Donald Trump, Partisan Motivated Reasoning, and Covid-19: A Partisan Filtered Theory of Backlash

James Bisbee,[1‡] Haohan Chen,[1] Richard Bonneau,[1,3,4]
Joshua A. Tucker,[1,2] Jonathan Nagler[1,2]

[1]Center for Social Media and Politics, New York University

[2]Politics Department, New York University

[3]Computer Science Department, New York University

[4]Center for Data Science, New York University

[‡]To whom correspondence should be addressed: james.bisbee@nyu.edu

August 2, 2024

Abstract

The polarization of attitudes on Covid-19 between liberals and conservatives in the United States was, and continues to remain, stark. But to what degree was Donald Trump individually responsible for this phenomenon? We exploit seven high-salience cues about the severity of Covid-19 from former President Trump over the course of 2020 to document a causal effect of Trump's efforts to downplay the seriousness of the pandemic on both liberal and conservative beliefs. We show that the most influential of these cues occurred very early on in the year when former President Trump compared Covid-19 to the seasonal flu. Subsequent cues, including the national State of Emergency, former President Trump's "liberate" tweets, the first time he wore a mask in public, and his positive Covid-19 diagnosis, all produced smaller albeit significant effects on public sentiment, which we measure with a bespoke sample of over 600,000 random Twitter users. We argue that the divergence in the public's concern about Covid-19's health risks following the flu comparison were driven largely by liberals growing more concerned about Covid-19, and propose an explanation for this "backlash" effect that recognizes elite cues are never experienced in a vacuum. Instead,

Trump's comparison was reported on by an ideologically diverse set of media accounts, whose coverage of Trump's cue influenced how their followers updated their concern, a framework we refer to as "Partisan Filtered Backlash". Our findings speak to the process by which opinions on Covid-19 became polarized, to the broader debate between the Folk Theory of representation and Partisan Motivated Reasoning, and propose a novel framework for understanding how elite cues are received in the information-dense reality of online social media environments.

Keywords: Covid-19, public opinion, elite cues, partisan motivated reasoning

Introduction

The process by which voters evaluate politicians is central to democratic representation, yet remains the point of much debate among political scientists. On one side of the debate lies a "folk theory" (Achen et al., 2017) in which voters compare the policy positions of political candidates to their own preferred policy, and select the candidate whose platform is closest to their own. On the other side of the debate is a "partisan motivated reasoning" (PMR) framework in which voters' policy preferences and political beliefs change to match the cues they receive from co-partisan politicians and media.

Adjudicating between these competing understandings is difficult because we rarely observe the evolution of a policy preference from its inception. For the vast majority of policy dimensions, they are either too obscure to matter to most voters, or are well-defined enough that it becomes difficult to untangle the causal flow between voter preferences and elite cues. As such, the mounting evidence of voters' policy preferences "following" elite cues (Lodge and Taber 2013; Lenz 2013; Kunda 1990; Bullock 2009; Achen et al. 2017; Freeder, Lenz and Turney 2019, but also see List et al. 2013; Fowler and Hall 2018) may simply capture the equilibrium of mainstream issues on which the party positions are well-known. In this paper, we leverage the outbreak of Covid-19 in the United States to evaluate this debate, building a novel dataset that captures the American public's perception of the seriousness of the virus beginning well before traditional surveys began including questions related to the pandemic.

Covid-19 is arguably a hard test of partisan motivated reasoning for three reasons. First, it is a surprising event in the sense that the American public was largely unaware of the virus as late as January of 2020. As such, there are very few partisan cues on which to rely prior to the outbreak. Second, Covid-19 is an extremely relevant and salient policy issue for everyone. Unlike many polarized topics that impact only a subset of the public, a

pandemic affects everyone. This means voters have more incentive to 'get it right' on the issue, and may seek out information beyond the party cue. Third, Covid-19 involved not just government policy decisions, where a rational voter might think they have no impact and thus little reason to bother to form an opinion, but also individual decisions. Voters did not just have to decide if government should mandate masks or social distancing, but they also had to make decisions of their personal behavior around masks and social distancing.

