_{ji}^l \leq \varepsilon$. (Proof on page 45.)*

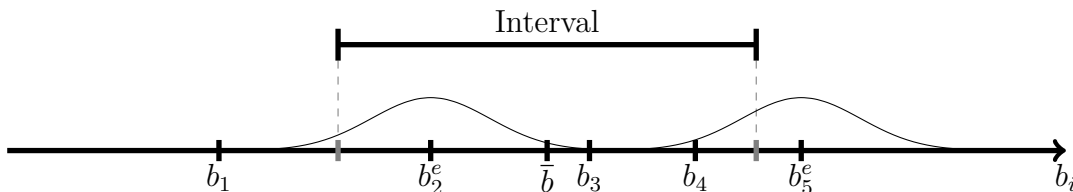


Figure 7: An illustration of propositions 5 to 7. (The biases are identical to the one in figure 1 except that b_2 and b_5 are now uncertain.)

Figure 7 illustrates propositions 5 to 7. 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 6). Player 5 was babbling, but will now sometimes send an informative message if his realized bias is close enough to \bar{b} (proposition 5).¹¹

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

¹¹This 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.

in one room. Of course, truth-telling between people with identical bias distributions is not necessarily easier – note that proposition 4 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 8. *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.*
- *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 45.)

To illustrate this result, let us return to the example on taxation from the 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.

6. Empirical Evidence from Twitter

The main mechanism in our model is that information transmission may be impossible if there is a large difference in preferences between a sender and his expected audience. All other results follow from people’s rational response to this mechanism. In this section, we consider a real-life communication environment in which people can be thought of as having both different ideologies (i.e. bias) and different information, and engage in debate. In particular, we will analyze data from the micro-blogging service Twitter.

Twitter allows its users to send short messages of 140 characters¹² either to people who have followed them (“tweets”), or to specific receivers (“replies”). Replies are also public, and are especially visible to followers of the sender, the receiver, or to people who are reading a specific “thread” that was started by a message.

This coexistence of different audiences creates a natural environment to study the messages that individuals send when they believe they are talking to an audience of mostly like-minded people, or to a mixed audience, or to an audience of people they disagree with. In particular, we will examine whether we can see signs of information transmission being harder across large ideological differences, and of how people rationally respond to this difficulty. To do so, we first need to find a way to measure people’s ideology, and can then consider interactions between different senders and receivers. The following paragraphs describe our data collection and analysis step by step.

6.1. Preliminary Steps

First step: Building a dictionary We analyzed the tweets of all 535 current members of the U.S. Congress (100 senators and 435 members of the house of representatives) to build a dictionary of partisan words and bigrams (groups of two words). For that, we counted how often each word or bigram was used by Democratic and Republican members of Congress, and isolated the words whose usage was (i) high enough and (ii) different enough between parties.

Table 1 has some examples for partisan words. Note that the differences in usage might derive from using different words for the same thing (talking about “Obamacare” vs “affordable care act”) or from different focuses (talking about “Iran” vs talking about “women”). We are agnostic about where the differences come from.

The table also shows that the most intensely partisan words are often those used less frequently – in fact, all the words in the table on the right are used only by one side. We weight words according to their frequency to avoid over-extrapolating from small samples.

Second step: Scoring accounts Armed with this partisan dictionary, we can identify a person’s political leanings purely based on their twitter feed. For each word or bigram

¹²Since November 2017: 280 characters. However, our data was collected before this change in Twitter’s policy.

Democrats	Republicans	Democrats	Republicans
access	spend	#stonewall	barrow
health	tune	mink	@goshock
invest	via	eastside	korda
women	Iran	#vayp	ocar
afford	Obama	kobach	#neag
worker	#senatemajldr	#taxseason	#ruleofflaw
opioid	Obamacare	#broadbandprivacy	beckley
Republican	Collins	#confirmlynch	@heralddispatch
Trump	McConnell	#killthebill	@fgpao
GOP	#obamacare	#repdankilde	ouachita

Table 1: Left: Most partisan words among the words that were used very often (more than 1000 times) in our sample. “#senatemajldr” is a hashtag for the (Republican) Senate Majority Leader McConnell; Susan Collins is a Republican senator. Right: Most partisan words among words that were used at least 10 times in our sample. (Note that these expressions are stemmed.)

that this person uses and which is found in our dictionary, we assign a score based on how differently the term is used between parties. In the end, we arrive at an overall score for that person, based on all partisan terms they have used.

To demonstrate the effectiveness of our partisan dictionary, we have created scores for a number of political journalists and pundits, whose political leaning is known but who are not part of our sample.¹³ If our dictionary works well at scoring, we should be able to separate the journalists and pundits into partisan camps, only based on their twitter feed. Table 4 on page 48 of the appendix shows that we are able to do so with about 80% accuracy.

Third step: Sampling random twitter users We randomly sampled a number of twitter users who (i) mostly or exclusively tweeted in English, (ii) had tweeted at least 3000 times, (iii) had at least 1000 followers, (iv) wrote some tweets of their own (and did not only re-tweet other people’s tweets), and (v) tweeted sufficiently often about political topics (i.e. used enough terms from our political dictionary).

We scored these random twitter users based on their original tweets, i.e. all tweets that were not replies to or retweets of other tweets, so that each user is assigned a location on a left-right scale.

