Political Debate on Social Media: Theory and Evidence*

Ole Jann — CERGE-EI

Christoph Schottmüller — University of Cologne and TILEC

March 14, 2024

Abstract

We investigate online political debate with a theoretical model and an original, large-scale dataset. In our model, debaters can use several types of strategic communication but also derive "expressive utility" from speaking their mind. We examine how social media users try to convince others and "win" debates by deploying arguments, hyperlinks, media and different styles of language. Our empirical analysis considers almost 140,000 Twitter interactions between users whose ideological stance we can estimate, using a novel methodology. We use our model to interpret this data and document patterns that are consistent with the predictions of the model.

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

Keywords: asymmetric information, polarization, debate, cheap talk, information aggregation, social media, 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. Some of the results in this paper were previously part of our working paper "Why Echo Chambers are Useful" but are now removed from there. We are grateful for helpful comments by Rachel Bernhard, Matteo Escudé, Satoshi Fukuda, Paul Klemperer, Vasily Korovkin, Niccolò Lomys, Meg Meyer, David Ronayne, Larry Samuelson, Peter Norman Sørensen and Peyton Young, as well as audiences at Bielefeld, CERGE-EI, East Anglia, Groningen, Konstanz, Munich, Oxford and Sussex and at EEA 2018 (Cologne), ESEWM 2018 (Naples), ECOP 2019 (Bologna), ASSA 2021 (virtual), EPSA 2022 (Prague), SAET 2023 (Paris) and ITES 2023 (Prague). This research was supported by Charles University Primus project 20/HUM/019 and GACR project 22-33162M.

Democratic societies need the exchange of information and opinions that happens in political debate and discussion. At least starting with the ancient Athenian democracy and via Mill (1859) to Arendt (1958) and Habermas (1981), philosophers and social scientists have emphasized the importance of debate for democratic societies.[1] Much focus has been on the necessary conditions for productive debate – for example, that it needs to take place between people on roughly equal footing and without the threat of coercion or violence, that it must be open to people with different viewpoints and must adhere to other broad principles of freedom of speech.

When people debate on social media, these features are mostly in place – more so than they have been at most other points in human history. But actual political debates on social media are often experienced as vitriolic and unproductive – a view that is intensifying: For example, large majorities of Americans state that "the tone and nature of political debate" has become more negative in recent years.[2] Millions of people have been harassed online for their political views[3]; in a survey on the toxicity of social media sites, none was rated below "medium toxicity"[4]. All of this, on platforms that are free, easily accessible and free from direct coercion, suggests that the dysfunctionality of online debate is due to factors that are inherent to the motivations and the consequent behavior of humans.

This paper combines theoretical and empirical methods to examine political debate on social media. In our theoretical analysis, we consider a stylized model of political debate in which participants are interested in gathering and transmitting information as well as in scoring partisan points, while also deriving direct payoffs and costs from expressing their opinions. Our empirical analysis considers a large set of interactions on the social media site Twitter to show patterns that are consistent with the central predictions of our theoretical work.[5] This combination of methods allow us to speak to how the motivations of self-interested debaters interact to create various patterns which we can then document in the real-life data.

We think that these methods complement each other to give a deeper understanding of the subject matter. The expansive and highly complex real-life data cannot be understood without first clarifying what to look for and which mechanisms to expect – which requires theoretical analysis. At the same time, the theoretical analysis of political debate needs applied impulses to clarify what its predictions are – and then can be informed by whether

[1] We use "debate" and "discussion" interchangeably, also to refer to what other works have called "deliberation". We briefly comment on this usage of words in our discussion of related literature. At times, it can be almost ironic how little debate there is about *whether* debate is good – though see e.g. Conover et al. (2002), p. 22, for a brief review of critical literature.

[2] https://www.pewresearch.org/short-reads/2019/07/18/americans-say-the-nations-political-debate-has-grown-more-toxic-and-heated-rhetoric-could-lead-to-violence/

[3] https://www.pewresearch.org/short-reads/2017/07/11/key-takeaways-online-harassment/

[4] https://simpletexting.com/most-toxic-social-media-apps/

[5] Our data was collected in February 2021, before the major changes in ownership and user structure at Twitter that began in 2022. Twitter has since been rebranded as X and has lost parts of its user base.

these predictions receive support from empirical analysis. No part of our analysis should be read on its own, and we see part of our contribution in our connection of methods – though, by necessity, we describe our methodological approaches one after the other.

Our analysis has three scientific contributions. First, we construct a model that conceptualizes how debate actually works when people have the varying and occasionally self-contradictory motivations outlined above. Second, this helps us understand the functions and effects of real-life communication tools such as verbal arguments, hyperlinks or profanity. Third, we show how the mechanisms and results of such a highly abstract model can be connected to real-life evidence collected directly from a social media site.

The ultimate goal of our analysis is to contribute to our understanding of how political debate works, how different methods can be combined to analyze it, and what may determine whether debate leads to a genuine change of information. This is of special interest considering that political debate on social media in many ways fulfils the oft-stated ideals of political deliberation of publicity, non-tyranny and equality [6] yet is not widely viewed as well-functioning or effective. The reasons for that, we suggest, lie in the debaters' varying motivations themselves – which is an issue that is inherent to all debate among human beings who are strategic actors. This, in turn, means that seemingly intuitive policies to improve debate may be ineffective or counterproductive if they do not proceed from an understanding of these aspects of debate.

Policies aimed at curbing aggressive speech, for example, may be counterproductive because they also make the *absence* of aggressive speech less informative – as we argue below. Giving easier access to supporting sources and media (or even using artificial intelligence to provide or check evidence) could similarly have indirect effects that contradict the direct ones.

The remainder of this introduction discusses our theoretical and empirical analysis in more detail and connects it to other studies.

A summary of our theoretical analysis To understand how different motivations may interact when people engage in debate, in section 1 we analyze a simple one-shot exchange between two individuals who are discussing an issue. One of them has stated an opinion; the other has information that the first does not have and is considering whether and how to convey it. In keeping with the conventions of models of strategic information transmission, we call the latter person the sender (S) and the former the receiver (R).

S and R are separated by some distance between their ideological stances. The larger that distance, the more S is interested in simply pushing R's view of the world in a certain direction (and thus "winning" the debate) rather than actually revealing any information (which could influence R's opinion in either direction). We hence think of this ideological distance as being akin to a preference: Even though someone's ideological position may

[6] As summarized by Conover et al., 2002, cf. also the "ideal speech situation" of Habermas, 1983.

shift over time, in a single interaction it remains approximately fixed. Besides caring about R's information and ex-post-view, S may also derive direct payoff from (or pay a direct cost for) expressing herself; this depends on the tone of S's message as well as the ideological distance between S and R.

We try to capture some of the complexity of real debates by assuming that S can make several choices when communicating with R. In particular, she can choose (i) whether and how to express her opinion about the world, (ii) whether to exert effort to support her opinion by arguments or references to evidence, and (iii) whether to freely show her emotions about R's position or to alter her tone. She can do the latter by either "biting her tongue" when she really feels like expressing her anger at R, or conversely choosing a sharp tone when she and R are mostly in agreement.

Technically, our model thus combines the assumptions and ideas of three different modeling approaches: Cheap talk (i.e. communication through common interest), signaling (i.e. communication through costly messages) and expressive utility (i.e. communication for expressive reasons, not to convey information).

Our main result is that S's optimal choice (and hence also how R forms beliefs) varies in the ideological distance between the two, with some monotonicity and some non-monotonicities. The main monotonicity is that truthful communication becomes harder as distance increases. At low distances, S and R can engage in simple cheap talk communication, in which S makes a statement about the world and R believes her. At larger distances, truthful information exchange in equilibrium is only possible if S makes some costly effort, either by incorporating "evidence", like complex verbal arguments or references to outside sources, or by changing the message's tone in a way that she would otherwise prefer not to do.

Crucially, we do not need to assume that complex arguments or hyperlinks to evidence are directly convincing (because they convey fully verifiable information). As anyone who has participated in real-life political debates knows, such an assumption would be rather questionable. Instead, such additions enhance the credibility of messages because of the *effort* involved in using them, and the information that is communicated by this observable effort. Similarly, we do not need to assume that aggressive tone directly influences whether R listens to a message or not – instead, it is again the effort involved in "biting your tongue" (or, conversely, strongly admonishing someone you feel close to) that endogenously lends some types of messages extra credibility.

Of course, our model is optimistic in some aspects – for example, we assume that all debaters are at least in principle interested in some notion of truth (though we discuss the effects of bots and trolls in section 3.1). We therefore still see our model as a limiting case of what is possible, and what is not, under the relatively stylized conditions we describe.

A summary of our empirical analysis In section 2, we examine evidence from about 140,000 interactions on the social media site Twitter.[7] We have two main goals in this analysis. First, we provide empirical content and interpretation to the abstract mechanisms of our model. Second, we document several patterns that are consistent with and support the main conclusions from our model.

Users of Twitter can use the platform to "tweet" to all of their followers or to "reply" to tweets by specific users. This difference in audiences allows us to observe interactions between specific users, as well as estimating a user's overall ideological stance by looking at the tweets they send to no one in particular.

The data that is at the core of our analysis is generated by the interactions of randomly chosen Twitter users that are based in the United States and at least occasionally discuss politics. We begin by developing a method to estimate their political and ideological stance by measuring how similar their tweets are to the tweets of contemporaneous or recent members of the U.S. Congress (whose ideological affiliation we know). When we then observe two users interacting, we take the difference of their ideological positions as the "bias" of our model: A difference in preferences that we take as given for that interaction. This ideological difference thus influences how much they care about winning over the other, and how much intrinsic enjoyment they may derive from being aggressive (or polite) towards each other.

Not all theoretical results have empirical content. In models of strategic communication, messages only acquire their meaning in equilibrium, which means that the exact same message can be an example of meaningful communication or of meaningless "babbling" – we cannot determine which is which from the data alone. But the combination of our theoretical framework and dataset allows us to make predictions about behavioral patterns, and then check for these patterns in the data.