Yet despite these qualities which make this a theoretically difficult test for PMR, we demonstrate that attitudes on Covid-19 quickly and durably became differentiated along partisan lines, and that this was inspired by elite cues. Using a novel panel dataset of 610,775 randomly sampled Twitter users – a subset of which we can calculate both ideology and state of residence for – and applying cutting edge transformer classifiers to label more than 90 million tweets since January 1st, 2020, we construct a tweet-based proxy of public attitudes on Covid-19 which we aggregate to the individual-day level.[1] We then implement generalized difference-in-differences methods using these data to causally identify the effect of seven high-salience shocks to the public information environment surrounding Covid-19 over the course of 2020, demonstrating that liberal and conservative users diverged in their publicly expressed concerns as early as February 26th, 2020 in reaction to cue by former President Donald Trump that first downplayed the seriousness of the virus by comparing it to the seasonal flu.

Notably, we find that this divergence went in both directions, with conservative users expressing increasing skepticism about the seriousness of Covid-19's health risks *and* liberals expressing increasing concern. We theorize that this "backlash" among out-partisans in response to Trump's cue reflects the nature of real world information environments in which elite cues are never experienced in a vacuum, and ground our intuition in an aug-

[1]We refer to the tweets from each account as the tweets of a single individual. We do this knowing *some* accounts may be bots.

mented version of a standard Bayesian model of belief formation, which we refer to as the "Partisan Filtered Backlash Theory". Supplementing our random sample of Twitter users with similarly detailed information on local and national media accounts, we show that (1) more liberal media accounts respond to Trump's cues by emphasizing Covid-19's health risks and (2) these media accounts are followed by predominantly liberal users. These patterns provide an important but overlooked reality of motivated reasoning in which a partisan cue's dissemination is bundled with co and out-partisan frames that themselves contain influential signals about the state of the world.

1 Theoretical Framework

To give structure to our expectations about attitude formation, we summarize a standard Bayesian model of belief formation and apply it to our substantive setting of Americans learning about Covid-19 through a mixture of real world facts and elite cues (Zechman, 1979; Achen, 1992; Bartels, 1993; Druckman and McGrath, 2019). To start with a qualitative description of the Bayesian framework, we assume that an individual holds some belief about the state of the world, called a "prior". For example, an individual might think that Covid-19 is not a health threat. This individual then receives new information – a "signal" – that causes her to update her belief. These signals can take the form objective facts or elite cues. For example, she might learn that a neighbor has fallen ill from Covid-19 (an objective fact) or she might hear President Trump claim that the virus is no worse than the common cold (an elite cue). We assume that both signals will influence the individual's posterior belief, but that the degree of influence is a function of the signal's "credibility". Continuing the example, if the individual is a Trump supporter and views his cues as credible, she might believe even more strongly that the health risks associated with the virus are overblown. Conversely, if her neighbor dies, she might revise her prior to believe that the health risks are indeed severe, due to the higher credibility of a death compared to a mild case of the

disease. We reproduce the formalization of this model used by Druckman and McGrath (2019) below.

For state of the world μ, the individual i's prior belief $\pi_i(\mu)$ can be expressed as a probability distribution:

$$\pi_i(\mu) \sim \mathcal{N}(\hat{\mu}_{i,0}, \hat{\sigma}^2_{i,0})$$

where $\hat{\mu}_{i,0}$ is the individual's belief about μ and $\hat{\sigma}_{i,0}$ is the individual's uncertainty surrounding her belief. Starting from this prior belief, an individual can encounter a new piece of information – referred to as a signal and denoted with x – which is similarly drawn from a distribution defined as $\mathcal{N}(\mu_x, \hat{\sigma}^2_{i,x})$. Note that μ_x can be equal to μ (meaning it is an unbiased reflection of the true state of the world), or it can be centered on some value $\mu_x \neq \mu$. Furthermore, note that $\hat{\sigma}^2_{i,x}$ is subjective and a function of two factors. The first factor is the salience of the signal, reflecting the fact that more salient signals are clearer, reducing uncertainty. The second factor is the credibility of the signal, capturing the extent to which individual i trusts the source of the signal. The individual's posterior belief is thus:

$$\pi_i(\mu|x) \sim \mathcal{N}\left(\hat{\mu}_{i,0} + (x - \hat{\mu}_{i,0})\left(\frac{\hat{\sigma}^2_{i,0}}{\hat{\sigma}^2_{i,0} + \hat{\sigma}^2_{i,x}}\right), \frac{\hat{\sigma}^2_{i,0}\hat{\sigma}^2_{i,x}}{\hat{\sigma}^2_{i,0} + \hat{\sigma}^2_{i,x}}\right)$$