Fourth step: Making use of different visibilities When “tweeting”, users’ texts are read by different audiences, based on what type of tweets they are.¹⁴ Simple tweets by

¹³We used the list of the 20 most influential journalists and blogger on the right and left, respectively, from StatSocial (2015).

¹⁴See here for how the company itself describes visibility: <https://help.twitter.com/en/using-twitter/types-of-tweets>

user X are shown in the timelines of all users who follow X. A reply by user X to user Y is shown in the timelines of users who follow *either X or Y*.

Given that we have scored random twitter users based on their original tweets (which are only shown to their own followers), we can now examine how these twitter users interact with other twitter users, given that such interactions (if they are replies) are visible to a different audience than the tweets based on which we have scored the user.

6.2. Actual difference-in-difference analysis

Using our work from the previous steps, we have generated a data set containing 12,043 reply tweets sent from 87 senders to 3,730 receivers. For each of these interactions, we can determine the political score of the sender and the receiver, as well as the properties of the interaction itself. This allows us to examine how the nature of communication changes in the ideological distance between sender and receiver.

If we apply the ideas of our model, we would predict that with a larger ideological distance, the communication of actual information becomes harder and babbling becomes more likely. It is, however, not entirely clear what “babbling” is in this context, and which observable criteria it would have. Any language obtains meaning only through the equilibrium interplay between the sender’s intention to communicate truthfully and the receiver’s belief in the truthfulness of messages. Statistically speaking, babbling could hence look like meaningful communication in any number of dimensions, or deviate only in some specific dimensions.

It is possible to take a slightly agnostic position on what babbling *looks like* and instead focus on the *consequences* of babbling. If we assume that people want to communicate actual information and that this becomes harder with larger ideological differences, we should expect people to adapt by changing the frequency or nature of their messages. In particular, we can think of three responses. First, people who find it hard to communicate with ideologically distant others will communicate more with those who are ideologically close. Second, if information exchange is hard, people will send more emotional and negative messages without much content – simply for emotional reasons and without trying to communicate any information. Third, if the communication of unverifiable information is hard, some people may try to communicate information anyway by making more complex arguments and providing verifiable information. We will take each of these three effects in turn, and show that they are all present in our data.

Less communication if ideological distance is larger We can show that there are fewer interactions across the ideological spectrum, compared to interactions between people with similar ideology. This is what our model would predict. Figure 6.2 shows this relationship including a linear best fit; table 5 in the appendix shows the exact regression results. The relationship is strongly positive, meaning that the more right-wing a Twitter

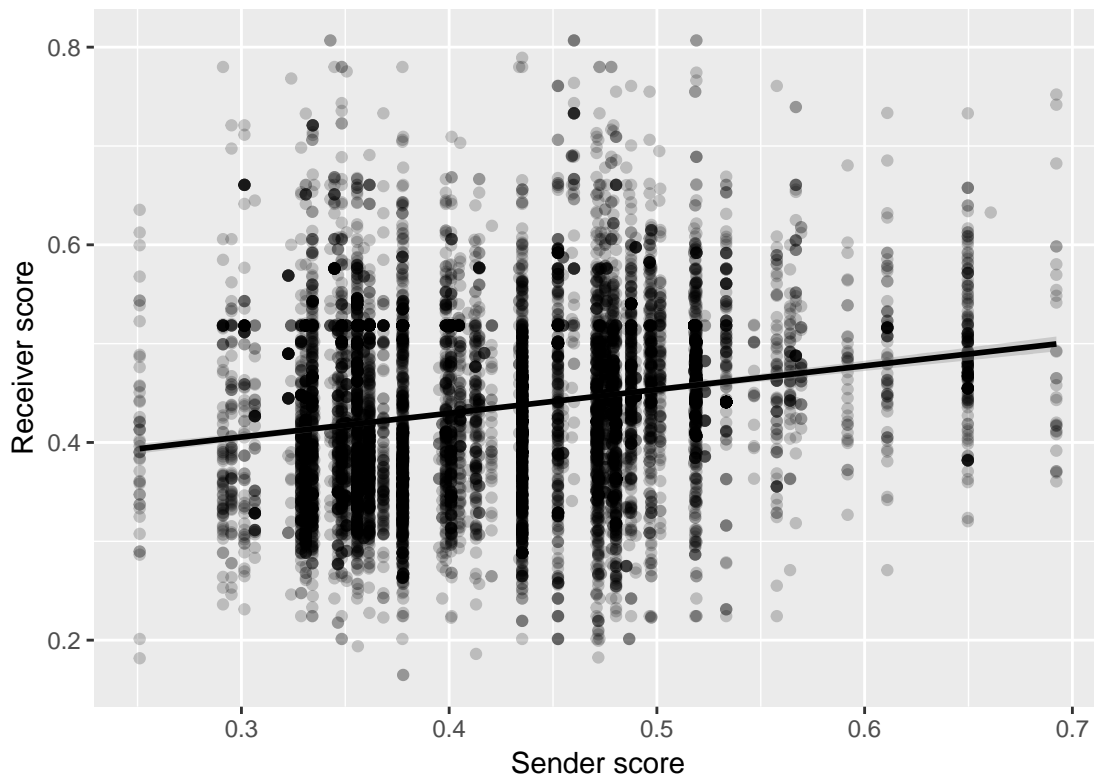


Figure 8: Political score of the sender and the receiver for 12043 interactions. The line shows the linear best fit (with a gray 99% confidence interval).

