Why Echo Chambers are Useful*

Ole Jann — CERGE-EI

Christoph Schottmüller — University of Cologne and TILEC

November 29, 2021

Abstract

Why do people appear to forgo information by sorting into "echo chambers"? We model a society in which information is dispersed and preferences are polarized. Segregation into small, homogeneous groups can then maximize the amount of communication that takes place, thus making such segregation both individually rational and Pareto-efficient. We examine the optimal communication structure and give sufficient conditions for when it is attainable through endogenous group formation. A major problem is that people tend to segregate inefficiently little. Using data from Twitter, we show several behavioral patterns that are consistent with the results of our model.

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

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

*Jann: CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Prague; ole.jann@cerge-ei.cz. Schottmüller: Department of Economics, University of Cologne; c.schottmueller@uni-koeln.de. We are grateful for helpful comments by Rachel Bernhard, James Best, Ben Brooks, Vince Crawford, Marcelo Fernandez, Ben Golub, Sanjeev Goyal, Paul Klemperer, Vasily Korovkin, Nenad Kos, Meg Meyer, Kirill Pogorelskiy, David Ronayne, Larry Samuelson, Bill Sandholm, Karl Schlag, Jakub Steiner, Peter Norman Sørensen, Kyle Woodward and Peyton Young, as well as audiences at Bar-Ilan, Bonn, CERGE-EI, Cologne, Copenhagen, Groningen, Konstanz, Munich, Oxford, Sussex, Vienna, Warwick and Wisconsin-Madison and at EEA 2018 (Cologne), TTW 2018 (Northwestern), ESEWM 2018 (Naples), ECOP 2019 (Bologna), ESWM 2020 (Milan/virtual) and ASSA 2021 (virtual).

Large parts of society are organized around the non-market exchange of information: People gather around breakfast and dinner tables, in meeting rooms, committees, cafés and bars, while keeping in 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 tend to 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? Is it simply because of irrational biases or fear of confrontation, or can there be informational reasons? It also raises the urgent question of whether the resulting segregation in fact limits the spread of useful information and thus hurts society, and whether policy should seek to influence who communicates with whom.[2]

In this paper, we consider a society that faces two problems: Dispersion of information and polarization of preferences. Dispersed information means that no single individual knows enough about the world to make very successful choices. Polarized preferences mean that even if everybody had perfect knowledge about the world, people would still disagree on which action each person should take.[3] We develop a general framework to model how people rationally communicate within groups in the presence of information dispersion and preference polarization, and how they sort into groups while anticipating what communication within the groups will be like.

Our analysis shows that segregation into small, homogeneous groups can result from rational choice and *maximize* the amount of information available to any single individual. In fact, such segregation can be efficient and even Pareto-optimal for society. The optimal amount of segregation increases in the degree to which preferences are polarized.

Why is that? Information dispersion and preference polarization work in opposite directions: The former makes individuals curious about learning each other's information so they can make better choices; the latter gives them an incentive to misrepresent their information in order to influence others' choices. If the polarization of preferences increases in any given group, individuals become more and more interested in misleading each other in cheap talk communication à la Crawford and Sobel (1982) and truthful communication

[1] See, 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).

[2] Consider, for example, claims that echo chambers are "dangerous" (Grimes, 2017) and have "Balkanised society" (Itten, 2018), as well as played a role in populist insurgencies in Western democracies such as the "Brexit" referendum (Chater, 2016) or the rise of Donald Trump (Hooton, 2016).

[3] We understand "polarization" as a measure of distribution, similar to Esteban and Ray (1994) and the following literature. To avoid misunderstandings, we will use "polarization" only when talking about exogenously given preferences – i.e. a primitive of our model – and not when referring to information or beliefs.

becomes harder – even though individuals are still just as much interested in learning from one another. This means that in a highly polarized society in which everybody tries to learn from everyone, very little truthful communication may be possible because everyone also tries to mislead (almost) everyone. If we are able to split this society into segregated groups, however, people within each one of these groups may be a lot less polarized, so that for the purposes of their communication the effects of information dispersion now dominate those of preference polarization. While segregation into echo chambers limits the amount of *potential* communication, it makes *actual* communication possible.

If it is left to individuals to decide with whom to communicate, an efficient allocation may nevertheless be achievable and we give sufficient conditions for when that is the case: If individuals care sufficiently about the decisions of others, or if the polarization of preferences is very large or very small compared to the dispersion of information. But people will also sometimes segregate *too little* compared to what a social planner would choose. This is because choosing a group to communicate with has an informational externality: A person may choose a group from which she can learn the most, without fully internalizing how much others learn from her.

Our theory (and our title) should not be understood to mean that echo chambers are unambiguously good for society. What we do in this paper is to identify and isolate a mechanism by which they can be useful and increase welfare. This may, in many instances, be outweighed by ways in which they can be detrimental, some of which we discuss in section 7.4. Overall, our model calls for a nuanced view: Division into small, homogeneous groups can facilitate honest debate and real mind changes just as it can narrow world views and shut out crucial information. Which of these mechanisms dominates may not be apparent from the simple fact that there are groups, and can only become clear when we understand the structures of polarization and mistrust that underlie the sorting behavior.

Given the usefulness of disjoint groups in our model, we also suggest that the main business model of social networking sites could be understood as providing the infrastructure for people to sort in the way that they want – and not, in fact, simply to connect them, as is commonly assumed. We discuss this idea in more detail in section 7.1.

This paper has three main contributions. First, in sections 1 to 3 we develop a highly tractable framework to analyze strategic communication among n players, where everybody has different information and different preferences and can talk and listen to everybody else. The framework can accommodate endogenous choice of communication structures, different communication protocols, different types of uncertainty and other modifications. Second, in section 4 we use this framework to develop the theory described in the previous paragraphs. Finally, in section 6 we provide evidence from data collected on the micro-blogging service Twitter that is consistent with some of the main predictions of our work. The remainder of this introduction explains our theoretical and empirical methodology and findings in detail.

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 each chooses an action. The state of the world and each person's preference are real numbers; a person's ideal point is simply the sum of the state of the world and her preference. Each person wants all players' actions, including her own, to be as close as possible to her own ideal point, and her payoff is concave in the distance between anyone's action and her ideal point. People may therefore take different actions based on differences in information and in preferences. Differences in either dimension are sufficient for disagreement, i.e. people would choose different actions even if they all had the same information but different preferences or vice versa. Differences in information can be bridged by communication, while differences in preferences cannot.