This formalization provides a structured framework for understanding the public's polarized response to Covid-19 in 2020 and, in turn, speaks to the debate between the folk theory of democracy and partisan motivated reasoning. At the heart of divergence lies the individual-specific credibility term $\hat{\sigma}^2_{i,x}$. This term allows for two individuals i and j to start from the same prior ($\pi_i(\mu) = \pi_j(\mu)$), receive the same signal from the same source (μ_x), but differ in the credibility they assign to this source ($\hat{\sigma}^2_{i,x} \neq \hat{\sigma}^2_{j,x}$). Even in this stylized setting where individuals start from the same prior and receive the same signal, their updated beliefs will diverge unless the credibility they assign to the signal is the same.

To give an example, consider a Democrat and a Republican who both start from the same prior that Covid-19 is dangerous, held with the same uncertainty, and receive the same information from a cue given Donald Trump. The Democrat assigns a very low credibility to this cue, while the Republican assigns it a very high credibility – i.e., $\hat{\sigma}^2_{Dem,Trump} >> \hat{\sigma}^2_{Rep,Trump}$. Accordingly, the Republican will update their beliefs more dramatically while the Democrat's beliefs will remain virtually the same, as visualized in Figure 1. Importantly, the posterior position becomes a new prior for ensuing information. As illustrated, the Republican is less uncertain about their new prior than is the Democrat, making it harder for them to update further.