user is, the more he will on average interact with other right-wing Twitter users.¹⁵ If we split the writers of tweets in our sample into quartiles depending on their political score, the lowest quartile will on average reply to tweets by people with a score of 0.4234, while the highest quartile will respond to people with an average score of 0.4752. (The standard deviation of receiver scores in our sample is 0.09.) This is consistent with people refraining from communication across ideological boundaries because it is harder, and it supports the existence of “echo chambers” in how people interact in our dataset.¹⁶

More short, negative tweets When cheap-talk communication is difficult, people may send messages to each other for other reasons than to transmit information. A person who is confronted with a political opinion they disagree with may simply wish to voice their disapproval by sending a short, emotional and negative message. We can check for evidence of such behavior by examining the interaction between the length of a tweet, its emotional content, and the ideological distance between sender and receiver. We measure a tweet’s emotional content using the sentiment dictionary by Hu and Liu (2004), which

¹⁵Of course, there is already some inbuilt bias in the ideological leaning of the people whom a user follows, and whose tweets he is hence most likely to see. This would in turn influence whom he responds to. But we would argue that since this bias results from the user’s choice, it is endogenous and therefore consistent with users following people with whom communication is easier.

¹⁶Of course, we are not the first to show segregation on twitter – consider, for example, the studies by Barberá et al. (2015) or Krasodomski-Jones (2017).

gives scores to certain words and phrases that mark positive or negative content. Consider the following linear model (in which α_i represents a sender fixed effect):

$$\text{sentiment} = \alpha_i + \beta(\text{absolute score difference}) + \gamma(\text{tweet length}) + \delta(\text{score difference} * \text{tweet length})$$

Table 2 shows that $\beta < 0$, $\gamma > 0$ and $\delta > 0$.¹⁷ In words: (i) Reply tweets are more negative the larger the ideological distance between sender and receiver; (ii) short tweets contain more negative emotional words than long tweets; and (iii) short tweets are more negative, the larger the ideological distance between sender and receiver. Figure 9 shows the same relationship graphically.

	<i>Dependent variable:</i>
	sentiment score
log(tweet length)	0.048*** (0.026, 0.070) p = 0.00002
abs(score difference)	-1.055** (-1.868, -0.241) p = 0.012
log(tweet length):abs(score difference)	0.233** (0.025, 0.440) p = 0.028
Sender fixed effects	Yes
Observations	5,473
R ²	0.061
Adjusted R ²	0.046
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Table 2: Tweets are more negative if they are shorter and sender and receiver differ ideologically; the latter relationship is especially strong for short tweets. (Values in brackets show the 95% confidence interval.)

More long and complex tweets with hyperlinks Of course, unverifiable cheap talk is not the only way that people can exchange information. An opponent who suspects me of wanting to mislead him has no reason to believe my statement that “I think you are wrong”. “Look at this report by the national statistical office which shows that you are wrong”, however, is another thing. Some minds can also be changed by language if it does not just transmit viewpoints, but complex arguments – consider, for example,

¹⁷The number of observations is smaller than for the previous steps of the analysis since only a fraction of tweets contain enough terms to conduct a sentiment analysis.

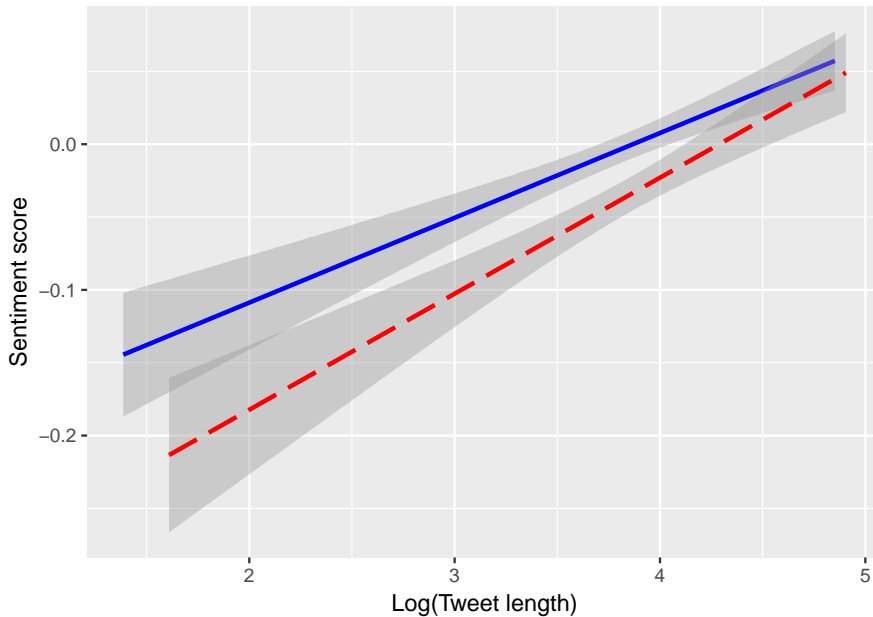


Figure 9: Linear model predictions of the emotional content of tweets as a function of tweet length for large ideological distance (red dashed line) and small distance (blue solid line). Long tweets are statistically indistinguishable, but short tweets between ideologically different people are much more negative than between similar people. (95% confidence interval in gray.)

that a reader may be unconvinced that our theorem 1 is correct until they have read the proof on page 37. In the context of internet debate, such “verifiable” information would usually take the form of hyperlinks (to presumably reliable sources of information) and more complex language.