We assume that people's preferences are common knowledge, though we relax this assumption in an extension. Before choosing whether (and what) to communicate, each person privately receives a binary signal about the state of the world. We make the simplifying assumption that each person's signal contains information about the state of the world, but no information about the information of others. Intuitively, different people observe different aspects of the world, and what one person observes does not tell her anything about what any other person knows. We will see how this assumption allows us to develop a much simpler and more tractable analysis than richer models in the literature.[4]

We assume that "cheap talk" communication takes place in disjoint "rooms", where each statement by an individual can be heard by anyone else within the room, but not by people in other rooms. If a person now finds herself in a room with a mixed group of others, she faces a trade-off between wanting to correctly inform those who have preferences close to her own (as that will bring their action closer to her own ideal point), and wanting to mislead those who have very different preferences. If most of her audience has a much lower preference parameter than her, for example, she would want to make them believe that her signal says that the state of the world is a high number, to counteract the fact that her audience will always choose an action that she deems too low.

We show that the question of who tells the truth and who babbles in equilibrium in any given room has a simple solution and only depends on the difference between a person's preference and the average preference of her audience (theorem 1).[5] This allows us to very easily compute how much communication can take place in the most informative equilibrium in any room. Following the backwards-induction logic, we can then analyze

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

[5]Unlike in Hagenbach and Koessler (2010), however, this distance result is driven by the trade-off between wanting to correctly inform some people and mis-inform others, and not by the wish to coordinate people's actions.

how a social planner would choose to allocate people to rooms, and how people can allocate to rooms in equilibrium.[6]

We show that the most informative equilibrium in any room is always in pure strategies, and that hence we can count information in discrete "pieces": Receiving a signal, or learning another player's signal through communication, are each equivalent to getting one "piece" of information (corresponding to 1 bit in Shannon entropy). Despite the fact that people face aggregate uncertainty, have different preferences and have concave preferences about their own actions and the actions of others, we can show that all payoffs (and hence also welfare) can simply be expressed in the integer amount of pieces of information that each player has after room choice and communication have taken place (proposition 1).

We consider two different types of preference polarization. First, we analyze a society that consists of people with two types of preferences. In section 4.2, we completely characterize the optimal room allocation for all possible group sizes and magnitudes of polarization in this case. We show that either the welfare-optimum is an equilibrium of the room-choice game, or people segregate *too little* in the welfare-optimal equilibrium.

In section 4.3 we consider a more general specification of polarization as "clustering" around certain values. We introduce a parameterization of polarization in such a model that is well-defined and well-ordered, and show the following result: If the relative polarization of preferences is large, full segregation by preferences is always welfare-optimal and an equilibrium, whereas integration is optimal and an equilibrium for low polarization. (theorem 2).

Finally, our analysis allows us to disentangle the welfare effects of polarization and segregation. An increase in a society's polarization leads to an increased desire for polarization as well as a welfare loss in equilibrium. An observer may hence be tempted to conclude that segregation itself has caused the welfare loss, but we can show that the opposite is the case. We show that polarization lowers welfare (proposition 2), but segregation actually mitigates the corrosive effects of polarization. Not allowing people to segregate in the presence of polarization would lower welfare. We can hence see segregation into echo chambers as not just an individually rational action, but as society's decentralized countermeasure against the welfare losses caused by polarized preferences.

In section 5, we show that our model remains tractable if we consider several extensions.

Empirical Evidence In section 6, we provide evidence that is consistent with 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

[6]Following the usual convention in the literature on strategic communication, we only consider the most informative equilibrium in each room and ignore less informative babbling equilibria.

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 those by current or recent 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. Such babbling may not be measurable in practice, since the same message can be informative or meaningless, depending on the sender's intention and receiver's expectation. Instead, we focus on trying to spot *consequences* of babbling.

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 such as hyperlinks.

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 emotionally negative and profanity-laden are the interactions we observe. 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.

As the basis of our analysis of communication, 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.[7] 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. Galeotti et al. (2013) analyze 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 (except in an extension), 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. Some of the results of our paper hence allow us to analyze "communication design" without commitment, and perhaps even without a designer.

In an extension of our model, we 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

[7] While 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.

of biases. Of 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 method of scoring 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 15 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), however, point out 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 directly by choice. 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. Even where echo chambers would have negative consequences that are not in our model, the effects that we describe would have to be reckoned with, and a nuanced discussion of how much segregation is optimal in debate is necessary. Most importantly, we argue that those who see ideological segregation as the ruin of societies are focusing on a symptom, not the cause. Polarization of preferences and mutual mistrust are doing the real damage; informational segregation can be a rational behavior that mitigates the harm they do.[8]

1. Model

There is an unknown state of the world $\theta = \sum_{i=1}^{n} \theta_i$. Each θ_i is independently drawn to be 0 or 1 with equal probabilities, so that θ is binomially distributed on $\{0, 1, \ldots, n\}$. n individuals each make an observation about the state. In particular, individual i receives a private signal $\sigma_i \in \left\{\sigma^l, \sigma^h\right\}$ of accuracy p about θ_i, i.e. $Pr\left(\sigma_i = \sigma^h | \theta_i = 1\right) =$

[8]Several recent papers mention the echo chamber phenomenon while focusing on exogenous news sources (or algorithm designers who do not care about the informational content of shared messages), e.g. Che and Mierendorff (2019); Martinez and Tenev (2020); Acemoglu et al. (2021). Our paper differs as strategic senders of information are at the heart of our argument.

$Pr\left(\sigma_i = \sigma^l | \theta_i = 0\right) = p > 1/2$. 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

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

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_i is the aspect of the world that player i has *information* about. The parameter α measures the relative weight players assign to other players' behavior – in other words, the sensitivity of i's payoff to the actions of other player. If $\alpha = 0$, i only cares about his own decision; if $\alpha = 1$ then every other player's decision is just as important to i as his own decision. 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.

More formally, let B_{R_i} be the vector of biases in room R_i and denote by $\mathbb{B} \in \Re \cup \cdots \cup \Re^n$ the set of all such bias vectors. A messaging strategy in stage 2 is then a map $m_i : \mathbb{B} \times \{\sigma^l, \sigma^h\} \to \{m^l, m^h\}$. Stage 3 actions assign to each combination of own signal σ_i and messages sent in R_i an action in $\Re$.[9]

We analyze the model by backwards induction: First we characterize the 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.[10]

[9] In principle, strategies could also depend on the composition of other rooms. However, ignoring this possibility is without loss of generality as a player cannot gain from such a dependence (given that messages are not payoff relevant).

[10] All 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.

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

$$a_i^* = b_i + \mathbb{E}[\theta | m_{R_i}, \sigma_i] = b_i + \sum_{j=1}^{n} \mathbb{E}[\theta_j | m_{R_i}, \sigma_i], \tag{2}$$

where m_{R_i} denotes the profile of messages sent in room R_i. In words, 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.