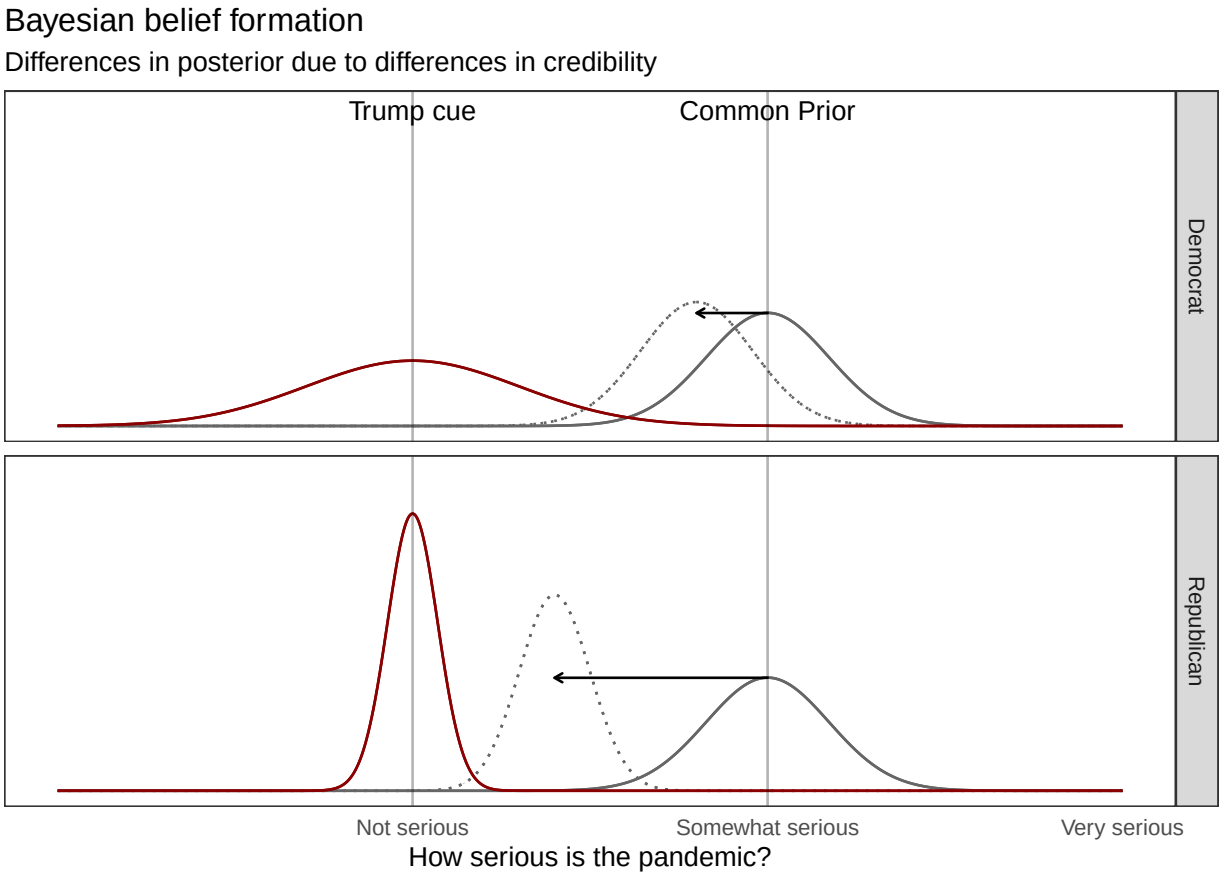

Figure 1: Two individuals with identical priors and who receive the same message update differently based on the credibility that they assign to the message. Democrats (top) assign low credibility to the cue provided by Trump, as indicated by the wide distribution. Republicans (bottom) assign high credibility to the same cue, as indicated by the narrow distribution. As such, Republicans move further toward Trump's cue than Democrats.

This framework can be applied to any information received by an individual. For example, knowing someone with the virus is also a piece of information with its own credibility, as is information from Fox News and CNN, as are the competing messages from other politicians, as are statewide policies restricting the size of gatherings. Greater divergence will occur when the signals themselves differ. If one set of signals from a source that individual i assigns high credibility to are centered on μ_x, and another set of signals from a source that individual j assigns high credibility to are centered on μ_y where $\mu_x \neq \mu_y$, the divergence in beliefs will be exaggerated. Partisan motivated reasoning occurs where there are partisan differences in the credibility assigned to different signals, and where these signals carry different information. [2]

The folk theory of democracy would expect that all voters assign greater credibility to objective facts than to elite cues, meaning $\hat{\sigma}^2_{i,facts} << \hat{\sigma}^2_{i,cues} \forall i \in \mathcal{P}$ where $\mathcal{P}$ is the voting population. Conversely, the essence of partisan motivated reasoning predicts that different voters assign different credibility to the same elite cues, such as when Democrats discount statements by President Trump while Republicans treat the same statements as gospel.

1.1 "Backlash" in Motivated Reasoning

An open question in the research on partisan motivated reasoning is whether individual's might ever update *away* from a signal. For example, does a liberal who learns about Trump's stance on an issue update their posterior belief in the opposite direction? Given the contentious nature of politics in the United States, it is not hard to imagine an attitude of "if they said it, it must be wrong" becoming ever more prevalent. Yet survey experimental evidence suggests that "while backlash may occur under some conditions with some individuals,

[2]The possibility of divergence is further exaggerated if the credibility is itself a function of an individual's prior ("prior attitude effect"). Prior attitude effect posits that people have a cognitive need for their priors to be upheld, meaning that $\hat{\sigma}^2_{i,x} = f(||\mu_x - \mu_{i,0}||)$ where $f'(||\mu_x - \mu_{i,0}||) > 0$. [ADD SOME CITES HERE]

it is the exception, not the rule." (Guess and Coppock, 2020, pg 1500)

Given the active nature of the debate over backlash, we take special care to define precisely what we can and will measure in this paper. While there are many theoretical models for expecting backlash (Zaller 1992; Lodge and Taber 2013; Kahan et al. 2007; see Guess and Coppock 2020 for a synthesis), we rely on the Bayesian framework for its flexibility, and start from the observation that elite cues such as Trump's statements on Covid-19 do not appear in a vacuum. They are re-interpreted and reacted to by media and other political elites such that the public rarely ever confronts a single elite cue in isolation, but rather as a sound bite in an ideological context. In the Bayesian setting, any raw cue centered on μ_x^{Trump} prompts an offsetting cue μ_x^{Libs} from out-partisan elites. Both of these cues are associated with a credibility parameter that, as above, is a function of the recipient. Thus, a liberal's Bayesian response to Trump downplaying the severity of the virus is a product of Trump's raw cue μ_x^{Trump} to which very little credibility is assigned ($\hat{\sigma}^2_{Dem,Trump}$ is large), *and* the liberal media's response cue μ_x^{Libs}, to which a great deal of credibility is assigned ($\hat{\sigma}^2_{Dem,Libs}$ is small). We visualize an example of this backlash theory in Figure 2.[3]