We can measure such attempt at persuasion (as opposed to cheap talk) by considering whether reply tweets grow longer, more complex (measured by average word length¹⁸) and contain more hyperlinks as the score difference between sender and receiver increases. Table 3 shows that this is the case.

7. Discussion

7.1. Who provides the Rooms?

In our model, we have assumed that the rooms are available in sufficient quantity so that players who want to segregate themselves can do so. In reality, that is of course not guaranteed. Information exchange could literally be impossible for lack of an empty room, such as when co-workers find themselves unable to discuss sensitive questions in an open-plan workspace. Bernstein and Turban (2018) have shown that the creation of open-plan offices tends to decrease the number of (public) face-to-face interactions and increase

¹⁸Average word length is an integral part of many widely-used readability scores, such as the Automated Readability Index or the Coleman-Liau-Index.

	<i>Dependent variable:</i>		
	log(tweet length) (1)	log(mean word length) (2)	link dummy (3)
absolute score difference	0.378*** (0.198, 0.559)	0.205*** (0.126, 0.284)	0.172*** (0.042, 0.302)
Sender fixed effects	Yes	Yes	Yes
Observations	12,019	11,975	12,034
R ²	0.137	0.082	0.199
Adjusted R ²	0.131	0.075	0.194
p-value	0.00005	0.00000	0.00964

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3: Tweets get longer, more complex, and contain more hyperlinks as the ideological difference between sender and receiver increases. (Values in brackets show the 95% confidence interval.)

the number of (segregated) electronic interactions among colleagues. Or the shortage of rooms could be more figurative, such as when a politician may want to discuss his doubts of a policy with colleagues but cannot find a forum in which to do so without potentially giving ammunition to his political opponents.

In both cases, we have seen that segregation may be in the interest of everybody involved. It benefits not just the sender and the receiver in the segregated room, but even those who end up being excluded – since their inclusion would render communication impossible and thus not benefit anyone. Since rooms provide such clear benefits and are not automatically available, those in need of them should be willing to pay for whoever can provide them. We could imagine a group of agents who are sufficiently polarized and caught together in one place, which makes them unable to exchange any information. If now a plucky entrepreneur opened a separate room and took a small entrance fee, it would be an equilibrium for one group of agents to each pay the fee, enter the room – and improve their own and everybody else’s situation.

We think that this fable provides a way to understand the success of social messaging platforms such as Facebook, Twitter, WhatsApp and Snapchat. Each of these allows its users to send messages (and other content) to certain groups of others, with varying possibilities of exclusion. It can seem from the outside as if the service that is provided is to connect people with each other, but our model suggests it is just as much to exclude some people and not others, while providing sophisticated ways to determine who should and should not be excluded.¹⁹ This has a strict economic logic to it: Once the Internet

¹⁹Facebook, for example, allows its users among other things to (i) choose which of their data is visible

is available and ubiquitous, simply connecting people is not a scarce resource or service. But connecting them in such a way that they want to communicate truthfully, and can exchange the information they want to exchange, is much harder, and those who do it well can make a profit.

7.2. Political Parties and “Safe Spaces”

Of course, the room structure need not be provided by the market, it could be created by the agents themselves so that they can communicate with others who share their interests and world view. Besides the obvious examples of clubs and societies, we think that this is one rationale for the existence of political parties. In a society that is polarized enough, political parties can help solve the problem of aggregating political views and opinions.

We should also note that while messages are meaningless if a player is not truth-telling in equilibrium, the messages that he is most reluctant to send are those that could be seen as being counter to his own interest. For example, if an agent’s b_i is much lower than the average of all b_j , he has no problem truthfully reporting σ^l , but is more reluctant after σ^h . This is how political parties can be useful: by providing a secluded forum in which, for example, members of a party can discuss the flaws and merits of their own candidates or programs. They would not be able to have this kind of discussion in the presence of members from other parties, where they would become overly defensive of “their” candidates and programs.

But the problem of defensiveness also provides an argument for so-called “safe spaces”, i.e. spaces in which minorities or marginalized groups can communicate without outside interference. Informationally, such safe spaces may provide opportunities to communicate that would otherwise not exist. Consider the problem of two vegetarians who privately doubt whether vegetarianism is indeed a sensible choice – yet they find themselves defending it whenever they talk to (or in the presence of) non-vegetarians. Providing a “safe space” for vegetarians would allow them to discuss freely, and would hence provide a Pareto-improvement.

7.3. Rooms as Commitment and Information Design

The main unit of groups in our model is the room, in which everybody communicates with everybody and which is disjoint from other rooms. This allows us not just to examine the influence of bias differences on communication, but also the disciplining as well as the destructive effects of room composition.