Specifically, our model predicts that interactions between Twitter users with larger ideological distances tend to feature more hyperlinks, more complex language and longer messages – all of which are costly ways of increasing the credibility of messages when truthful communication is hard. This is also what we observe in the data. While we cannot observe directly whether the ideas or sources contained in complex arguments or hyperlinks are more convincing, our theory suggests that their ability to convince may not (or only partially) depend on their content and at least partially stem from the observable effort that was spent on deploying them. "Verifying" an argument, or a source, hence also does not require following it in every nuance or reading a linked article, but merely checking that they relate to the issue at hand in a way that suggests they had to be written or found for this particular case (i.e., at some effort).

At the same time, we also observe that interactions between users with a larger ideo-

[7] Our data was collected before the platform changed ownership, name and user base – we will hence call it "Twitter" throughout since this is what it was called at the time that our data was collected.

logical distance are more negative in tone and contain more profanity and more hashtags.[8] This is consistent with an emotional need to be angry or unfriendly that increases in ideological distance – and (as is the nature of strong language) is highly observable to the recipient. An analysis of interaction effects, however, shows that aggressive language and evidence do not tend to get deployed together – which is one of the main predictions of our model that holds across many parameter specifications.

Connection to other research Our work is related to theoretical as well as empirical approaches to debate as strategic communication, and ties into a wider literature and societal discussion on the nature and structure of political debate.

We use the term "debate" throughout to describe an exchange of viewpoints between actors who are not purely interested in information exchange, but also in expressing their opinions and being more convincing than the other. This could also be called "discussion", though that usually suggests a more open-ended conversation than the constrained back-and-forth that takes place on social media. "Deliberation" is a term that is used in particular by political scientists (cf. Bächtiger et al., 2018). While this is a much wider term it is often used to describe the type of interaction we are considering here (cf. Strandberg and Grönlund, 2018).

Philosophers and social scientists have emphasized the importance of debate and deliberation throughout history (see Chambers, 2018, and Conover et al., 2002, for overviews from Aristotle to now). While our study considers specific aspects of a specific setting, we see our work in the tradition of asking *when* debate can succeed in revealing and transmitting – which in itself may not be sufficient, but probably necessary for successful and effective debate.

Our theoretical model combines two canonical approaches to the analysis of strategic communication (i.e. the theory of communicating for instrumental reasons) with assumptions about intrinsic motivations to communicate. Our agents are engaged in a "cheap talk" style situation (in the style of Crawford and Sobel, 1982), in which some informative communication is possible if the interests of sender and receiver are sufficiently aligned. Models of cheap talk have been used widely to analyze political debate, starting with Austen-Smith (1990, 1992). Senders in our model also have access to costly messages (similar to "signals" of the literature following Spence, 1973) which allow them to credibly signal about their private information. In particular, a sender can signal by making a costly effort or not taking actions they would like to take, similar to the "money burning" of Austen-Smith and Banks (2000).

While there is a fair amount of research on people's motivations to engage in public debate, the terminology of such studies cannot always be easily adapted into a model with

[8]Hashtags are a way of adding topics to tweets, but are often used to express opinions – such as by adding the hashtag "#fakenews" when responding to someone.

rationally choosing players. The players in our model have three broad motivations when engaging in public debate: (i) changing the minds of others, (ii) transmitting information and (iii) deriving direct pleasure from certain types of expression (or deriving direct discomfort, which they try to avoid). These broadly map into the "six motivations for political discussion" that Morey and Yamamoto (2020) explore in an online survey.[9] While Conover et al. (2002), in their surveys of American and British focus groups, found that "the political motives of expressing preferences and persuading others [...] are regarded as among the least important", we should note that this refers to the motives for *engaging* in political discussion, rather than the incentives that people face within the discussion. (Still, Conover et al. also noted the "the personal pleasure of expressing their views" that some participants derived from public debate.)

Similarly, our motivations (i) and (ii) would be called "civic motivations" ("the need to gain information, express opinions and persuade others") by Gil de Zúñiga et al. (2016) in their panel survey study, whereas motivation (iii) would be among the "social motivations [...] stemming from the sheer entertainment and relational goals achieved through informal political conversations". The distinction between "intrinsic" and "extrinsic" motivations made by Lilleker and Koc-Michalska (2018) also has some similarity to the mix of direct pleasure and instrumental utility that the players in our model derive from public debate.

An overview of the literature on "expressive" utility (which has mostly been studied for voting though also for verbal expressions and corresponds closely to our motivation iii) can be found in Hamlin and Jennings (2011). Relevant for our study is also the discussion by Akerlof and Kranton (2000) on how individuals lose utility if they act against their identity. In our context, such expressive utility creates additional possibilities to send costly signals by either foregoing positive expressive utility or by experiencing actual displeasure at having to communicate in a certain style. While (to our knowledge) this mechanism is relatively novel to the literature, we see it as a natural and necessary consequence of any model where agents can observably communicate in a way that provides them with expressive utility. It would require a peculiar set of assumptions for this *not* to become another opportunity to signal about hidden information.

We begin our empirical analysis by 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.

[9]Specifically, "influence others" is our point (i); "educate oneself" and "learn about others" fall under our point (ii) and "be social" and "build relationships" give rise to our motivation (iii).

The quality and effects of political debate, on and off social media, have been the focus of widespread debate themselves in recent years. Much of the research has focused on perceived "pathological" aspects of debate, such as disinformation, polarization and segregation – see Persily and Tucker (2020) for a collection of overviews. In contrast, we mostly assume that people who engage in online debate have some serious interest in engaging with other opinions – at the same time that they want to win arguments, convince others they are right and sometimes just express their frustration and humiliate those they disagree with. The messy product of this combination of motivations is the focus of our theoretical and empirical analysis. We briefly discuss what would happen in the presence of "trolls" or "bots" in section 3.1.

By combining different theoretical approaches in one model, considering their empirical content and showing evidence for their predictions, we are also contributing to the discussion of whether and how theoretical models can help us understand real-life phenomena like debate. See, for example, Little (2023) for a discussion of different approaches and the insights they provide.

1. Theory

1.1. Model

We will start by setting out the basic assumptions of our theoretical model; the following section comments on how and why we think they are appropriate simplifications in our applied setting.

A user S (sender) can reply to another user R's (receiver) tweet. S privately observes a state of the world $\theta \in \{0, 1\}$ which is a priori equally likely to be 0 or 1. S then sends a reply message $m \in \{0, 1\}$ to R; that message can also include evidence and can use aggressive or friendly language. After observing S's message (including whether it contains evidence and which language it uses), R takes an action $a \in \mathbb{R}$. S and R differ in their ideology; their ideological difference is given by b and is common knowledge. Without loss of generality we assume that $b > 0$.

R's payoff is

$$U_R = -(a - \theta - b)^2$$

while S's payoff is

$$U_S = -(a - \theta)^2 - \mathbb{1}_e c + \mathbb{1}_a g\left(b - \hat{b}\right).$$

Here $\mathbb{1}_e$ is an indicator function that is 1 if S has included evidence and 0 otherwise. c is the cost of using evidence. $\mathbb{1}_a$ indicates whether S has used aggressive language (otherwise we call S's language "friendly"). Using aggressive language either has a benefit (if b is larger than the exogenous threshold $\hat{b}$) or a cost (if $b < \hat{b}$). The threshold $\hat{b}$, which we assume to be common knowledge, describes that S would prefer to communicate in neutral tone

with people who are ideologically close to her, but prefers using aggressive language with those who have very different ideological views from her. g (for "gratification") measures the size of this benefit.

Note that evidence and aggressive language can be used regardless of θ (and their direct costs and benefits are independent of θ); they therefore do not convey any inherent information about θ. In equilibrium, as we will see, they can become informative as S may make their use conditional on θ.

We are interested in the most informative Perfect Bayesian Equilibrium ("equilibrium" in the following). As with any communication model, there usually exist equilibria where for at least some parameters, less information is transmitted, but focusing on the most informative equilibrium allows us to describe what *can* be achieved through communication. If several PBE transmit the same amount of information, we are interested in the sender-preferred PBE among the most informative equilibria, since it transmits the most information at the least cost.

This sender-preferred most-informative PBE (or *SPMI PBE*) will be our main solution concept.

1.2. What do these assumptions mean?

Given the applied nature of our paper, we briefly want to discuss why we think these assumptions are useful abstractions to think about the real-life interactions we are interested in. Readers who are familiar with the theoretical literature on communication or find the assumptions immediately plausible may prefer to skim through this subsection. We already discussed how the motivations of players in our model can be related to and motivated by empirical studies in the literature section of the introduction.

The basic communication problem R's payoff and the first part of S's payoff establish a conflict of interest: Both of them care about R's action, but they differ in their preferences about what the optimal action is. R will optimally choose $a = \mathbb{E}[\theta|m] + b$, while S would prefer R to take the action $a = \theta$. a could have some concrete interpretation (and be a vote, or some other real-life action), but more broadly we want to interpret it as R's "posterior opinion" about what is the right policy or the right political choice, after having seen S's message.

S cares about the posterior opinion – but S and R disagree on what "good" posterior opinions are. θ represents some aspect of the world that S knows more about than R, but even if θ was known, S and R would not perfectly agree (and b represents the size of this "absolute", ideological disagreement). S thus wants to represent θ in a way that moves a as close as possible to the action S considers desirable.

As an example, imagine that S and R are debating which amount of unemployment benefit optimally improves welfare of the overall population (and higher numbers roughly

mean "more unemployment benefit"). S and R have some ideological difference on how much support people should get who do no work, but S also has some knowledge on e.g. the effects of unemployment benefits on economic growth. We can note that (i) S only cares about R's final opinion, which is the sum of R's ideology and R's belief about the effect of unemployment on growth, and (ii) S knows that if θ were common knowledge, R would advocate a benefit level that S considers too high on purely ideological grounds. For these reasons, S might have an incentive to misrepresent θ if and only if $\theta = 1$, and pretend that it is 0 instead. However, we can see that S only wants to mislead R if b is large enough, since e.g. if $b = 0.01$ and $\theta = 1$, convincing R that $\theta = 0$ would mean that R supports an unemployment level that is too low by 0.99, whereas convincing him of the truth would mean that R supports a level that is too high by 0.01 from S's perspective. (This is precisely the content of our lemma 1 below.)