It is worth underscoring that our framework and empirical results are not incompatible with the existing consensus that a given cue does not, on its own, produce a backlash effect (Coppock, 2023; Wood and Porter, 2019; Porter and Wood, 2022). Most notably, our observational setting sacrifices the tight internal validity of the existing evidence against backlash for the ecological validity of the saturated information environments that individual's actually exist in. In this setting, backlash effects can be observed even when, at the level of the individual-signal dyad, we would never theorize that the individual's posterior can move away from the signal. Put differently, while it may almost always be true that an

[3]Even raw cues might produce a backlash if out-partisans respond by searching for new information to contradict the signal, or even simply respond rationally. For example, liberals who grow more concerned about Covid-19 in response to Trump downplaying the health risks may be rationally anticipating that the spread of the virus will be more destructive if half of the population refuses to take necessary precautions.

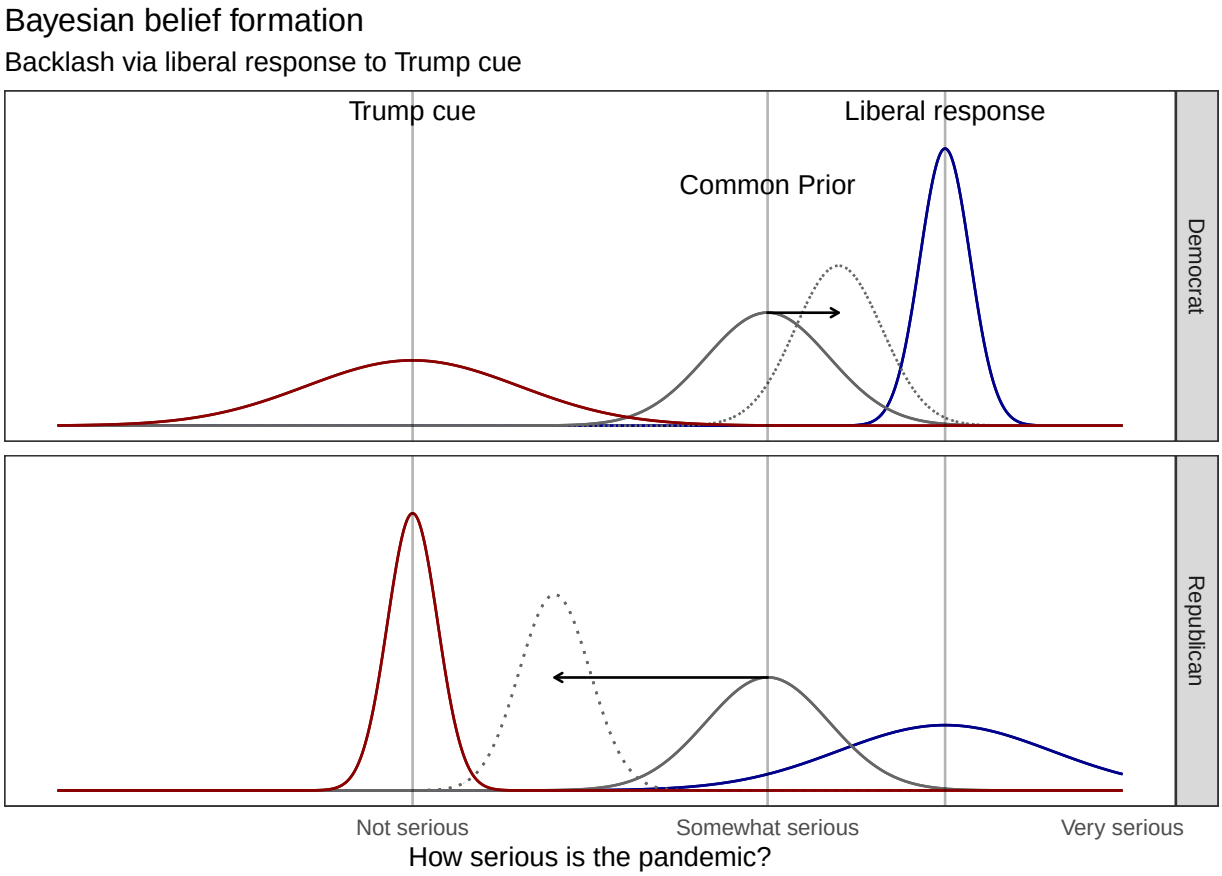

Figure 2: Two individuals with identical priors and who receive the same message update differently based on the credibility that they assign to the message, and the liberal response to Trump's cue. Democrats (top) assign low credibility to the cue provided by Trump, as indicated by the wide distribution, and high credibility to the liberal response. Republicans (bottom) assign high credibility to Trump's cue, as indicated by the narrow distribution, but low credibility to the liberal response. As such, Trump's cue appears to backfire among Democrats.