Consider, for example, a setting in which person 1 is willing to tell the truth to 2, and 2 is willing to tell the truth to 3, but 1 is not willing to tell the truth to 3 (and all

to search engines, (ii) choose for each post and image whether it is visible to everybody or just friends or friends of friends or even select group of friends (iii) block individual other users from seeing certain content (iv) create public or private events or groups to which members can be invited, (v) message directly with selected users or groups of users. All of these are tools of intelligent segregation, not connection.

relationships hold vice versa). If we were to simply allow 1, 2 and 3 to form bi-directional communication links, that would be the end of it. But if they need to communicate within rooms, the effects are more interesting: 2's presence could discipline 1 or 3's message, and hence make communication possible between people who would not otherwise be able to exchange information. Or 3's presence could prove to be a centrifugal force to the whole room and make communication between 1 and 2 impossible without inducing 2 to tell the truth to 3, thus shutting down any communication.

We can hence think of rooms as an information design tool, and room allocation as an information design problem – where other room members provide a source of commitment. The results we have derived in section 2.2 show exactly how and when that is possible. For example, a social planner that wants to achieve a maximum of communication between ideologically distant groups could rely on the addition of moderates. In situations where individual information can never be observed – neither *ex ante* nor *ex post* – this may well be the only information design tool available.

7.4. When are echo chambers bad?

Criticisms of segregated debate or echo chambers commonly rely on a combination of informational and behavioral arguments. The most common informational point is that diversity of information sources increases the accuracy of information. Behavioral arguments usually hold that people do not learn correctly if only faced with some opinions. Our model suggests that we should take the informational arguments with a grain of salt, and that even if we believe in the behavioral factors, they do not necessarily amount to an argument for full integration.

Diversity. Our model considers gains from diversity in the sense that one's information gets more accurate (and hence one's decision better), the more people one hears from. We can thus weigh a well-known benefit of diversity (more information) against its less-discussed cost (problems with communication). An additional line of argument may assume that information is more closely correlated between people with similar biases – so that interaction with people with different biases becomes more valuable. Even that, however, does of course not solve the problem that communication across large preference differences may still be impossible, no matter how valuable the information that the other side holds.²⁰ Overall, there is simply no use in meeting people with a very diverse set of opinions and very useful information, if there is no way to get that information out of them.

²⁰We consider an extension of a model in which there is only one state, and people with similar bias receive correlated information about it, in the supplementary material.

Behavioral arguments. Once they hear only from people who are like them, people may fail to account for the correlation between the messages they receive.²¹ Or they may fail to correctly learn in other, less well-defined ways, all of which make it harder for them to infer the state of the world from hearing only one side of the story. None of this, however, means in itself that a person would learn more if also exposed to viewpoints that they would not normally encounter, if their interlocutor rationally adjusts the informativeness of his message depending on whom he wants to inform and whom not.

Segregation by taste. There are two ways of applying the insights of this paper. The first, which we have used in developing our argument, is to see segregation as an informationally rational and welfare-optimal choice. Another perspective would be to assume that people segregate for exogenous or emotional reasons, or simply for reasons of taste. For example, rich people live in rich neighborhoods because of nicer houses and better infrastructure, and the segregation of types is only a secondary effect. But is such segregation necessarily informationally inefficient and bad for welfare? Our model suggests that this need not be the case. While rich people could surely learn from exchanging information with people whose lifestyle is different from theirs, it is far from given that such communication successfully takes places if we simply bring rich and poor together.²² Even taste-based homophily can end up improving everyone’s information.

8. Conclusion

Modern democratic societies have three main mechanisms to aggregate information: Debate, markets, and votes. Of the three, debate is arguably the oldest – and while the other two require an organized framework and somebody who can enforce the rules, debate just needs an ability to speak and to listen.

But when will people speak truthfully (and hence have reason to listen)? In this paper, we have argued that if people have different preferences as well as different information, segregation into like-minded, homogeneous groups can be individually rational and Pareto-efficient. Echo chambers are not necessarily as destructive as popular discourse can make them seem. But even more importantly, we have shown that if segregation happens, it is not in itself the *cause* of an inability to debate. Instead, the existence of echo chambers is the *consequence* of differences in preferences, and of uncertainty and mistrust about other people’s motives.

This has implications for how to improve debate. Society has a lot to gain from getting people with diverse backgrounds, experiences and opinions to exchange their views. But

²¹C.f. the experimental work by Kallir and Sonsino (2009) and Eyster and Weizsäcker (2011) on “correlation neglect”.

²²Policies that may be more successful, following the results of our model, are: Narrowing the conflict of interest between rich and poor; convincing them that they have common goals; or reducing the uncertainty about each other’s interests.

this can not simply be achieved by forcing or cajoling people to interact who would not do so out of their own choosing. In fact, that could be counter-productive, as it could destroy disjoint echo chambers in which communication works, in favor of large integrated groups in which it does not. Our research, which we have set out in this paper, suggests that meaningful debate can only happen if the participants feel that they have sufficiently much in common and they trust each others' motives. That may be a taller order than simply putting people into a room and expecting them to come out smarter and in agreement. But functioning debate requires consideration for the motivations of the debaters.

Appendix

A. Proofs for the Model with Certainty

Proof of lemma 1 on page 10.