The use of evidence We think of "evidence" as e.g. the use of links to news stories or explanatory articles, finding statistics or graphs, or writing longer arguments to communicate about θ. That all of this is costly is relatively straightforward. The crucial assumption we are making, however, is that evidence does *not* convey any verifiable information about the state of the world θ.[10] Instead, the only thing that is firmly observable when S uses evidence is that S has put effort into finding and using evidence (given by cost parameter c). Nevertheless, we will see that this can convey information about θ since S may condition her use of evidence on θ in equilibrium.

In our example of S and R debating unemployment benefits, it is unlikely that S can provide definitive proof for the growth consequences of unemployment benefits within the confines of a brief social media interaction – especially given that R knows in which direction S wants to influence him. What S may be able to communicate by using evidence, however, is the strength of her feeling about R's opinion (i.e. her cost depending on how much R gets it "wrong" from her perspective), which in turn depends on θ, the knowledge that S has about the world.

The use of aggressive language "Aggressive language" is any kind of hostile or unfriendly language that e.g. insults, belittles or ridicules someone. We assume that S derives some direct, "expressive" payoff from using such language and that this payoff varies in b, the ideological distance between S and R. Using aggressive language against people that are ideologically close can be painful, as these are people that S perceives as friends or allies and therefore does not like to alienate. Using aggressive language against people with very different ideological leanings, however, can be satisfying – because it allows S to "let off steam", affirms her identify and makes her feel part of a group with shared values.

[10] This is why we use the term "evidence", in the sense of "there is evidence for and against", rather than "proof", which is definitive.

We should note that evidence is mixed whether it actually makes people happy or content to be unfriendly to someone else on social media. For our analysis and for the purposes of ascribing economic "preferences", it is enough if people behave *as if* it did give them joy, i.e. they seek and take opportunities to behave aggressively towards people whom they disagree with. This is a widely-documented phenomenon – whether this behavior actually generates well-being for the actor is another question and outside our analysis.

We assume that the direct payoff from aggressive language changes linearly in b which leads to a threshold $\hat{b}$ such that aggressive language is painful if b is below the threshold and satisfying if it is above it. This linearity, however, is not crucial – our main results only rely on the payoff from aggressive language increasing in ideological distance b, and being negative for very small b and positive for sufficiently large b.

S and R's shared knowledge We assume that b is common knowledge, i.e. S and R know their ideological difference exactly and only θ is unknown to R. This may not always be given in real life (where people may enter into arguments with people whose position they do not fully understand), but we argue that in the context of social media it is not a completely misleading assumption: Many of the interactions we observe are people who "follow" each other, and thus have some idea about where they stand. Even when interacting with someone whose ideology is not known ex-ante, this can often be inferred from either the user's tweet or their profile.[11] Since we are mainly concerned with S's messaging choice, it is also enough for us that S *believes* she knows b, not that she indeed knows it perfectly. Given the judgment mentality of social media, this is perhaps not an unreasonable assumption.

Our assumption that θ is 0 or 1 with equal probability is, of course, also simplifying our analysis – but richer assumptions would not qualitatively change our results. If e.g. there were more states which occurred with different probabilities, the result would still be that costless "cheap talk" messages can work for small but not for larger values of b.

1.3. Analysis

We will first consider each of the different tools with which S can communicate about θ on their own; section 1.3.4 then combines the insights from these sections and formulates hypotheses for our empirical analysis.

In the following analysis, we will use subscripts to note when S is using evidence and aggressive language, e.g. $m(0) = 0_e$ and $m(1) = 1_a$ describes the strategy "if the state is 0, send the message 0 in friendly language and attach evidence; if the state is 1 send the message 1 using aggressive language and no evidence." We will also call any messaging

[11] The fraction of social media users who directly state their political identity in their profile has increased markedly in recent years – cf. Rogers and Jones (2021)

strategy *truthful communication* if it is $m(0) = 0$ and $m(1) = 1$ with any combination of evidence or language used. A PBE in which communication is truthful is an *informative* PBE.

1.3.1. Pure cheap talk

The following lemma establishes that θ can be communicated truthfully without any need for evidence or aggressive language if b is small enough. We denote R's belief, i.e. the probability that he assigns to the state being 1, by μ.

Lemma 1. *For $b \leq 1/2$, truthful cheap talk communication, i.e. $m(0) = 0$, $m(1) = 1$, $\mu(0) = 0$, $\mu(1) = 1$ and $a(m) = \mu(m) + b$, is the SPMI PBE. (Proof on page 30.)*

If b is too large, S is tempted to send the cheap talk message 0 indicating the low state of the world regardless of the actual state of the world and therefore this message 0 is no longer credible.

1.3.2. The use of evidence

In this subsection, we consider the use of evidence in isolation, i.e. we assume for now that S can use evidence but cannot vary her language. Following lemma 1, we can concentrate on interactions in which $b > 1/2$, for which no informative equilibrium exists in which S uses a simple cheap-talk message to signal that the state is 0.

Message $m(0) = 0_e$, i.e. combining message 0 with costly evidence, can still be credible due to the fact that S is more eager to induce a low action if the state is actually low than when it is high (this is the Spence-Mirrlees condition of signaling models). The following proposition completely classifies for which set of parameters there exists a PBE in which S uses evidence to credibly communicate θ.

Proposition 1. *If S can only use evidence and $b > 1/2$, the SPMI PBE is as follows:*

1. *For $c \in [2b - 1, 2b + 1]$, S credibly signals θ with $m(0) = 0_e$, $m(1) = 1$, and R's beliefs are $\mu(0_e) = 0$, $\mu(1) = \mu(0) = \mu(1_e) = 1$ and $a(m) = \mu(m) + b$.*

2. *For $c \notin [2b-1, 2b+1]$, no meaningful communication is possible in equilibrium, i.e. there is no equilibrium in which information is transmitted from S to R.*

(Proof on page 30.)

Figure 1 illustrates the lemma. For $b > 1/2$, there exists a band of parameter values (b and c) such that there exists an equilibrium in which θ is communicated truthfully by the use of evidence. If, for a given b, c is too low, either type would find it optimal to expend c if $\mu(0_e) = 0$, and hence there cannot be an equilibrium in which $\mu(0_e) = 0$. If c is too large, then no type finds it optimal to expend c even if $\mu(0_e) = 0$, and hence evidence is never used.

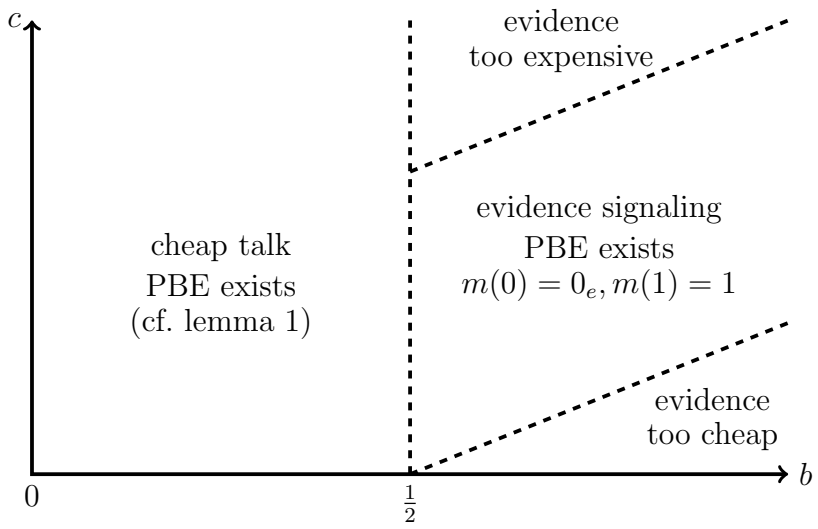

Figure 1: Most informative equilibria if S can only use evidence

1.3.3. The use of aggressive language

This section considers the use of aggressive language in isolation, i.e. we assume for now that S can only vary her language to be either friendly or aggressive (and cannot use evidence). Since b and $\hat{b}$ are common knowledge, R knows $b - \hat{b}$ and hence knows whether S would derive a direct benefit from using aggressive language or not. S can thus use aggressive language (or its absence) as a costly signaling tool (similar to the use of evidence in the previous section).

We can distinguish two ways in which S can do so:

- If $b < \hat{b}$, aggressive language is costly for S and could therefore be used to directly support the message 0 (similarly to how costly evidence was used in section 1.3.2). We call this effect "tough talk among friends".

- If $b > \hat{b}$, S enjoys using aggressive language, and its costly *absence* can hence be used to support the message 0. We call this "biting your tongue".

The following proposition establishes when each of the two strategic uses of aggressive language is possible. (We assume that $\hat{b} > 1/2$; if that is not the case, S will never engage in "tough talk among friends" as can also be seen intuitively from figure 2.)

Proposition 2. *If S can only vary her language and $b > 1/2$, the SPMI PBE is as follows:*

- *If $b < \hat{b}$, S credibly signals θ with $m(0) = 0_a$ and $m(1) = 1$ ("tough talk among friends") if $\frac{2b-1}{\hat{b}-b} \leq g \leq \frac{2b+1}{\hat{b}-b}$.*

- *If $b > \hat{b}$, S credibly signals θ with $m(0) = 0$ and $m(1) = 1_a$ ("biting your tongue") if $\frac{2b-1}{b-\hat{b}} \leq g \leq \frac{2b+1}{b-\hat{b}}$.*

For all other values of g, there is no truthful communication as long as $b > 1/2$. (Proof on page 30.)