In the following, we will denote by $\mu_{ij} = \mathbb{E}_i[\theta_j | m_{R_i}, \sigma_i]$ i's expectation 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 expectations $(\mu_{ij})_{j=1}^n$, 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 payoff relevant 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, so that m_i is either perfectly uninformative or perfectly informative about σ_i.*
- mixed *if for some signal σ^k, $k \in \{l, h\}$, both messages are sent in equilibrium with positive probability 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.[11] The following lemma implies that the most informative equilibrium is in pure strategies.

Lemma 1. *Let $(m_1, \ldots, m_n)$ 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 35.)*

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 that induces lower actions by the players in his room. This means that the actions induced by m^l are somewhat too low from i's point of view and the actions induced by m^h are somewhat too high. 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 expectation of the other players 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 strategies and the resulting strategy profile remains an equilibrium. This new equilibrium is more informative as the truthful strategy is most informative (in the Blackwell sense) and therefore best for the receivers.

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 $\overline{b} = \frac{\sum_{k \in R} b_k}{n_R}$ 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*

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

[11] The 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.

and babbles otherwise. (Proof on page 36.)

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 $\overline{b}$ (which includes b_i), we have to add this correction.[12] 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' 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.

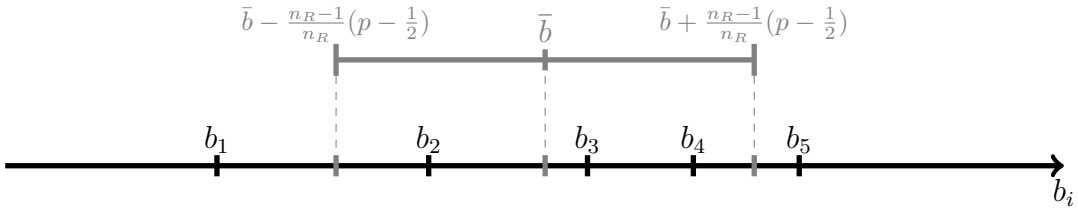

Figure 1: Finding the most informative equilibrium in a room consisting of players 1 to 5. We find the average bias and construct a symmetric interval around it. Players 1 and 5 babble in the most informative equilibrium, since their biases are too far from $\overline{b}$. Players 2, 3 and 4 tell the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

3. Room Choice

We can now analyze room choice, under the assumption that the most informative equilibrium will be played in any room. 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.

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.

[12] Intuitively, 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.

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 $\zeta_i \in \{1, 2, \ldots, n\}$ 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. *(i) Player i's payoff is a linear and increasing function of $\zeta_i + \alpha \sum_{j \neq i} \zeta_j$.*
(ii) Welfare is a linear and increasing function of $\sum_i \zeta_i$.
In both cases, the coefficients of the linear functions are given by model parameters. *(Proof on page 37.)*

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 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 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. It also leads to the following corollary:

Corollary 2. *If $\alpha = 1$, the welfare-optimal room allocation is also an equilibrium of the room choice game.*

Proposition 1 means that we can quickly compare the welfare of 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 and might have to consider many possible room allocations before concluding which one is optimal. 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. However, we derive general results on optimal and equilibrium room allocations in the next section.

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 minimize a weighted sum of their own uncertainty and that of the other players. In this section, we will use these results to draw a connection between the polarization of players' preferences, and the question of which room allocations are optimal, and which allocations can be achieved in equilibrium.

We will begin by giving a simple, non-technical example in which segregation is both efficient and an equilibrium. We then generalize the intuitive insights from this example to all possible models in which there are two bias types. Some insights from this model can be generalized again to all conceivable generic bias configurations with an arbitrary number of biases and players. Finally, we show that the welfare effects of polarization work despite segregation, not through segregation.

In the supplementary material, we also give lower bounds for how large polarization needs to be so that segregation becomes optimal, by considering bias configurations with large numbers of players.

4.1. A Non-Technical Example

Consider a set of biases as in panel (i) of figure 2: 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 any equilibrium none of them reveals any information. The number of pieces of information generated is 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. 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 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.

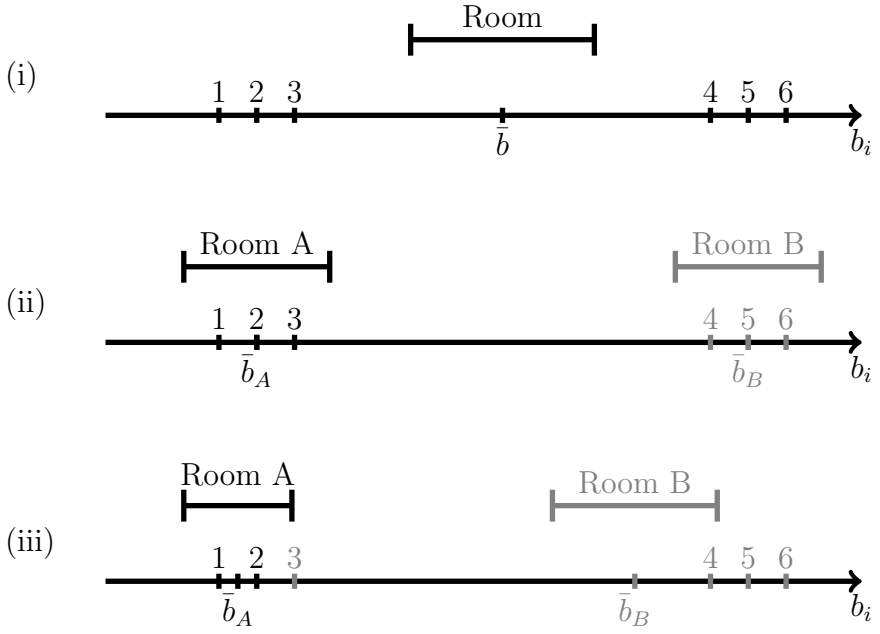

Figure 2: Truth-telling intervals for (i) the fully integrated room, (ii) two segregated rooms, (iii) player 3's best deviation from the segregated room.

4.2. 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. We can make this more precise in the following result, which emerges from the detailed derivations in the supplementary material:

Result 1. *If all $b_i \in \{0, b\}$ and the welfare-optimal room allocation is not an equilibrium, then the welfare-optimal equilibrium allocation involves too little segregation, i.e. welfare could be improved by moving players from mixed rooms into rooms that contain only their own bias type.*

Intuitively, if segregation is welfare optimal, players might have an incentive to switch to the room which contains players with the opposite bias 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 3.