Let (m_1, \dots, m_n) be an equilibrium. Player i 's expected payoff when sending message m_i to players in room R_i can be written as

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

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

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

Specifically, sending message m^h gives expected payoff

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

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

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

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

$$\begin{aligned}
\Delta U_i(\sigma_i) &= (U_i(m^h) - U_i(m^l))/\alpha \\
&= - \sum_{j \in R_i, j \neq i} \mathbb{E} \left[\mu_{ji}^{h^2} - \mu_{ji}^{l^2} + 2(\mu_{ji}^h - \mu_{ji}^l) \left(b_j - b_i + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right) \middle| \sigma_i \right] \\
&= -2(\mu_{ji}^h - \mu_{ji}^l) \sum_{j \in R_i, j \neq i} \left[\frac{\mu_{ji}^h + \mu_{ji}^l}{2} + b_j - b_i - \mathbb{E}[\theta_i | \sigma_i] \right] \\
&= 2(\mu_{ji}^h - \mu_{ji}^l)(n_{R_i} - 1) \left[-\frac{\mu_{ji}^h + \mu_{ji}^l}{2} - \frac{\sum_{j \in R_i, j \neq i} b_j}{n_{R_i} - 1} + b_i + \mathbb{E}[\theta_i | \sigma_i] \right] \tag{3}
\end{aligned}$$

where n_{R_i} denotes the number of players in room R_i . (For the transformation to line 3, we make use of the fact that μ_{ji} is the same for all $j \in R_i$.)

Player i is only willing to choose a mixed strategy after receiving signal σ_i if $\Delta U_i(\sigma_i) = 0$. From expression (3) it is clear that this can only be true for at most one signal as $\mathbb{E}[\theta_i | \sigma_i]$ varies in the σ_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 σ^h . Then, $U_i(\sigma^h) = 0$ implies $U_i(\sigma^l) = 2(\mu_{ji}^h - \mu_{ji}^l)(n_{R_i} - 1)(1 - 2p) < 0$ 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 σ^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 $U_i(\sigma^h) > 0$ in the equilibrium candidate (m_i^t, m_{-i}) , i.e. truthful reporting is optimal for i after receiving signal σ^h . In the equilibrium candidate (m_i^t, m_{-i}) , truthful messaging is still optimal after signal σ^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 $-(\mu_{ji}^h + \mu_{ji}^l)/2 + p = \sum_{j \in R_i, j \neq i} b_j / (n_{R_i} - 1) + b_i$, we get that $-1/2 + 1 - p < \sum_{j \in R_i, j \neq i} b_j / (n_{R_i} - 1) + b_i$ 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 σ^l is analogous. \square

Proof of theorem 1 on page 11.

Consider again the difference between lying and truth-telling for player i that we considered in equation (3) in the proof of lemma 1. Following corollary 1, we only consider pure strategies and therefore for every non-babbling player $\mu_{ji}^h = p$ and $\mu_{ji}^l = 1 - p$ which

implies that $\Delta U_i(\sigma^h) \geq 0$ simplifies to

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

If this inequality does not hold, player i will not use the truthful strategy in the most informative equilibrium and by corollary 1 this implies that he will babble in the most informative equilibrium.

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

Proof of proposition 1 on page 12.

Denote the sets of babbling and truthful players in room R_j as R_j^{bab} and R_j^{truth} , respectively. For a given room allocation, the expected payoff of player i in room R_i is

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

For any $i \neq j$, the two values of θ_i and θ_j are independent; the same is true for μ_{ij} and μ_{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

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

Now note that $\mathbb{E}[(\mu_{jk} - \theta_k)^2]$ can have two possible values: If $k \in R_j^{truth} \cup \{j\}$, i.e. if j has received information about θ_k , then $\mathbb{E}[(\mu_{jk} - \theta_k)^2] = p(1 - p)$. If j has not received information about θ_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 means that if i is telling the truth, we can write

$$U_i^{truth} = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - \frac{1}{4} [n + \alpha(n - 1)n] \\ + \left(\frac{1}{4} - p(1 - p) \right) \left[n_{R_i}^{truth} + \alpha \sum_R \{ n_R^{truth} n_R^{truth} + (n_R - n_R^{truth})(1 + n_R^{truth}) \} - \alpha n_{R_i}^{truth} \right] \quad (4)$$

The first term represents the loss that i suffers because other players choose a decision that is by $b_j - b_i$ too high from i 's point of view. The second term represents the (theoretical) loss that would result if no player had any information and all μ 's were simply $\frac{1}{2}$. The factors n and $(n - 1)n$, which sum up to n^2 , represent the total number of possible pieces of information in the model: If everybody's signal was available to everyone, n people would receive n pieces of information. The term hence represents, for each potential piece of information, the loss to i of that information not being available.