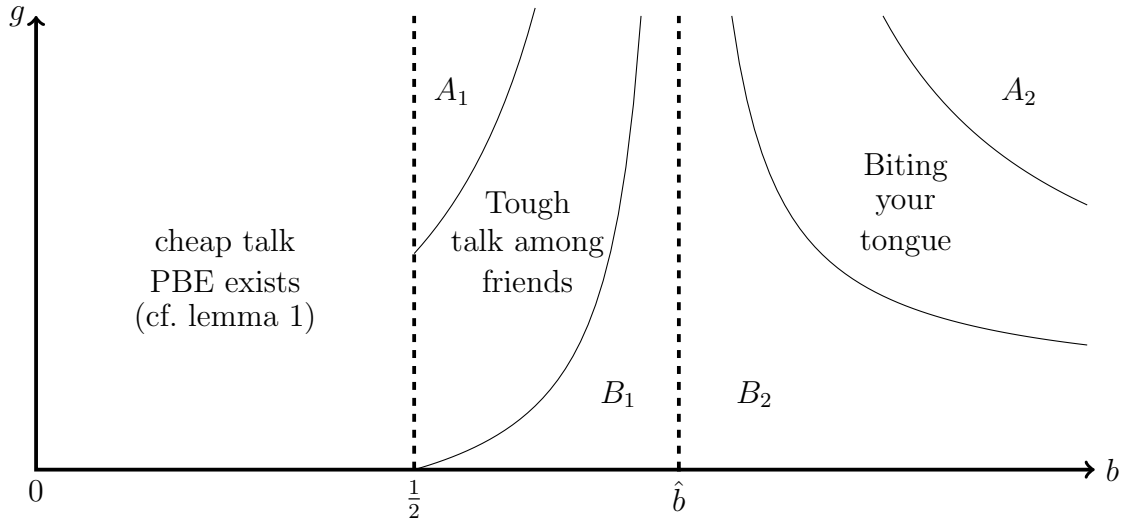

Figure 2: Most informative equilibria if S can only vary language.

Figure 2 illustrates the proposition: In areas labeled "A", the costs and benefits of using aggressive language are too high to lead to informative separation based on types. In areas labeled "B", the costs and benefits are too low, so that as soon as the presence or absence of aggressive language becomes informative, either type would find it optimal to always send message 0 by using (or not using) aggressive language. In areas labeled with subscript 1, no type of S uses aggressive language; in areas labeled with subscript 2, every type of S uses aggressive language. Only in the intermediate areas does there exist a PBE in which θ is credibly communicated through the use of aggressive language by exactly one type.

1.3.4. How does communication evolve as b increases?

In our main model, we assume that S can simultaneously choose (i) Message m, (ii) whether to use evidence and (iii) whether to use aggressive language. The previous sections have outlined how each of these choices can transmit information and contribute to the existence of informative PBE. In the general model, there can be PBE in which S uses several of these methods at the same time. For example, for $b < \hat{b}$, S could be able to credibly signal that $\theta = 0$ by both using evidence and using (costly) aggressive language (i.e. using strategy $m(0) = 0_{ae}$ and $m(1) = 1$).

In this section, we will derive some general results about what the SPMI PBE can look like for different magnitudes of b. Which PBE exists (and is SPMI) for any given b depends on c and g, and giving a general classification would therefore involve a multitude of case distinctions.

Instead, we will focus on broad insights about how S uses evidence and aggressive language in the SPMI PBE for different intervals of b – and in particular, how often (and in which combination) S uses these tools. We will do so by establishing a series of lemmas.

All these lemmas assume that $\hat{b} > 1/2$ – if that is not the case, lemma 2 does not apply anywhere, and lemma 3 applies for $b > 1/2$ analogously (while lemma 1 applies unchanged.)

Lemma 2. *If $1/2 < b \leq \hat{b}$, S uses one of the following messaging strategies in any SPMI PBE:*

1. $m(0) = 0_e$ *and* $m(1) = 1$
2. $m(0) = 0_a$ *and* $m(1) = 1$
3. $m(0) = 0_{ae}$ *and* $m(1) = 1$
4. $m(0) = 0$ *and* $m(1) = 0$ *(or any other uninformative strategy of S that does not involve evidence or aggressive language))*

(Proof on page 31.)

Lemma 3. *If $b > \hat{b}$, S uses one of the following messaging strategies in any SPMI PBE:*

1. $m(0) = 0_e$ *and* $m(1) = 1$
2. $m(0) = 0$ *and* $m(1) = 1_a$
3. $m(0) = 0_e$ *and* $m(1) = 1_a$
4. $m(0) = 0_{ea}$ *and* $m(1) = 1_a$
5. $m(0) = 0_a$ *and* $m(1) = 0_a$ *(or any other uninformative strategy of S that does not involve evidence and where both types use aggressive language)*

(Proof on page 31.)

For very large b, we can narrow down the set of SPMI PBE even further for generic values of c and g:

Lemma 4. *Assume $g \neq 2$. For every combination of c and g, there exists a $\bar{b}$ such that for any $b > \bar{b}$, the only PBE is that both types use aggressive language and no information is transmitted. (Proof on page 31.)*

1.4. Turning theoretical results into empirical predictions

Which PBE will exist (and be SPMI) at which point in these intervals depends on the exact values of c and g. These values will of course differ from person to person, and in any case we think of this model as being at a relatively high level of abstraction, so that c and g are not parameters that have a direct equivalent in reality (or could usefully be estimated from actual data).

What we think our model *does* allow us, however, is to derive broader statements about how the behavior of S depends on b.[12] We will make three observations about equilibria that follow directly from the results we have derived; we will then turn these into hypotheses for our empirical work.

Observation 1: For small b, evidence is never used. For intermediate b, evidence is sometimes used. For very large b, evidence is never used. This leads us to:

Hypothesis 1. *The use of evidence first increases, then decreases in ideological distance.*

Observation 2: For small b, aggressive language is never used. For intermediate b, at most half of the senders use aggressive language. For sufficiently large b, all senders use aggressive language. This leads us to:

Hypothesis 2. *The use of aggressive language is increasing in ideological distance.*

Observation 3: In most cases in which evidence is used, it is not used by the same person who uses aggressive language.

Hypothesis 3. *Aggressive language is used less often together with evidence than without it, and the use of aggressive language increases less in ideological distance for replies that use evidence compared to the replies that do not use evidence.*

2. Empirical Evidence from Twitter

2.1. What are the empirical equivalents of our model?

Starting from our theoretical analysis and the hypotheses that we have formulated, we will now consider a dataset of sender-receiver communication that we have collected on the social networking site Twitter in early 2021.[13]

At the time of our data collection, Twitter allowed its users to send short messages of 280 characters to people who have followed them ("tweets"). Furthermore, users could respond to other users' tweets with similar short messages ("replies").

We think of the latter interaction as being a close real-life equivalent to what we have modeled in section 1: A user has mentioned some topic or voiced some opinion; now

[12] Of course, R's beliefs and behavior also depend on b, but we cannot usefully observe either of them from the data we have, while we can observe S's behavior.

[13] Twitter has since been rebranded as X and has seen changes in its functionality and user base; our descriptions all apply to the time at which we collected our data.

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

Democrats	Republicans
endgunviol	kssen
trumpshutdown	arkansan
actonclim	countymeet
protectourcar	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, i.e. have been reduced to their grammatical stems.)

another user can choose whether and how to respond. The person replying to a tweet is hence the sender, S, of our model; the person who wrote the original tweet is the receiver, R. Learning about the ideological distance between S and R will then allow us to examine how the message from S to R depends on that ideological distance.

The following paragraphs describe how we collected our data, how we measure the ideological distance between S and R, and how we conduct our actual analysis.[14]

2.2. Preliminary Steps

First step: Building a dictionary of partisan words We analyzed the tweets of all members of the 116th and 117th U.S. Congress (by early 2021) 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 in tweets by Democratic and Republican members of Congress, and isolated the words whose usage was (i) high enough and (ii) different enough between parties.[15]

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.

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

[15]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 random 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 their 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.[16] 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.[17] 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 8 on page 34 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 (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 will receive score 0, while a user who only uses words that are only used by Republicans will 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.[18]

Fourth step: Collecting data about interactions We then collected data about all replies written by these random Twitter users whom we had scored in the previous step. For each reply by a user in our sample, we can observe the identity of the user they have replied to, and we can then determine the ideology score of that user. (Again we are excluding users for whom there is too little data to assign an ideology score.)

This leaves us with the basic unit of our subsequent empirical analysis: A dataset of reply tweets, where for each interaction we know: (i) the ideological score of the sender, (ii) the ideological score of the receiver (i.e. the author of the tweet that is replied to) and

[16] We call a tweet *original* if it is not a reply or a retweet, i.e. we score users based on tweets that are "unprovoked".

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

[18] More information and an exact formula can be found in the supplementary material.

hence also the ideological distance between sender and receiver, (iii) the content of the reply tweet, (iv) further data about whether a link or a hashtag was used.

2.3. Descriptive statistics

Before turning to the actual analysis we first present some descriptive statistics in table 2. Here, *absolute score difference* is the absolute score difference between sender and receiver, *number links* gives the number of links in a reply (and *link dummy* whether there is a link), *media* describes whether media (pictures or video) are used, *profanity* gives the frequency of profanity and *sentiment* the sentiment score. *tweet length* and *word length* are given in characters.

	Obs	Mean	Std. Dev.	Min	Max
absolute score difference	139075	0.101	0.085	1.885e-06	0.465
receiver score	139075	0.473	0.116	0.099	0.829
sender score	139075	0.453	0.072	0.240	0.739
number links	139075	0.034	0.202	0	6
link dummy	139075	0.031	0.172	false	true
media	139075	0.071	0.257	false	true
profanity	139075	0.151	0.228	5.145e-04	1.000
sentiment	139075	-0.031	0.570	-0.998	1.000
hashtags	139075	0.259	0.918	0	29
tweet length	139075	153.293	78.027	0	293
word length	138878	5.477	0.638	1.000	34.000

Table 2: Descriptive statistics

There are no important differences between senders of different ideologies in how much they reply or in how their replies vary in the variables that interest us. Such differences only emerge when we consider interactions, i.e. who replies to whom and how large is the ideological distance between them – this is the main focus of our analysis below. Descriptive graphs in the appendix (figure 6 on page 40) show the main variables of our analysis for senders of different ideologies. To control for differences between senders and the effects of certain times or events, our main analysis will use sender and day fixed effects. Figure 3 shows the distribution of the score difference between sender and receiver.