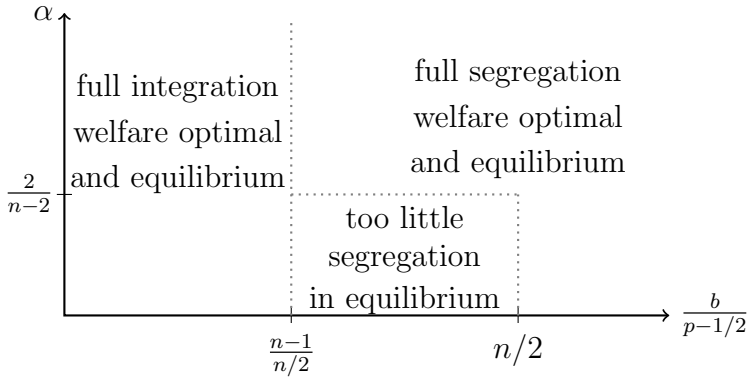

Figure 3: Welfare and equilibria for equally sized bias groups.

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.[13] 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 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 4 schematically illustrates welfare optimal and equilibrium room allocation. We refer the reader to the supplementary material for a full analysis.[14]

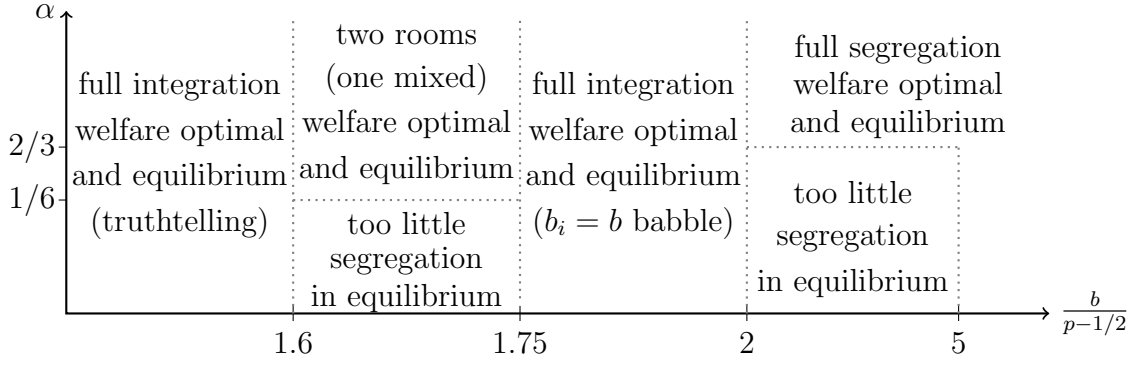

Figure 4: Welfare and equilibria when $n_0 = 5 > 4 = n_b$. (not to scale)

[13] That 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.

[14] There 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.

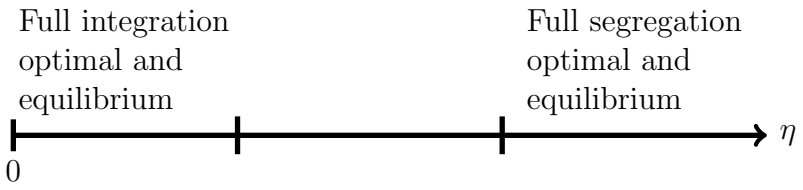

Figure 5: Welfare-optimal allocations that are also equilibria for large and small η.

4.3. When is Segregation Optimal?

The previous section has shown that integration and segregation are, respectively, optimal if preferences are little polarized or very polarized. We can generalize this insight to arbitrary bias configurations with arbitrarily many biases. Let $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$ be a bias configuration. (Note that this is not a set, as several people can have the same bias.) Assume that $\mathcal{B}$ is generic in the sense that no bias is the average of any set of other biases (except in cases where several people have the same bias).[15] Now we can consider an alternative bias configuration $\mathcal{B}_\eta$, with $\eta \in (0, \infty)$, which for every b_i in $\mathcal{B}$ contains ηb_i. Intuitively, η parameterizes the polarization of preferences compared to the differences in information between players.[16] 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 for bias configuration $\mathcal{B}_\eta$.*

(ii) If η is sufficiently large, full segregation by bias types is generically welfare-optimal and a room-choice equilibrium for bias configuration $\mathcal{B}_\eta$. (Proof on page 39.)

Figure 5 summarizes the result. We can intuitively explain it in the following way: If biases are clustered very closely relative 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.

[15] More precisely, the assumption is that $b_i \neq \sum_{b_j \in \mathcal{B} \setminus \{b_i\}} \frac{\tilde{n}_{b_j}}{\sum_k \tilde{n}_{b_k}} * b_j$ for any vector of $\tilde{n}_{b_j} \in \{0, 1, \ldots, n_{b_j}\}$ where n_{b_j} is the number of players with bias b_j.

[16] One way to think about η is as the polarization measure proposed by Esteban and Ray (1994) (theorem 1) with an appropriate scaling parameter.

Note that in this general case, a welfare-optimal equilibrium that is not the welfare-optimum may involve too much as well as too little segregation.[17] We give an example for a case in which there is too much segregation in equilibrium in the supplementary material.

4.4. 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-parameterization under which we derived our general results on integration and segregation in section 4.3, we can show that both welfare and the amount of communicated information are weakly decreasing in η, i.e. our measure of polarization.

Proposition 2. *Denote expected welfare in the welfare optimal room assignment with bias configuration $\mathcal{B}_\eta$ by $W(\eta)$ and the total number of pieces of information in the welfare optimal room assignment by $\mathcal{Z}(\eta)$. $\mathcal{Z}(\eta)$ and $W(\eta)$ are both decreasing in η. (Proof on page 40.)*

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 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.

[17] Also note that the notions of "too much" or "too little" segregation may not necessarily be well-defined if there are arbitrarily many bias groups.

5. Extensions: Uncertainty, Public Information, and Follower Networks

This section considers three extensions to our model. We describe each extension and present the results. All derivations and further details are, however, relegated to the supplementary material.

5.1. Public Information

Besides the private information they get from σ_i, players may also have access to common, public information that is relevant for their decision. Assume that instead of our usual assumption on θ, it was now $\theta = \tau\theta_0 + (1-\tau)\sum_{k=1}^{n}\theta_i$. In addition to the private signals σ_i that give player i information about θ_i, there is now also a public signal of accuracy p_0 that is informative about $\theta_0 \in \{0,1\}$. The parameter $\tau \in [0,1]$ gives the relative importance of public information. Our main interest in this extension is the comparative static with respect to τ: Will more public information, for example due to progress in information technologies or higher-quality news outlets, lead to more or less informative communication and will this segregate society more or less?

The main mechanisms of our model remain unchanged in such an extension. Public information, however, crowds out incentives to tell the truth: If we increase the importance of public information τ, it becomes more tempting to mislead players with different biases. As private information is relatively less important for high τ, other players respond less strongly to one's message (if the message is believed to be truthful) and consequently players are less disciplined by the danger of misleading their audience "too much".