This loss is mitigated by information, which we see in the second line: i receives his signal and $n_{R_i}^{truth} - 1$ truthful messages, which means that instead of $\frac{1}{4}$, on each of these pieces of information i loses only $p(1 - p) < \frac{1}{4}$. Other players, about whose decisions i cares with weight α , also receive some signals/messages: in any given room R , n_R^{truth} players receive their own signal and $n_R^{truth} - 1$ truthful messages while $n_R - n_R^{truth}$ players (those that babble in R) receive n_R^{truth} truthful messages and their own signal. (We have to subtract the correction term $-\alpha n_{R_i}^{truth}$ for room R_i in which there are only $n_{R_i}^{truth} - 1$ other players who tell the truth – in other words, i cannot count himself again as one of the players who receive information.) Analogously, we can write

$$U_i^{bab} = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n - 1)n] \\ + (1/4 - p(1 - p)) \left[1 + n_{R_i}^{truth} + \alpha \sum_R \{ n_R^{truth} n_R^{truth} + (n_R - n_R^{truth})(1 + n_R^{truth}) \} \right. \\ \left. - \alpha(1 + n_{R_i}^{truth}) \right]. \quad (5)$$

In both the expressions for U_i^{truth} and U_i^{bab} , the second lines are adjusting the (pessimistic) expression in the first line for the reduction in variance by information. We can

simplify both expressions by simply writing

$$U_i = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 [n + \alpha(n-1)n] + (1/4 - p(1-p)) \left[\zeta_i + \alpha \sum_{j \neq i} \zeta_j \right] \quad (6)$$

and express welfare as

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

In this expression, all terms are model parameters except for the sum over all ζ_i , which shows that welfare is linearly increasing in $\sum_i \zeta_i$. \square

Proof of proposition 2 on page 14.

Let $\Delta\zeta_i$ be the change in ζ_i that results from a deviation, and $\Delta \sum_{j \neq i} \zeta_j$ the change in $\sum_{j \neq i} \zeta_j$ resulting from the same deviation. We know that in the welfare-optimal room allocation, there can be no deviation by a single player that would increase welfare. That means that in the welfare-optimum, it must be

$$\Delta\zeta_i + \Delta \sum_{j \neq i} \zeta_j \leq 0. \quad (7)$$

The condition that there is no profitable deviation for player i is

$$\Delta\zeta_i + \alpha \Delta \sum_{j \neq i} \zeta_j \leq 0, \quad (8)$$

which is identical except for the factor α . From this we can immediately see that 7 implies 8 if $\alpha = 1$. \square

Proof of theorem 2 on page 17.

Recall that a truth-telling equilibrium exists if and only if for every player i it is

$$\left| \sum_{k \neq i} \{b_k / (n-1)\} - b_i \right| \leq \frac{1}{2}.$$

This can be rewritten as $|\sum_k \{b_k\} - nb_i| / (n-1) \leq \frac{1}{2}$. If η is sufficiently small, 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, and no player can gain from

leaving the room.

If $\left| \sum_{k \in R_i, k \neq i} \{b_k / (n - 1)\} - b_i \right| > \frac{1}{2}$, then i will not be truthful when receiving either signal σ^l or σ^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 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 η while the right hand side is not. That is, for η sufficiently high, player i will report the highest (lowest) signal in all rooms in which $\sum_{k \in R_i, k \neq j} b_k < n_{R_i} b_i$ ($\sum_{k \in R_i, k \neq j} b_k > n_{R_i} b_i$). 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 η is sufficiently high. For high enough η , this holds true for all players and it is then obvious that full separation is both welfare maximizing and an equilibrium. \square

Proof of proposition 3 on page 20

Take two values of η , namely η' and $\eta'' > \eta'$. Denote a welfare optimal room assignment under η'' by R'' . Consider the same room assignment R'' with biases η' . In each room the number of pieces of information is weakly higher with set of biases $\mathcal{B}_{\eta'}$ than with set of biases $\mathcal{B}_{\eta''}$: By theorem 1 a player i is truthtelling if and only if $\eta \bar{b} - \frac{n_{R_i} - 1}{n_{R_i}} (p - \frac{1}{2}) \leq \eta b_i \leq \eta \bar{b} + \frac{n_{R_i} - 1}{n_{R_i}} (p - \frac{1}{2})$. Hence, player i will be truthtelling in room R_i'' with biases in $\mathcal{B}_{\eta'}$ if he is truthtelling in R_i'' with biases $\mathcal{B}_{\eta''}$ by $\eta' < \eta''$. Consequently, there is weakly more information transmitted in every room given assignment R'' under η' than under $\eta'' > \eta'$. This implies $W(\eta') \geq W(\eta'')$ by proposition 1. \square

B. Detailed Analysis and Proofs for the Model with Uncertainty

B.1. Preliminary Analysis

Similarly to the derivation of expression (3), 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 μ_{ji}^h and μ_{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 (3) 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. λ^h and λ^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 σ^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):

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

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

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 σ_i , i should never want to misrepresent σ^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 2. *Assume that $\lambda^h + \lambda^l > 1$. Then i always strictly prefers to truthfully reveal (i) σ^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) σ^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 λ^h and λ^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 λ^h and λ^l as the marginal probabilities of truth-telling by i .

Lemma 3. *There exists an equilibrium in which i truthfully reveals σ^h with marginal probability λ^h and truthfully reveals σ^l with marginal probability λ^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 $\Delta U_i(\sigma^l) \leq 0 \Leftrightarrow$

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

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 3 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 2 imply $\lambda^l + \lambda^h = 1$ and therefore uninformative messages.²³

B.2. Proofs

Proof of proposition 4 on page 22.

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 3 imply $\lambda^h + \lambda^l = 1$, and hence there exists no equilibrium in which any of the players tells the truth. \square

²³To be precise, both points at which F_i is evaluated in lemma 2 lie in the interior of the interval $\left[\frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} - (2p - 1), \frac{\sum_{j \in R_i, j \neq i} b_j^e}{n_{R_i} - 1} + (2p - 1)\right]$ 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 2 then hold with equality, they imply $\lambda^h + \lambda^l = 1$ which in turn implies $\mu_{ji}^h - \mu_{ji}^l = 0$.