2.4. Difference-in-differences analysis

Using our work from the previous steps, we generated a data set containing 139,075 reply tweets sent from 2,401 senders to 28,796 receivers.[19] 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

[19]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. Our results remain qualitatively unchanged when doing so.

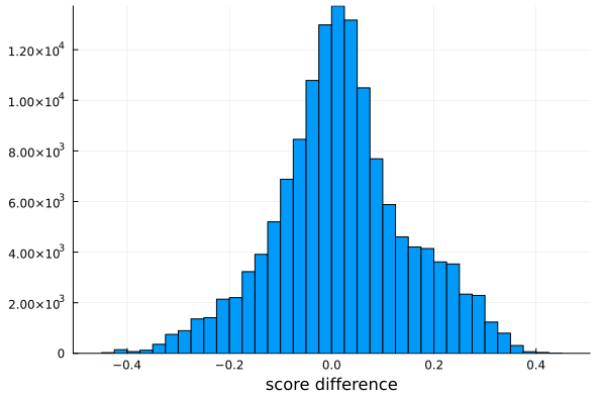

Figure 3: Distribution of the difference between receiver score and sender score.

in the ideological distance between sender and receiver. Formally, we will use OLS to estimate equations of the following form:

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

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 sender $S(i)$, $+\text{FE}_{day}$ is a fixed effect for the day of the reply, and ε_i is an error term (we cluster error terms at the sender level). Due to the sender fixed effect, we effectively use variation in the score of the receivers to estimate the parameter of interest β. For each regression that is in the main text of the paper, tables in the appendix (starting on page 34) explore how the coefficients change under different combinations of fixed effects and standard errors.

Signaling with evidence Our model predicts that with a larger ideological distance, the communication of actual information between S and R requires that S uses signaling tools such as evidence. Hypothesis 1 in section 1.4 summarizes our results, which is that the use of evidence should increase in ideological distance, and eventually decrease again as informative communication becomes impossible.

The evidence of our model can take many different forms in this real-life setting, but they all must have in common that their usage is costly for S – usually, this cost will be the time that S uses to find and deploy the evidence. On Twitter, there are three main ways to use "observable" evidence:

1. Through the use of hyperlinks (to news articles, statistics, fact checks, research etc). We can directly observe in our dataset whether a reply includes a hyperlink.

2. Through making complex and sophisticated arguments (which, by the way, is also what we are trying to do in this paper). While we cannot directly observe or quantify how complex and sophisticated an argument is, we can measure two dimensions that

must at least be positively correlated with our variable of interest: the length of words, and the length of tweets. The former is an integral part of many widely-used readability scores, such as the Automated Readability Index or the Coleman-Liau-Index (where longer words indicate more complex and sophisticated language). The latter follows from the simple consideration that arguments take space.[20]

3. Through the use of media (mostly pictures), which can be a costly signaling tool for two reasons. First, it is time consuming to search for a fitting picture (such as a statistic, graph, meme or screenshot) that supports one's argument. Second, twitter users often share longer texts as photos instead of typing (and hitting the 280 character limit).

Table 3 shows that all signaling tools are more likely to be used if the ideological distance between sender and receiver is large. This is consistent with the first part of hypothesis 1. The predicted inverse U-shape exists for some but not all variables, see table 4. We discuss possible explanations in section 3.2.

	number links	link dummy	tweet length	word length	media
	(1)	(2)	(3)	(4)	(5)
absolute score difference	0.015*	0.016*	23.234***	0.215***	0.072***
	(0.007)	(0.007)	(4.313)	(0.030)	(0.015)
sender fixed effects	Yes	Yes	Yes	Yes	Yes
day fixed effects	Yes	Yes	Yes	Yes	Yes
N	139,035	139,035	139,035	138,838	139,035
R^2	0.451	0.333	0.275	0.154	0.374

Table 3: 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%)

Aggressive language In our theoretical model we saw that aggressive language can help with credible communication, and hypothesis 2 summarizes the main predictions. Our data allows us to observe whether larger ideological distance leads to more aggressive language.

We measure aggressive language in three ways: first, we check whether a tweet contains profanity.[21] Second, we measure the sentiment of a reply using the sentiment dictionary

[20] Of course, we are not assuming that complex words or longer tweets are *sufficient* to make a more complex or costly argument, but for our purposes it is enough if they are *necessary* at least some of the time, so that there is a correlation between word and tweet length and the complexity of one's arguments.

[21] 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).*

	number links	link dummy	tweet length	word length	media
	(1)	(2)	(3)	(4)	(5)
absolute score difference	0.056**	0.055**	10.552	0.380***	0.015
	(0.019)	(0.018)	(10.084)	(0.088)	(0.032)
absolute score difference^2	-0.146**	-0.138*	44.783	-0.584	0.202
	(0.056)	(0.054)	(35.662)	(0.321)	(0.113)
sender fixed effects	Yes	Yes	Yes	Yes	Yes
day fixed effects	Yes	Yes	Yes	Yes	Yes
N	139,035	139,035	139,035	138,838	139,035
R^2	0.451	0.333	0.275	0.154	0.374

Table 4: An inverse U-shape relation between ideological distance and evidence exists for some but not all variables interpreted as evidence. (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

by Hu and Liu (2004), which gives scores to certain words and phrases that mark positive or negative sentiments.[22] Third, we check whether hashtags are present in a reply as these are often used in a declarative, emotional fashion to "make a point". (For example, our dataset contains many examples of accounts simply replying "#fakenews" to accounts they – presumably – disagree with.)

The left panel in Figure 4 illustrates a strongly positive relationship between ideological distance and the use of profanity in interactions. The central panel of the same graph shows that a larger ideological distance between sender and receiver is associated with a more negative sentiment. The right-hand panel shows that there is also a positive relationship between ideological distance and hashtag frequency in replies. Table 5 gives the corresponding regression results. All these results are in line with hypothesis 2.

	profanity	sentiment	hashtags
	(1)	(2)	(3)
absolute score difference	0.164***	-0.308***	0.280***
	(0.013)	(0.028)	(0.045)
sender fixed effects	Yes	Yes	Yes
day fixed effects	Yes	Yes	Yes
N	139,035	139,035	139,035
R^2	0.149	0.101	0.422

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

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

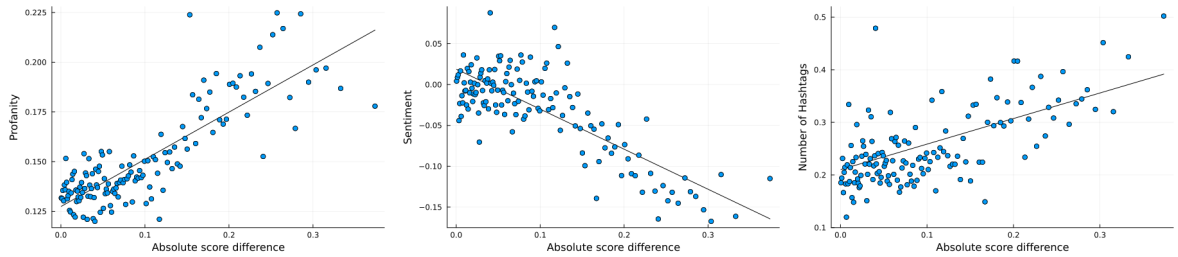

Figure 4: A larger (absolute) difference between sender and receiver score is associated with a lower sentiment score and higher usage of profanity and hashtags. (Binned scatterplot, one dot corresponds to roughly 1000 observations which are grouped according to absolute score difference. One outlier removed in the third plot.)

Interaction effects between evidence and aggressive language Hypothesis 3 predicts a particular relationship between evidence and aggressive language: Namely, that evidence is less likely to occur with strongly aggressive language than without, and that the "growth" in aggressive language as ideological distance increases is smaller in interactions that use evidence. Table 6 shows that indeed the effect of ideological distance on aggressive language is weaker for the variables *profanity* and *hashtags* if we restrict the sample to those replies containing a link. The same cannot be said of *sentiment*. To explain this latter finding, it is important to keep in mind that even in cases in which the sender uses evidence she does so in order to change the receiver's mind. That is, there is still a disagreement at the heart of the conversations and presumably the disagreements are more likely and more pronounced the larger the difference in ideology. Notions of disagreement, however, are typically associated with a negative sentiment. In other words, trying to convince with evidence in a respectful manner – and without using aggressive language – might still be associated with a negative sentiment due to the underlying disagreement.[23] Profanity, on the other hand, is not compatible with respectful, non-aggressive language.

Homophily We briefly want to document another effect in our data: A tendency towards homophily in who interacts with whom. Such an effect has already been demonstrated – albeit with different methods than ours – by other studies, see for example Barberá et al. (2015) or Krasodomski-Jones (2017).

We show that in our data (and with our scoring) the more right-wing Twitter users are, the more likely they are to interact with other right-wing Twitter users; see the regression results in table 7.[24] Figure 5 shows the distribution of absolute score differences in our data which also show that most communication takes place between users of similar ideological

[23]For example, sentences like "While I get your idea, I think you may have missed the increase in last month's unemployment rate." are not exactly aggressive but get a negative sentiment score (-0.4215 in the example sentence).

[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.

	profanity	sentiment	hashtags
	(1)	(2)	(3)
absolute score difference	0.098	-0.413**	-0.468
	(0.063)	(0.158)	(0.394)
sender fixed effects	Yes	Yes	Yes
day fixed effects	Yes	Yes	Yes
N	3,822	3,822	3,822
R^2	0.296	0.395	0.760

Table 6: Effects of ideological distance on aggressive/emotional language in the subsample of replies with links (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

leaning.

This finding is consistent with the existence of so called "echo chambers" 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)	0.408***
	(0.018)
sender score	0.144***
	(0.038)
N	139,075
R^2	0.008

Table 7: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

3. Discussion

In this section, we will discuss some of our assumptions and results in more detail.

3.1. Costs and benefits that are missing from our model

A cost of being treated aggressively (and other costs or benefits of the receiver). We think it is highly plausible that such a cost exist (few people enjoy being shouted at or insulted), but it would not change R's behavior in our model, which focuses on the

[25]For a discussion of the potential benefits of echo chambers for debate, see Jann and Schottmüller (2023).