Formally, we can show that the truth-telling interval within any room (the equivalent to the interval from theorem 1 above) is

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

The length of this interval is decreasing in τ, which means that for larger τ less players in a given room are truthtelling. Consequently, it is rational and efficient to segregate more if τ is higher. In particular, there exists a $\bar{\tau} < 1$ such that full segregation is optimal and an equilibrium for all $\tau \geq \bar{\tau}$.

This suggests an additional mechanism for why segregation occurs and how it may differ over time and between settings. In communication settings, both private or professional, where almost all relevant information is private information of the participants, it may be easier to achieve communication and hence segregation is less useful. But when the discussion is about politics, for example, where almost all information is public and people's private knowledge and experiences are only a small facet of a larger whole, more segregation may be desirable.

The results also suggest that progress in information technologies, which make information publicly accessible that in earlier times was held only by experts, will lead to less

(truthful) private communication and more segregation. Note, however, that this does not necessarily imply that players make less informed decision as the additional public information can more than outweigh the informational loss from less private communication.

5.2. Uncertainty

So far, we have assumed that all biases b_i are common knowledge. This may not always be the case, especially in environments where communication is somewhat anonymous, such as on the internet. In such cases, it seems reasonable to assume that both the state of the world and the types of all players are subject to uncertainty.

Assume that all biases b_i are 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. The main results of our model generalize to this setting with a few modifications. Players' motivations to tell the truth, similar to theorem 1, now depend on the distance between a player's realized bias and $\bar{b}^e$, the average of the expected biases of all players in the same room. Ex ante, the probability with which player i tells the truth hence depends on how likely it is that the realization of b_i lies within that interval around $\bar{b}^e$. An increase in mean-preserving uncertainty can increase or decrease truth-telling, depending on whether it shifts probability mass of b_i into the relevant interval around $\bar{b}^e$ or out of it. In general, however, we can show for several partial orderings of uncertainty that a sufficiently large increase in uncertainty will eventually erode all truth-telling. Such uncertainty would be most prevalent in anonymous, one-shot interactions such as in online public comment sections.

Uncertainty also has implications for whether segregation is efficient and individually optimal. Consider two bias groups (as in section 4.2 above) that are close enough to each other so that full integration is optimal and an equilibrium. Even a small increase in uncertainty can drastically reduce how much information is exchanged in the fully integrated room, as players are now with a high probability too far from the average bias in the room to tell the truth. Segregation, however, may restore much of the information exchange (or even full truth-telling) in two segregated rooms. This may be welfare-optimal, especially given that the benefits of truth-telling are not linear in its probability.[18]

5.3. (Overlapping) Follower Networks

Our main model restricts how players can associate by only allowing players to join exactly one room, and only to communicate with the other players in that room. We can

[18] To illustrate this point, consider a player with a relatively low bias who truthfully reveals σ^l half of the time. This means that 75% of the time, he sends the relatively uninformative message l. His messaging strategy thus partitions the state space much worse than truth-telling. This means that listening to two players who tell the truth "half of the time" in this way reduces the variance of one's belief less than listening to one players who fully tells the truth.

soften this assumption by considering a modified model that allows a freer choice of whom to learn from. Imagine that instead of the room choice stage, all players decide simultaneously to "follow" as many of the other players as they like. In other words, players create a directed communication network where links can be unilaterally created by the receiver of messages. In the communication stage, players then each send one message that is received by all of their followers.

Many of our main results carry over to this extension in a modified way. Similarly to theorem 1, a player will now tell the truth if and only if his own bias is in the symmetric interval $\left[\overline{b} - \frac{n_{F_i}-1}{n_{F_i}}(p-\frac{1}{2}),\ \overline{b} + \frac{n_{F_i}-1}{n_{F_i}}(p-\frac{1}{2})\right]$ around the average bias of his followers. (n_{F_i} is the number of followers that i has.) This means that player i always wants to follow player j unless the very act of following makes j babble. This feature of the best response implies that the notions of most informative equilibrium and welfare optimal follower-assignment coincide. We can again show that if polarization increases, segregation becomes more desirable and it becomes optimal for players to segregate more. If polarization is low, it is efficient and an equilibrium for everyone to follow everyone – similar to the fully integrated room in our main model.

This extension has the interesting feature that there are differences between players with moderate and extreme preferences in how isolated they are from others. Players with moderate preferences can in equilibrium be followed by much of the population but still tell the truth, because different players' influences on $\overline{b}$ at least partially neutralize each other. Extremist players, however, can only be followed by other extremists of the same persuasion, as they would babble if followed by too many moderates or even by extremists at the other end of the spectrum.

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 280 characters 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. Consequently, a given sender addresses different audiences when replying to different people.

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

Democrats	Republicans
endgunviol	ccp
trumpshutdown	rubio
actonclim	arkansa
protectourcar	schiff
defendourdemocraci	hawley
forthepeopl	communist
climatecrisi	prolif
justiceinpol	chuckgrassley
getcov	oklahoma
lgbtq	hoosier

Democrats	Republicans
protectourcar	kssen
endgunviol	arkansan
trumpshutdown	countymeet
actonclim	nevergiveup
defendourdemocraci	bornal
lowerdrugcost	nebraskan
climateactionnow	dakotan
whatsatstak	secureourbord
equalpay	republicanstudi
homeisher	buildthewal

Table 1: Left: Words with most partisan usage difference among the words that were used very often (more than 1000 times) in our sample. "ccp" is an abbreviation for "Chinese Communist Party".
Right: Most partisan words among words that were used at least 10 times in our sample. "kssen" is an abbreviation for the senator of Kansas. (Note that these expressions are stemmed.)

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.[19]

6.1. Preliminary Steps

First step: Building a dictionary We analyzed the tweets of all members of the 116th and 117th U.S. Congress to build a dictionary of partisan monograms (i.e. words) and bigrams (groups of two consecutive 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.[20]

Table 1 has some examples for partisan words. Note that the differences in usage might derive from using different words for the same thing (e.g. in our sample 80% of those referring to Donald Trump's Twitter handle "realdonaldtrump" are Republicans while 80% of those referring to "trump" are Democrats) or from different focuses (talking about "ending gun violence" vs talking about "Chinese Communist Party"). We are agnostic about where the differences come from.

[19]Further details concerning data collection and scoring are provided in the supplementary material.

[20]In order not to over-extrapolate from small samples and also to restrict the size of our dictionary, we only use monograms/bigrams that are used at least 50 times by members of Congress and that are used at least twice as often by the members of one party than by members of the other party.

Second step: Scoring accounts Armed with this partisan dictionary, we can identify a person's political leanings purely based on how similar their Twitter feed looks to that of a Democrat or a Republican member of Congress. For each monogram/bigram that this person uses in his original tweets and which is found in our dictionary, we assign a score based on how differently the term is used between members of Congress. In the end, we arrive at an overall score for that person, based on all partisan terms they have used. We do this for mono- and bigrams separately and construct the overall score by averaging between the mono- and the bigram score.