Proof of proposition 5 on page 22.

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 3 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 3 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 6 on page 23.

By the symmetry of F , all F^κ have the same expected value. We can find a $\bar{\kappa}$ small enough so that F^κ 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 3 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 7 on page 23.

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 2 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 2 and the continuity of μ_{ji} in λ^h and λ^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 $\overline{\sigma_{F_i}^2}$ has to have more than ε' probability mass above \bar{b}^ε or below $\underline{b}^\varepsilon$ as $\overline{\sigma_{F_i}^2}$ 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 $\overline{\sigma_{F_i}^2}$ will lead to $\mu_{ji}^h - \mu_{ji}^l \leq \varepsilon$. \square

Proof of proposition 8 on page 24.

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 2 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

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

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

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

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

C. Empirical Work: Tables and Figures

Account	Score
DavidCornDC	0.3515
RBReich	0.3672
ezraklein	0.3682
Lawrence	0.3763
ariannahuff	0.3951
chrislhayes	0.3968
mattyglesias	0.4062
mtaibbi	0.4228
nycjim	0.4244
NickKristof	0.4249
NateSilver538	0.4272
AnnCoulter	0.4316
ggreenwald	0.4373
jaketapper	0.4475
stephenfhayes	0.4576
MHarrisPerry	0.4576
KatrinaNation	0.4589
maddow	0.459
megynkelly	0.4624
jdickerson	0.4662
secupp	0.4764
greta	0.4779
EWErickson	0.4898
greggutfeld	0.4923
michellemalkin	0.4936
DLoesch	0.4962
glennbeck	0.4969
camanpour	0.5002
brithume	0.5051
AHMalcolm	0.5078
MajorCBS	0.5229
seanhannity	0.534
tavissmiley	0.5354
AnnCurry	0.536
AndreaTantaros	0.5384
andersoncooper	0.5667
DanaPerino	0.5695
krauthammer	0.5969
BretBaier	0.6036
FareedZakaria	0.6138
TuckerCarlson	0.6157
edhenry	0.6457
Judgenap	0.6645
kinguilfoyle	0.7134

Table 4: The scores of the twitter feeds of 40 prominent American political pundits. The higher the score, the more Republican-leaning a pundit is.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.3336	0.0044	75.27	0.0000
scoreOriginalRel	0.2399	0.0103	23.19	0.0000

Table 5: The political scores of sender and receiver are correlated. The table shows the results of estimating the equation receiver score = intercept + β sender score.

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.
- Bernstein, E. S. and S. Turban (2018). The impact of the ‘open’ workspace on human collaboration. *Philosophical Transactions of the Royal Society B* 373(1753), 20170239.
- Chater, J. (2016). What the EU referendum result teaches us about the dangers of the echo chamber. <https://www.newstatesman.com/2016/07/what-eu-referendum-result-teaches-us-about-dangers-echo-chamber>. Accessed: 2018-02-10.
- Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica* 50(6), 1431–1451.
- Del Vicario, M., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113(3), 554–559.
- Eyster, E. and G. Weizsäcker (2011). Correlation neglect in financial decision-making. *DIW Discussion Papers* 1104.
- Farrell, J. and R. Gibbons (1989). Cheap talk with two audiences. *American Economic Review* 79(5), 1214–1223.
- Galeotti, A., C. Ghiglino, and F. Squintani (2013). Strategic information transmission networks. *Journal of Economic Theory* 148(5), 1751–1769.
- Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica* 78(1), 35–71.
- Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *Quarterly Journal of Economics* 126(4), 1799–1839.
- Hooton, C. (2016). Social media echo chambers gifted Donald Trump the presidency. <https://www.independent.co.uk/voices/>

- donald-trump-president-social-media-echo-chamber-hypernormalisation-adam-curtis-pro.html. Accessed: 2018-02-10.
- Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. In *AAAI*, Volume 4, pp. 755–760.
- Kallir, I. and D. Sonsino (2009). The neglect of correlation in allocation decisions. *Southern Economic Journal* 75(4), 1045–1066.
- Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies* 76(4), 1359–1395.
- Krasodonski-Jones, A. (2017). Talking to ourselves. <https://www.demos.co.uk/project/talking-to-ourselves/>. Accessed: 2018-07-30.
- Krishna, V. and J. Morgan (2001). A model of expertise. *Quarterly Journal of Economics* 116(2), 747–775.
- Lawrence, E., J. Sides, and H. Farrell (2010). Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspectives on Politics* 8(1), 141–157.
- Li, M. and K. Madarász (2008). When mandatory disclosure hurts: Expert advice and conflicting interests. *Journal of Economic Theory* 139(1), 47–74.
- Morgan, J. and P. C. Stocken (2003). An analysis of stock recommendations. *RAND Journal of Economics* 34(1), 183–203.
- Quattrociocchi, W., A. Scala, and C. R. Sunstein (2016). Echo chambers on Facebook. Available on SSRN.
- StatSocial (2015). The most influential political journalists and bloggers in social media. <https://www.statsocial.com/social-journalists/>. Accessed: 2018-02-10.
- Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.
- Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.