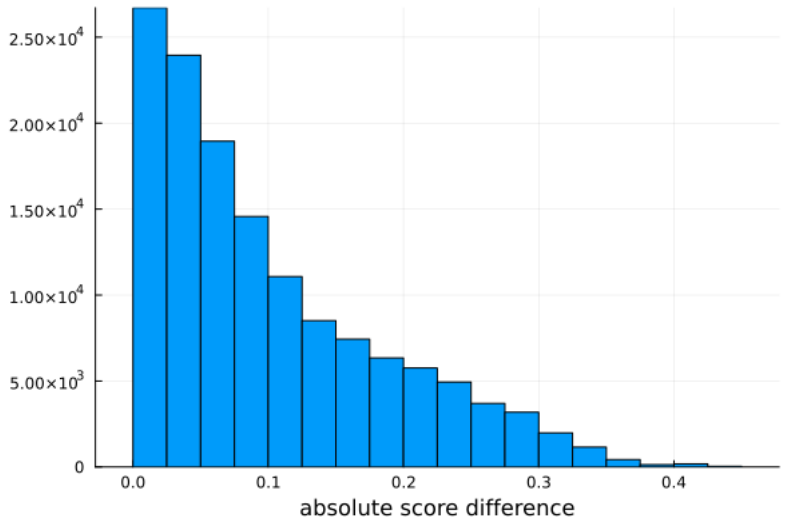

Figure 5: Histogram of the absolute score difference.

informational inferences that R draws. (This does mean that R is often less likely to change his mind after hearing aggressive language, but this is an endogenous effect of when S chooses to use aggressive language, not an instinctive reaction to aggressive language.)

The same would be true for other direct costs and benefits of R: As long as they do not influence R's inference problem, they do not change S's behavior, which is what we are mainly interested in (and are able to measure).

Trolls and bots. All of the senders of our model are interested in transmitting real information to R, at least in principle. For the case of online trolls and disinformation bots, this is clearly not the case – these are players with a completely different utility function that do not exist in our model.

One way to think about such players would be that they "dilute" information transmission: Imagine that S is a "real" person (like in our model) with some probability $1-\lambda$, and a disinformation bot (which sends random messages, or always the same message that does not depend on θ), with probability λ. R would then have to consider that some (or all) types of messages are with some probability $p(\lambda)$ sent by a bot, which would lower the impact of messages on R's belief. This, in turn, changes S's incentives to use evidence or aggressive language and could, for example, make it unprofitable to use evidence because the expected payoff (which results from a change in R's posterior belief) shrinks, while the cost stays the same. Conversely, if e.g. c is so small that without the presence of bots, credible signaling with evidence is not possible, the presence of some bots could make it possible. The dilution of information has the same effect as an increase in the costs of the signal which can increase informativeness if the initial costs are low but also decrease informativeness if the initial costs are high.

3.2. What happens at a large ideological distance?

Our model makes two predictions about what happens at large ideological distances that differ to some extent from what the data shows. We will briefly discuss these differences in this section, and why we think they represent effects that are outside (but not contradictory to) our model.

First, our model suggests that as b gets very large, sending uninformative messages with aggressive language becomes very attractive to S. S should then be very interested in sending aggressive messages to all kinds of receivers that are ideologically distant as those lead to a high emotional benefit. Empirically, however, we see that there are fewer interactions between ideologically more distant users than between users that are ideologically closer.

One element that is missing from our theoretical analysis is the question of who *gets* to reply to whom. In our model, we consider a simple pair of S and R who are connected through no action of their own. In reality, users of Twitter (just like any other social media) can only reply to what they see, and what they see is determined to whom they follow. But "who follows whom" shows a huge tendency of homophily, as e.g. also studies by Barberá et al. (2015) or Krasodomski-Jones (2017) have shown. Such homophily can e.g. occur endogenously if there is an additional cost to being exposed to people one considers obnoxious, or of being subjected to aggressive language (cf. section 3.1).

Other studies (cf Goyanes et al., 2021) suggest that aggressive language itself can lead to "unfriending" or "unfollowing" (and also "blocking") between social media users, which would mean that connections that persistently lead to aggressive language are less likely to exist than ones in which aggressive language is used less often.

We therefore think it plausible that such homophily strongly limits the set of potential interactions with ideologically distant people that any one user can have. Even though users do often enjoy using aggressive language and uninformative messages in such interactions, they are not as plenty as they would be if all interactions were equally likely to be possible.

Second, our model (in hypothesis 1) predicts that the use of evidence is inversely U-shaped in ideological distance. In our data, we see this for some variables but for others we only see a monotonic increase. We think this is related to the previous point: The second part of the inverse U-shape would be driven by interactions with a large ideological distance, in which no information is transmitted and no evidence used. But since a large number of interactions with large ideological distance never occur due to the homophily mentioned above, we do not observe these messages.

We also need to keep in mind that c and g, the costs of evidence and cost/benefits of aggressive language, are likely to be at least somewhat person- and context-specific: It may be easier to find evidence or abstain from aggressive language on a well-documented topic that does not have strong emotional connections. In some cases, however, c could

be large, and we would then see some interactions with evidence even for relatively large b.[26]

The inverse U-shape of the model is also driven by our assumption that c, the cost of using evidence, is fixed – since this is what makes it impossible to credibly use evidence at sufficiently large b (cf. figure 1 on page 13). It might be plausible that S can choose c in a verifiable way, i.e. by using several links, writing an even longer argument that considers more cases etc. Credible communication could then be possible even for quite large b.

3.3. Are arguments and links costly signals or "verifiable information"?

We have made the assumption that when people use evidence (either through verbal arguments or references to sources), this results in *costly signals*. These carry information because only some people – those with specific knowledge or convictions – would have invested the time in finding and deploying this evidence. But one may wonder why we do not treat arguments or hyperlinks as "verifiable information" (in the sense of Milgrom and Roberts (1986)), which conclusively proves a point as it can only be sent if true.

We believe that such verifiable information exists (and, in fact, hope that readers will see the proofs in our appendix as such). But it does not strike us as a plausible assumption for what happens in most online political debate. First, such discussions often take place under enormous constraints of time and attention. People may write out a brief explanation of their thinking or provide one hyperlink; they will not provide detailed point-by-point arguments or refutations with reference lists that would rise to the standard of verifiable information.

Second, and perhaps more importantly, political debates rarely center on simple and easily established facts about which it would be possible to present incontrovertible evidence. Major political questions, such as about the effects of immigration or the state of a country's democracy, are amenable to evidence in the form of arguments or stories that support one's views – but these rarely constitute unambiguous and unmistakable proof of a larger point, in a form that someone who disagrees would be willing to accept.

If we did assume that users could deploy verifiable evidence at some cost, the patterns predicted by our model would change. For large bias differences, people would always send verifiable messages and we would see a lot of high-quality information over large distances. This is not what our data suggests. But it would also raise the practical problem of more sophisticated forgeries: The larger the bias difference, the larger the benefit of being believed – and hence the larger the incentive to invest a lot of time into fabricating something that looks like verifiable information. To overturn our results, verifiable information would have to be available, affordable and virtually tamper-proof. If it is not, it makes more sense to think of it as a signaling device, as we do in our model.

[26] It seems not unreasonable that people are more likely to overcome their tendency for homophily in those cases where meaningful communication is still possible which would, for example, be the case for interactions in which c is relatively high.

Arguments and references could also be persuasive for completely different reasons – for example, if there was a mental or (potential) reputational cost to using them if one knew them to be false. Listeners would then reason: "Well, if she makes the argument, there must be some truth to it." Such measures might be unnecessary among those who feel they mostly agree, and would fail among those who feel they have nothing in common – and so the practical implications of such an assumption would be very similar to what our model derives.

3.4. Who is the audience of a reply tweet?

Given that all interactions on Twitter were public during the time period that our data covers, one might wonder whether it is correct to consider the author of the original tweet as the reply's "audience" – as is the basis for our analysis in section 2. While this is of course a simplification that will not always apply, we have two main justifications. First, since our main dataset covers the interactions of a random sample of Twitter users (and not celebrities or similar), a lot of interactions do not play out in front of a large audience, and many replies are by far most likely to be seen by the person that is being replied to.

Second, even if the reply is seen by more people, these people are likely to be followers of R, given that these are the people most likely to be looking at R's original tweet. We could therefore think of S as not just addressing R, but addressing the entire group of people that follow R. Due to the homophily of Twitter followership that various studies have documented, we would expect that group's ideology to be similar to R's ideology. But then there would not be too much of a difference between S messaging R and caring about how the message is received and understood, or S messaging R's entire group of followers and caring about how the message is received and understood.

4. Conclusion

We have combined theoretical and empirical methods to understand how political debate works on social media: How different motivations combine or contradict each other, how people use different tools at their disposal to achieve the different goals they are interested in. While we do not claim to causally identify any of these theoretical mechanisms in our dataset, we do find effects, connections and tendencies that are compatible with what our theoretical model describes.

We started our introduction by pointing to widespread unhappiness with the state of online political debate. It may therefore be a natural question to ask whether our results can point out any ways to improve debate – either such that more information is exchanged, or participants are happier, or both.

Our analysis suggests that mistrust between people with different views, and the fear that the other may be discussing "in bad faith", makes communication harder. Tools of strategic persuasion such as making detailed arguments, referring to sources or making

an effort to stay polite can only partially compensate for this difficulty. It is likely that the problem of mistrust is especially large in online contexts, where the basic ideological "bias" of another user may easily be inferred from their statements or user profile, but it is much harder to build a reputation for good-faith communication. A general increase of partisan mistrust does not help – such as in the United States, where party supporters have become more likely to describe members of the other party as "close-minded" and "immoral" in recent years.[27]

Our analysis suggests that the ability to use aggressive and hurtful language can have an important strategic benefit. But this benefit really only exists if debaters have every freedom to be nasty to each other: If strong content moderation, for example, would punish – or technically prevent – the use of aggressive language, then the absence of aggressive language bears no cost, and hence "biting your tongue" contributes less (or nothing) to your persuasiveness. Even if aggressive messages bear no information (and come with a welfare cost to those who read them), they can have the benefit of making non-aggressive messages more informative.