To check whether the scoring method that we have constructed returns sensible (out-of-sample) results, we scored the Twitter accounts of journalists and pundits who were popular with either the American left or right.[21] If our scoring method works well, we should be able to separate these Twitter accounts into partisan camps, purely based on their word usage. Table 6 on page 42 of the appendix shows that we are indeed able to do so with more than 85% accuracy.

Third step: Sampling random Twitter users We randomly sampled a number of Twitter users who (i) tweet from inside the geographic areas of the U.S., (ii) had tweeted a tweet containing one of the words "Trump", "Biden" or "Congress" during a week in February 2021 (iii) have written at least 500 tweets of their own (and not counting replies and retweets), and (iv) have written replies to at least two people and at least 20 replies in total.

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 $[0, 1]$. A user who only tweets words that are only ever used by Democrats would receive score 0, while a user who only uses words that are only used by Republicans would receive the score 1. A user who uses words used by both sides receives a score that is based on how often she uses each word and how partisan the usage of that word is.

Fourth step: Making use of differences in audience composition When "tweeting", users' texts are read by different audiences, based on what type of tweets they are.[22] Simple tweets by user X are shown in the timelines of all users who follow X. A reply by user X to user Y can be 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.

[21] We used the list of the most influential journalists and bloggers on the right and left, respectively, from StatSocial (2015).

[22] See here for how Twitter itself describes visibility: https://help.twitter.com/en/using-twitter/types-of-tweets *(Accessed: May 04, 2021)*.

6.2. Difference-in-differences analysis

Using our work from the previous steps, we generated a data set containing 149,161 reply tweets sent from 2,418 senders to 30,075 receivers.[23] 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. Formally, we will use OLS to estimate equations of the following form:

$$\text{property}_i = \beta \left|\text{score}_{S(i)} - \text{score}_{R(i)}\right| + \text{FE}_{S(i)} + \varepsilon_i$$

where property_i is the property of interaction i that we are interested in, $S(i)$ and $R(i)$ are the sender and receiver in interaction i, respectively, $\text{FE}_{S(i)}$ is a fixed effect for $S(i)$, and ε_i is an error term (we cluster error terms at the sender level).

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 immediately obvious 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 would expect three responses:

1. Cheap talk communication: People who find it hard to communicate with ideologically distant others will communicate more with those who are ideologically close.

2. Emotional communication: Instead of trying to exchange information across an ideological divide, people may simply communicate to express emotions and provoke each other (or react to provocations).

3. Hard evidence: 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.

[23] These are observations for which both mono and bigram scores and therefore also the average of the two exist. We have more observations if we replicate our analysis using either only bigrams or only monograms. All our results remain qualitatively unchanged when doing so.

While the first response is a direct implication of our model, the second and third go slightly beyond (though the supplementary material to this paper contains an extension of our model that allows for the provision of hard evidence). We will discuss each of these three effects in turn, and show that they are all present in our data.

Cheap talk communication We show that there are fewer interactions across the ideological spectrum, compared to interactions between people with similar ideology. This is what our model predicts. The positive relationship means that the more right-wing a Twitter user is, the more they will on average interact with other right-wing Twitter users.[24] 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.[25] We also see, of course, that there is a huge amount of unexplained noise in our dataset – which is not surprising, given that we consider *all* interactions by people who have at some point used a potentially political term, and make no further pre-selection into our dataset.

	receiver score
(Intercept)	0.418***
	(0.019)
sender score	0.134***
	(0.040)
Estimator	OLS
N	149,161
R^2	0.008

Table 2: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

Emotional communication 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 an emotional and negative message ("letting off steam"). We can check for evidence of such behavior by examining the interaction between its emotional content and the ideological distance between sender and receiver. We first measure a tweet's emotional content using the sentiment dictionary by Hu and Liu (2004), which

[24]Of 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.

[25]Of course, we are not the first to show segregation on Twitter – see, for example, the studies by Barberá et al. (2015) or Krasodomski-Jones (2017).

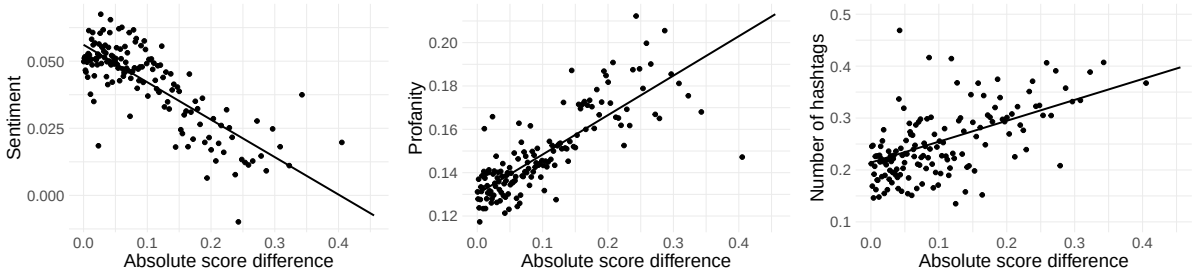

Figure 6: A larger (absolute) difference between sender and receiver score is associated with a lower sentiment score and higher usage of profanity and hashtags. (Binned scatterplot, one dot corresponds to roughly 1000 observations which are grouped according to absolute score difference. One outlier removed in the third plot.)

gives scores to certain words and phrases that mark positive or negative content.[26] The left panel in Figure 6 illustrates that a larger ideological distance between sender and receiver is associated with a more negative sentiment. The central panel of the same graph shows a strongly positive relationship between ideological distance and the use of profanity in interactions.[27] Finally, hashtags are often used in a declarative, emotional fashion to "make a point". (For example, our dataset contains hundreds of examples of accounts simply replying "#fakenews" at accounts they – presumably – disagree with.) The right-hand panel in the figure shows that there is also a positive relationship between ideological distance and hashtag use. Table 3 gives the exact regression results.

	sentiment	profanityCheck	hashtags
	(1)	(2)	(3)
absolute score difference	-0.104***	0.127***	0.238***
	(0.008)	(0.013)	(0.043)
sender fixed effects	Yes	Yes	Yes
Estimator	OLS	OLS	OLS
N	149,085	149,129	149,129
R^2	0.086	0.135	0.388

Table 3: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

Hard evidence communication 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 "I think you are wrong". "Look at this

[26] Our results are robust to using other sentiment analyzers like the popular VADER-Sentiment introduced by Hutto and Gilbert (2014).

[27] We measure the presence of profanity by the Python package profanity-check; see `https://pypi.org/project/alt-profanity-check/` for details. (*Accessed: May 27, 2021).*

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, that a reader may be unconvinced that our theorem 1 is correct until they have read the proof on page 36. 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.[28]

We can measure such attempts at persuasion (as opposed to cheap talk) by considering whether reply tweets grow longer, more complex (measured by average word length[29]) and contain more hyperlinks as the score difference between sender and receiver increases. Table 4 shows that this is indeed the case.

	tweet length	word length	links
	(1)	(2)	(3)
absolute score difference	10.991*	0.222***	0.044***
	(4.748)	(0.030)	(0.008)
sender fixed effects	Yes	Yes	Yes
Estimator	OLS	OLS	OLS
N	149,129	145,051	149,129
R^2	0.275	0.148	0.251

Table 4: Tweets get longer, more complex, and contain more hyperlinks as the ideological difference between sender and receiver increases. (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

Followers The extension of section 5.3 allowed people to "follow" several senders, as is e.g. possible on Twitter. In equilibrium, this means that individuals with extreme biases have fewer followers than those with moderate biases. Table 5 shows that this is indeed the case in our dataset. However, the difference in mean score between moderates and extremists is only statistically significant for those with a very right-wing score. The stronger results for right extremists can be explained by the distribution of scores, which has positive skew – see figure 7. Consequently, the average score distance from, for example, the mean is higher for right than for left extremists. The mechanism restricting the audience of

[28] One might wonder whether hard evidence could crowd out the cheap talk communication we model. However, one should not forget that cheap talk is easy while finding and providing hard evidence is costly, e.g. finding a convincing report by the national statistical office takes time (and such a report may even not exist). Because of its lower cost, cheap talk communication is therefore preferable as long truthful reports are possible. If we added costly hard evidence provision to our model, it would only be relevant in those cases where (i) cheap talk communication is impossible, (ii) hard evidence is not prohibitively costly to obtain and (iii) players care sufficiently about other players' actions. We formally show this intuitive conclusion in an extension to our model in the supplementary material.

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

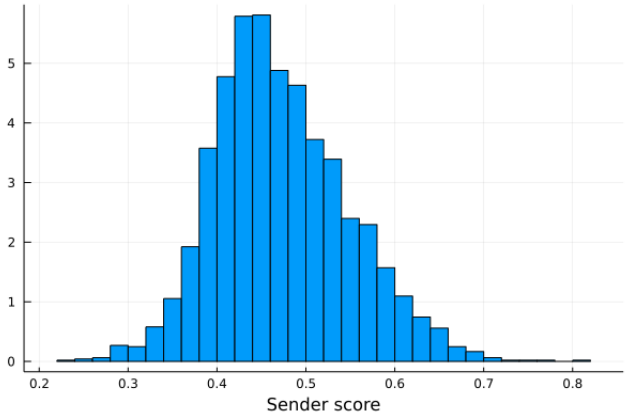

Figure 7: Histogram of sender scores (normalized).

extremists in our model is that more moderate followers (or, even worse, followers from the other extreme) would destroy the truthtelling incentives of an extremist. Given the positive skew in sender scores, this mechanism is stronger for right than for left extremists.

	extreme left	moderates	extreme right
average followers	4218.17	5730.42	2511.24
group size	344	1680	394
p-value Welch's t-test	0.352		0.027*

Table 5: Mean number of followers for senders with a score below the mean score minus one standard deviation (extreme left), score above the mean score plus one standard deviation (extreme right) and the rest (moderates). p-values of Welch's t-test that each extreme group has the same mean number of followers as the moderates. (significance levels: * 5%, ** 1%, *** 0.1%)

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 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 subgroup 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.[30] This has a strict economic logic to it: Once the Internet 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. Additionally, the resulting group structures are much less portable than files or contact lists (and often not at all), thus contributing to networking sites' market power.

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

[30]Facebook, for example, allows its users among other things to (i) choose which of their data is visible 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 as well as connection.

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. Room Choice as Communication Design

A large literature has recently analyzed the problem of designing socially optimal information structures – see, for example, Bergemann and Morris (2019). Such "information design" commonly assumes that a designer can set a rule by which messages about private information are chosen. Alternatively, players may themselves be able to commit to such a disclosure rule, which allows them to communicate truthfully despite a conflict of interest with the receiver (as in models of "Bayesian Persuasion"). Any such design therefore requires that players can either be forced to follow such rules, or that rule-breaking can be monitored and punished. But in some settings, no commitment, monitoring or punishment may be available.

Our model shows that truthful communication can still be made possible even between people who prefer lying to each other, if there are other people in the same room to whom both players want to tell the truth. Crucially, room composition acts as a commitment device by making players *want* to tell the truth, which means that no objective mechanism to later compare their messages to the truth is needed. The tools we have developed in sections 2.2 show how and when such "communication design" is possible.

The term "communication design", however, should not be understood to mean that a designer is always needed. As we have shown in section 3, players can often sort into an efficient allocation themselves (though they may need help in coordinating on one of many equilibria).

7.4. When are echo chambers bad?

Our argument that echo chambers can be useful does not necessarily mean that they are beneficial on balance and in every setting. Besides the mechanism that we analyze, echo chambers may have many other effects. Some of them can be informational, some behavioral, and some may only occur in settings that are slightly different from ours.

While we believe that the mechanism we describe is very general, a complete assessment of echo chambers in a given context may well conclude that our mechanism is present but outweighed by other, detrimental effects. We will discuss some such potential effects; there are of course others and it is outside the scope of this paper to provide a full assessment of all effects that echo chambers can have.

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 credible 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.[31] 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.

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.[32] 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.

Endogenous Polarization One could also assume that preferences, which we take as given in our model, are actually the result of an endogenous process that depends on whom each person communicates with. Imagine, for example, that the game in this paper is played several times in a row, and between stages everybody's preferences move closer to the average of the preferences of the people they communicated with in the most recent stage. Segregated communication could then lead to further polarization, as the preferences of people who are in different rooms move further and further apart. As long as this process requires actual communication, however, some segregation may still be optimal, and such endogenous polarization would simply add another trade-off between getting people to communicate with each other (which is better than having them all babble) and causing

[31] We 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.

[32] C.f. the experimental work by Kallir and Sonsino (2009) and Eyster and Weizsäcker (2011) on "correlation neglect".

further polarization down the line. In the long run, we might expect a stabilization of preferences around a few points, and consequently segregation into rooms, in the welfare optimum of such a repeated game.

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.[33] Even taste-based homophily can end up improving everyone's information.

Malicious actors. Our model is optimistic in the sense that while players want to mislead each other, they have an abstract interest in a well-informed society since it reduces the variance of people's mistakes. In reality, online "bots" and "trolls" may be interested in simply increasing uncertainty and chaos, both on behalf of state- and non-state actors. Similarly, (social) media companies may find that misinformation creates engagement even though it does not decrease (or even increases) the variance of people's actions. In both of these cases, segregation into echo chambers could help these actors in spreading misinformation, in particular if receivers cannot correctly deduce the motivations of a message's sender.