Other steps, such as making sources and evidence more easily available online, may also have ambiguous effects. While it might allow for more informed debate, it could also lower the cost of introducing all kinds of sources and thus also reduce their argumentative weight and persuasiveness. At the time of writing this paper, the fact that everyone with a phone has most of the world's knowledge at their fingertips has not yet quite created an enlightened public sphere in which well-informed citizens debate in good faith. That would require more than technology.

[27] https://www.pewresearch.org/politics/2019/10/10/how-partisans-view-each-other/

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:

The only relevant deviation is the 1-type pretending that $\theta = 0$ by sending $m(1) = 0$. This gives payoff $-(b-1)^2$, whereas the equilibrium strategy gives payoff $-b^2$. The 1-type hence optimally sticks to the equilibrium strategy if $-b^2 \geq -(b-1)^2$ or $b \leq 1/2$. □

Proof of proposition 1:

For the 0-type, following the equilibrium strategy gives payoff $-b^2-c$, while any deviation gives payoff $-(1+b)^2$; this gives an IC constraint of $c \leq 2b+1$. For the 1-type, the equilibrium strategy gives payoff $-b^2$, while the only potentially profitable deviation is $m(1) = 0_e$ (as all other deviations do not change R's beliefs) which gives payoff $-(b-1)^2-c$. This gives the second IC constraint $c \geq 2b-1$; together the two constraints prove the proposition. □

Proof of proposition 2:

First, consider $b < \hat{b}$ and combine the messaging strategy $m(0) = 0_a$ and $m(1) = 1$ with the beliefs $\mu(0_a) = 0$ and $\mu(0) = \mu(1) = \mu(1_a) = 1$. The 0-type's payoff from the equilibrium strategy is $-b^2 + g(b-\hat{b})$, while the payoff of any deviation is at most $-(1+b)^2$. This gives the first IC constraint: $g \leq \frac{2b+1}{\hat{b}-b}$. Similarly, the 1-type's payoff from the equilibrium strategy is $-b^2$ while the best possible deviation payoff is $-(b-1)^2 + g(b-\hat{b})$. This gives the second IC constraint: $g \geq \frac{2b-1}{\hat{b}-b}$. Together, these two constraints imply the first part of the proposition.

Now consider $b > \hat{b}$ and combine the messaging strategy $m(0) = 0$ and $m(1) = 1_a$ with the beliefs $\mu(0) = \mu(1) = 0$ and $\mu(1_a) = \mu(0_a) = 1$. The 0-type's payoff from the equilibrium strategy is $-b^2$, while the payoff from the best deviation is $-(1+b)^2+g(b-\hat{b})$. This gives the first IC constraint: $g \leq \frac{2b+1}{b-\hat{b}}$. The 1-type's payoff from the equilibrium strategy is $-b^2 + g(b-\hat{b})$, while the payoff from the best deviation is $-(b-1)^2$, which gives the second IC constraint $g \geq \frac{2b-1}{b-\hat{b}}$. Together, these two constraints imply the second part of the proposition.

If these constraints are not fulfilled, there exists no PBE in which the 1-type finds it too costly to imitate the 0-type; this implies the third part of the proposition. □

Proof of lemma 2:

A PBE in which S uses evidence if $\theta = 1$ cannot be SPMI: If it is a PBE and S's message is uninformative, then there is also an uninformative equilibrium in which no evidence is used and which is therefore sender preferred. Hence, consider the case that it is a PBE and S's equilibrium message 1_e is informative. But if S can transmit information by acquiring evidence if $\theta = 1$, there also exists a PBE in which S transmits the same information by not acquiring evidence (which relaxes the IC-constraint), and which is sender-preferred since S does not incur the cost c if $\theta = 1$.

Analogously, a PBE in which S uses aggressive language if $\theta = 1$ cannot be SPMI. Since $b > 1/2$, there is no informative PBE in which $m(0) = 0$. This leaves only the set of PBEs given in the lemma. □

Proof of lemma 3:

Analogously to the proof of lemma 2, a PBE in which S uses evidence if $\theta = 1$ cannot be SPMI. A PBE in which S uses aggressive language *only* if $\theta = 0$ cannot be SPMI, since it is either uninformative (in which case S has a profitable deviation to also using aggressive language if $\theta = 1$), or it is informative in which case there must also exist a sender-preferred PBE in which S uses aggressive language in both cases (as this relaxes the 1-type's IC-constraint). The remaing PBE are the ones listed in the lemma. □

Proof of lemma 4:

Wlog we assume $\bar{b} > \hat{b}$. Consider the PBE listed in lemma 3. It follows from propositions 1 and 2 that the first and second of these PBE do not exist if b is large enough. An analogous argument applies to the fourth PBE. For the third PBE, the IC constraints are given by:

$$(g-2)b < g\hat{b} + 1 - c$$

(for the 0-type) and

$$(g-2)b > g\hat{b} - 1 - c$$

(for the 1-type). This means that for $g < 2$, the PBE in which the 0-type uses evidence and the 1-type uses aggressive language does not exist for

$$b > \frac{1 + c - g\hat{b}}{2 - g},$$

and for $g > 2$ it does not exist for

$$b > \frac{g\hat{b} + 1 - c}{g - 2}.$$

Only the PBE in which no information is transmitted and both types use aggressive language remains. □

B. Additional Tables and Figures

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

	number links				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.039***	0.039***			
	(0.001)	(0.009)			
absolute score difference	-0.054***	-0.054	-0.046	0.015*	0.015*
	(0.006)	(0.038)	(0.035)	(0.007)	(0.007)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.001	0.001	0.010	0.446	0.451

Table 9: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	number links				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.040***	0.040***			
	(0.001)	(0.011)			
absolute score difference	-0.075***	-0.075	-0.067	0.057**	0.056**
	(0.020)	(0.091)	(0.085)	(0.019)	(0.019)
absolute score difference^2	0.075	0.075	0.073	-0.148**	-0.146**
	(0.067)	(0.201)	(0.193)	(0.057)	(0.056)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.001	0.001	0.010	0.446	0.451

Table 10: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	link dummy				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.034***	0.034***			
	(0.001)	(0.006)			
absolute score difference	-0.036***	-0.036	-0.031	0.016*	0.016*
	(0.005)	(0.026)	(0.025)	(0.007)	(0.007)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.000	0.000	0.010	0.327	0.333

Table 11: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	media				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.050***	0.050***			
	(0.001)	(0.005)			
absolute score difference	0.204***	0.204**	0.197**	0.072***	0.072***
	(0.008)	(0.066)	(0.063)	(0.015)	(0.015)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.005	0.005	0.040	0.369	0.374

Table 12: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	tweet length				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	151.368***	151.368***			
	(0.324)	(1.471)			
absolute score difference	19.062***	19.062	16.935	23.785***	23.234***
	(2.448)	(11.786)	(11.001)	(4.381)	(4.313)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.000	0.000	0.015	0.267	0.275

Table 13: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	tweet length				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	151.042***	151.042***			
	(0.439)	(1.668)			
absolute score difference	27.199***	27.199	26.497	9.202	10.552
	(7.796)	(32.070)	(30.290)	(10.248)	(10.084)
absolute score difference^2	-28.298	-28.298	-33.272	51.450	44.783
	(25.743)	(118.178)	(112.189)	(36.280)	(35.662)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.000	0.000	0.015	0.267	0.275

Table 14: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	word length				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	5.445***	5.445***			
	(0.003)	(0.010)			
absolute score difference	0.318***	0.318***	0.316***	0.219***	0.215***
	(0.020)	(0.079)	(0.076)	(0.030)	(0.030)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	138,878	138,878	138,878	138,838	138,838
R^2	0.002	0.002	0.016	0.146	0.154

Table 15: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	word length				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	5.443***	5.443***			
	(0.004)	(0.011)			
absolute score difference	0.369***	0.369*	0.373*	0.356***	0.380***
	(0.064)	(0.169)	(0.155)	(0.093)	(0.088)
absolute score difference^2	-0.177	-0.177	-0.199	-0.482	-0.584
	(0.211)	(0.503)	(0.470)	(0.341)	(0.321)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	138,878	138,878	138,878	138,838	138,838
R^2	0.002	0.002	0.016	0.146	0.154

Table 16: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	profanity				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.127***	0.127***			
	(0.001)	(0.003)			
absolute score difference	0.236***	0.236***	0.239***	0.165***	0.164***
	(0.007)	(0.035)	(0.034)	(0.013)	(0.013)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.008	0.008	0.019	0.142	0.149

Table 17: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	sentiment				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.019***	0.019**			
	(0.002)	(0.006)			
absolute score difference	-0.489***	-0.489***	-0.502***	-0.311***	-0.308***
	(0.018)	(0.054)	(0.049)	(0.028)	(0.028)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.005	0.005	0.021	0.090	0.101

Table 18: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

	hashtags				
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.210***	0.210***			
	(0.004)	(0.026)			
absolute score difference	0.489***	0.489**	0.475**	0.275***	0.280***
	(0.029)	(0.166)	(0.161)	(0.045)	(0.045)
day Fixed Effects			Yes		Yes
sender Fixed Effects				Yes	Yes
Clustered SE		Yes	Yes	Yes	Yes
N	139,075	139,075	139,075	139,035	139,035
R^2	0.002	0.002	0.014	0.416	0.422

Table 19: Regression results (standard errors in parenthesis, standard errors are clustered on sender level, significance levels: * 5%, ** 1%, *** 0.1%)

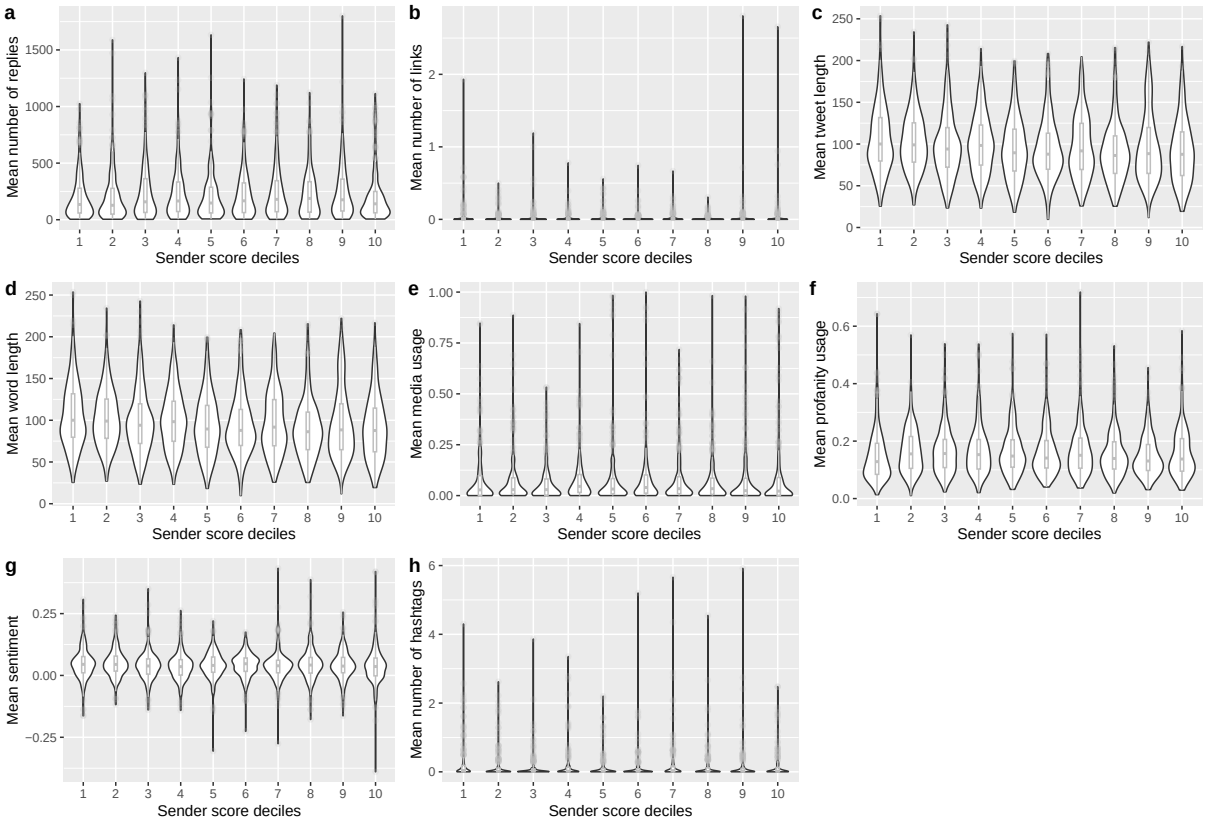

Figure 6: Violin plots of our main variables by sender decile, ranging from the most left-wing senders (decile 1) to the most right-wing ones (decile 10).

C. Supplementary Material: Documentation of data collection

In our data collection and analysis, we rely on the open-source programming languages R, Julia and Python, most notably on the rtweet package by Michael W. Kearney and the Julia-TextAnalysis package.

C.1. Tweets of members of Congress

C.1.1. Collection

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

static Twitter API. The collection took place between February 26 and March 3, 2021.[28]

C.1.2. Generating relative monogram/bigram frequencies

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

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

C.2. Main sample and reply sample

In the period February 12 to February 19, 2021 (a period that included the end of the second impeachment of Donald Trump) we used the Twitter Streaming API to stream tweets published by users that had activated geolocation and were within the GPS-coordinates between 65°W and 125°W as well as between 26°N and 49°N (which is, roughly speaking, the continental US). From the obtained 789 688 tweets, we discarded all retweets and reply tweets as well as those identified to be coming from users in Canada or Mexico. We then selected all remaining original tweets that included the words "Trump", "Biden" or "Congress". 12 467 tweets remained. We used the accounts that published those tweets and downloaded for each of those accounts all their tweets published after January 1st, 2019, using the static Twitter API. We restricted this sample further by excluding all accounts that had fewer than 500 original tweets or fewer than 20 replies.

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

[28] At the time of collection, the accounts of 11 Democratic and 15 Republican members of the 116th Congress were no longer available.

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

C.3. Scoring

We scored the political stance of the accounts in the main sample first by using using the bigram dictionary described in C.1.2: All original tweets of a user transformed to bigrams and those bigrams present in our frequency dictionary are used. The score of a user is then the average frequency among those bigrams, weighted by usage. More precisely,

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

where *text* is a set of all bigrams constructed from the original tweets of the user, *dict* is the set of bigrams in the bigram dictionary constructed from the Members of Congress tweets, $freq(bg)$ gives the frequency of Republican usage of bg in this dictionary.[29] We then use the monogram dictionary to produce a second score in the same manner and finally average the two scores to obtain our final score of a user's ideological position.

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

We then restricted our main sample slightly by deleting (i) all reply tweets for which the user replied to could not be scored (i.e. because his original tweets did not contain a monogram and a bigram from our respective dictionaries) and (ii) all reply tweets by those senders for which all users to which they replied had the same score (which in practice means that these users only replied to one single account). The reason for the latter is that in these cases there is no "within user variation" that we could exploit. This leaves us with 139,075 reply tweets sent from 2401 senders to 28,796 receivers.

C.3.1. Sentiment scores

Sentiment scores for each reply tweet were generated using the Vader sentiment score of Hutto and Gilbert (2014) from the Python package "vaderSentiment" that is geared towards sentiment analysis in the social media context (the variable in the dataset is *sentiment_vader*). We also created another sentiment score using the R-package "SentimentAnalysis" that implements the QDAP score of Hu and Liu (2004) in order to check all sentiment results for robustness.

[29] In case $text \cap dict = \emptyset$, the assigned score would have been 0.5. However, this did not occur for any user in our main sample.

C.3.2. Variables of interest

variable	explanation
receiver_score	receiver score
sender_score	sender score
abs_score_difference	absolute value of the difference between receiver and sender score
tweet_length	length of unstemmed tweet after removing links (words starting with "http://", "https://" or "www."), hashtags (words starting with "#") and directs (words starting with "@"), punctuation and additional whitespace
word_length	number of characters in a tweet divided by the number of words; tweet is not stemmed but links (words starting with "http://", "https://" or "www."), hashtags (words starting with "#") and directs (words starting with "@"), punctuation and additional whitespace have been removed
nHashtags	number of hashtags given by the Twitter API
nLinks	number of links given by the Twitter API
linkDummy	dummy that is 1 if at least one link is present
link_frequency	nLinks divided by tweet length
profanity_check	probability that a reply tweet contains profanity as judged by the Python package profanity-check; see `https://pypi.org/project/alt-profanity-check/` for details on profanity-check (*Accessed: May 27, 2021*).
sentiment_vader	sentiment score following the methodology of Hutto and Gilbert (2014)
sentiment_QDAP	sentiment score following the methodology of Hu and Liu (2004)
media	dummy that is 1 if a tweet contains media elements as given by the Twitter API (images and videos)
day	date of the tweet

References

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *Quarterly Journal of Economics 115*(3), 715–753.

Arendt, H. (1958). *The Human Condition*. University of Chicago.

Austen-Smith, D. (1990). Information transmission in debate. *American Journal of Political Science 34*(1), 124.

Austen-Smith, D. (1992). Strategic models of talk in political decision making. *International Political Science Review 13*(1), 45–58.

Austen-Smith, D. and J. S. Banks (2000). Cheap talk and burned money. *Journal of Economic Theory 91*(1), 1–16.

Bächtiger, A., J. S. Dryzek, J. Mansbridge, and M. E. Warren (2018). *The Oxford Handbook of Deliberative Democracy*. Oxford University Press.

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.

Chambers, S. (2018). The philosophic origins of deliberative ideals. *The Oxford Handbook of Deliberative Democracy*, 55–69.

Conover, P. J., D. D. Searing, and I. M. Crewe (2002). The deliberative potential of political discussion. *British Journal of Political Science 32*(1), 21–62.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica 78*(1), 35–71.

Gil de Zúñiga, H., S. Valenzuela, and B. E. Weeks (2016). Motivations for political discussion: Antecedents and consequences on civic engagement. *Human Communication Research 42*(4), 533–552.

Goyanes, M., P. Borah, and H. G. de Zúñiga (2021). Social media filtering and democracy: Effects of social media news use and uncivil political discussions on social media unfriending. *Computers in Human Behavior 120*, 106759.

Habermas, J. (1981). *Theorie des kommunikativen Handelns, 2 volumes*. Frankfurt am Main: Suhrkamp.

Habermas, J. (1983). *Moralbewusstsein und kommunikatives Handeln*. Frankfurt am Main: Suhrkamp.

Hamlin, A. and C. Jennings (2011). Expressive political behaviour: Foundations, scope and implications. *British Journal of Political Science 41*(3), 645–670.

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.

Jann, O. and C. Schottmüller (2023). Why echo chambers are useful. *Working Paper*.

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

Lilleker, D. G. and K. Koc-Michalska (2018). What drives political participation? motivations and mobilization in a digital age. In *Digital Politics: Mobilization, Engagement and Participation*, pp. 21–43. Routledge.

Little, A. T. (2023). Bayesian explanations for persuasion. *Forthcoming in Journal of Theoretical Politics*.

Milgrom, P. and J. Roberts (1986). Relying on the information of interested parties. *The RAND Journal of Economics*, 18–32.

Mill, J. S. (1859). On liberty.

Morey, A. C. and M. Yamamoto (2020). Exploring political discussion motivations: Relationships with different forms of political talk. *Communication Studies 71*(1), 78–97.

Persily, N. and J. A. Tucker (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Rogers, N. and J. J. Jones (2021). Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time? *Journal of Social Computing 2*(1), 1–13.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics 87*(3), 355–374.

StatSocial (2015). The most influential political journalists and bloggers in social media. `https://www.statsocial.com/social-journalists/`. Accessed: 2021-07-02.

Strandberg, K. and K. Grönlund (2018). Online deliberation. *The Oxford handbook of deliberative democracy*, 365–377.