8. Conclusion

Modern democratic societies have three main mechanisms to aggregate information: Debates, 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

[33] Policies 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.

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 this cannot 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 groups in which communication works, in favor of large integrated groups in which it does not. Our research suggests that meaningful debate can only happen if the participants feel that they have enough in common and they trust each others' motives. Debate is more than putting people into a room and expecting them to come out smarter.

Appendix

A. Proofs

This appendix contains only the proofs for results that are explicitly given in the main text; all other results and their proofs can be found in the supplementary material.

Proof of lemma 1 on page 11.

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

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

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

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

Specifically, sending message m^h gives expected payoff

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

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

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

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

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

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 $\mathbb{E}\left[\theta_i | \sigma_i\right]$ varies in 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, $\Delta U_i(\sigma^h) = 0$ implies $\Delta U_i(\sigma^l) = 2(\mu^h_{ji} - \mu^l_{ji})(n_{R_i} - 1)(1 - 2p) < 0$ and therefore $m_i(\sigma^l) = m^l$ which implies $\mu^h_{ji} = p$ as a m^h is only sent by i after receiving signal σ^h. Consequently, $(\mu^h_{ji} + \mu^l_{ji})/2 \geq 1/2$ as $\mu^l_{ji} \geq 1 - p$. Now consider the equilibrium candidate (m^t_i, m_{-i}). With the truthful strategy m^t_i, $\mu^{th}_{ji} = p$ and $\mu^{tl}_{ji} = 1 - p$ and therefore $(\mu^{th}_{ji} + \mu^{tl}_{ji})/2 = 1/2$. This implies that $\Delta U_i(\sigma^h) > 0$ in the equilibrium candidate (m^t_i, m_{-i}), i.e. truthful reporting is optimal for i after receiving signal σ^h. In the equilibrium candidate (m^t_i, m_{-i}), truthful messaging is still optimal after signal σ^l as well: From $p > 1/2$, $\mu^h_{ji} \leq p$ and $\mu^l_{ji} \leq 1/2$ it follows that $-1/2 + (1 - p) < -\left(\mu^h_{ji} + \mu^l_{ji}\right)/2 + p$. As in the original equilibrium (m_i, m_{-i}) we had $\Delta U_i(\sigma^h) = 0$ and therefore $-\left(\mu^h_{ji} + \mu^l_{ji}\right)/2 + p = \sum_{j\in R_i, j\neq i} b_j/(n_{R_i} - 1) + b_i$, 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^t_i, m_{-i}). Hence, truthful messaging is i's best response in the equilibrium candidate (m^t_i, 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^t_i. Hence, (m^t_i, m_{-i}) is an equilibrium.

The argument in case i's strategy is mixed after signal σ^l is analogous. □

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 $\mu^h_{ji} = p$ and $\mu^l_{ji} = 1 - p$ 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 $\Delta U_i(\sigma^l) \leq 0$ and get the interval used in the theorem. □

Proof of proposition 1 on page 13.

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

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

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

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

Now note that $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right]$ can have two possible values: If $k \in R_j^{truth} \cup \{j\}$, i.e. if j has received information about θ_k, then $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right] = p(1-p)$. If j has not received information about θ_k, then $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right] = \frac{1}{4}$. (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

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

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

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

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

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

and express welfare as

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

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

Proof of theorem 2 on page 18.

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 $\left|\sum_k \{b_k\} - n b_i\right| / (n-1) \leq \frac{1}{2}$. 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. □

Proof of proposition 2 on page 19

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 $B_{\eta'}$ 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. □

screen name	score
RBReich	*0.339*
MHarrisPerry	*0.347*
ariannahuff	*0.37*
DavidCornDC	*0.379*
TheRevAl	*0.39*
ChrisCuomo	*0.406*
ezraklein	*0.407*
donnabrazile	*0.422*
NateSilver538	*0.429*
anamariecox	*0.429*
paulkrugman	*0.429*
sullydish	*0.431*
CharlesMBlow	*0.431*
camanpour	*0.432*
Lawrence	*0.433*
HardballChris	*0.435*
maddow	*0.441*
jdickerson	*0.447*
markos	*0.449*
KirstenPowers	0.457
AnnCoulter	0.457
NickKristof	*0.458*
chrislhayes	*0.471*
KatrinaNation	*0.471*
costareports	0.474
nycjim	*0.479*
stephenfhayes	0.484
MajorCBS	0.497
mkhammer	0.498
megynkelly	0.507
brithume	0.507
EWErickson	0.508
secupp	0.509
greggutfeld	0.516
ggreenwald	*0.538*
mtaibbi	*0.539*

FareedZakaria	*0.54*
seanhannity	0.552
jaketapper	0.557
RichLowry	0.56
michellemalkin	0.567
glennbeck	0.574
DLoesch	0.583
greta	0.585
AHMalcolm	0.594
ericbolling	0.599
TeamCavuto	0.602
DanaPerino	0.617
Peggynoonannyc	0.639
kimguilfoyle	0.646
edhenry	0.686
MonicaCrowley	0.73

Table 6: Scoring of most influential journalists and bloggers on the right and *left* according to StatSocial (2015).

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. National Bureau of Economic Research Working Paper, No. 28884.

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.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

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: 2021-07-02.

Che, Y.-K. and K. Mierendorff (2019). Optimal dynamic allocation of attention. *American Economic Review 109*(8), 2993–3029.

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.

Esteban, J.-M. and D. Ray (1994). On the measurement of polarization. *Econometrica 62*(4), 819–851.

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.

Grimes, D. R. (2017). Echo chambers are dangerous – we must try to break free of our online bubbles. https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles. Accessed: 2021-07-02.

Hagenbach, J. and F. Koessler (2010). Strategic communication networks. *Review of Economic Studies 77*(3), 1072–1099.

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: 2021-07-02.

Hu, M. and B. Liu (2004). Mining opinion features in customer reviews. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, Volume 4, pp. 755–760.

Hutto, C. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 8, pp. 216–225.

Itten, A. (2018). Coming undone: How echo chambers balkanised society. https://www.politics.co.uk/comment-analysis/2018/08/16/coming-undone-how-echo-chambers-balkanised-society. Accessed: 2021-07-02.

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.

Krasodomski-Jones, A. (2017). Talking to ourselves. https://www.demos.co.uk/project/talking-to-ourselves/. Accessed: 2021-07-02.

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.

Martinez, G. and N. H. Tenev (2020). Optimal echo chambers. arXiv preprint arXiv:2010.01249.

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: 2021-07-02.

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.