individual will never backlash against a given signal, it is rarely ever true that an individual encounters a given signal in isolation from a broader information environment of conflicting interpretations of that signal.[4]

[4]Insofar as our study better captures the reality of the United States in 2020 as the country grew increasingly divided along political lines about the appropriate response to Covid-19, we argue that our trade-off between internal and ecological validity is defensible. Nevertheless, we are sensitive to the inferential challenges we must confront by relying on observational data, instead of the more tightly controlled survey experiments used in the existing research (Guess and Coppock, 2020). Where appropriate, we flag these limitations and encourage future research to find ways to overcome both limitations in a single study.

2 Data and Methods

2.1 Tweet-Based Measures of Beliefs

Traditional public opinion surveys suffer from three limitations that complicate our goals. First, most surveys didn't start asking questions about Covid-19 until March 2020 at the earliest.[5] Second, public opinion surveys typically measure random samples of US adults re-drawn for each survey, preventing us from documenting within-individual changes in attitudes and exposure that would augment our causal interpretation. Finally, most public opinion surveys are fielded sporadically over time, preventing us from narrowing our analyses to just prior to, and immediately following, high-salience signals which would further augment our causal interpretation.

As such, we turn to a novel source of information on public attitudes: Twitter. We rely on a bespoke panel random sample of over 600,000 Twitter accounts which was drawn in 2018, prior to public knowledge about Covid-19. We scraped the timelines of these accounts every two months over the course of 2020. In addition, we estimated both the ideology and the geographic location (the US state) for each account.[6] Our full dataset is comprised of over 90 million tweets written in 2020, associated with 610,775 unique Twitter accounts randomly sampled from the 2019 universe of US-based English speaking accounts.

Nine research assistants manually labeled a random sample of 40,000 of these tweets that contained one or more Covid-19 keywords.[7] Specifically, they were asked to read the

[5]The earliest reference to the "coronavirus" appears in a PBS poll from January 31st, 2020. More systematic time-series cross sectional data comes from Nationscape, which first asked about the "coronavirus" on March 19, 2020.

[6]Detailed descriptions of the data and methods used to estimate ideology and location can be found in the SI.

[7]We relied on Twitter's own Covid-19 keywords for this filtering step, the full list of which can be found in the SI.

tweet and determine whether it expressed concern about Covid-19's health risks, or expressed skepticism about them. We then trained a BERT-based transformer model on these labeled data, achieving good out-of-sample performance on both categories [HC TO FILL IN F-1 SCORES]. Our final measure of interest is the algorithm's predicted probability that a given tweet expresses concern, minus the predicted probability that it expresses skepticism, resulting in a continuous number that ranges from -1 (meaning extreme skepticism) to +1 (meaning extreme concern). We refer to this measure as NET COVID CONCERN in the remainder of our analysis.

2.2 The construct validity of our measure

A natural concern with our reliance on tweets as proxies for public beliefs on Covid-19 is that the manner in which people present themselves online is potentially inconsistent with their true attitudes. Expressive goals such as social desirability or group signaling might be elevated in public but pseudo-anonymous forums such as Twitter. Furthermore, there is the potential for substantial selection bias, despite our reliance on a random sample of Twitter users, since Twitter itself is not representative of the broader US population [CITE PEW]. Finally, even if we could overlook these limitations, there is the final selection bias implicit in who chooses to tweet about Covid-19, and thus appear in our data. Despite the advantages of our rich time series panel discussed above, one might be skeptical about what we can actually learn about belief formation using Twitter data.

We acknowledge this concern, and engage with it in the framework of validity, which we break into three components: internal validity, external validity, and construct validity. On the first two dimensions, we acknowledge the magnitude of the patterns we document may be biased upwards or downwards depending on how one interprets the social desirability and selection biases inherent in our measure. For example, if tweets are written to signal group affiliation (i.e., liberals express anger and disdain in response to conservative cues

more readily online than they would in reality), this would exaggerate the magnitude of ideological polarization on Covid-19. Relatedly, by relying on Twitter users who are known to be younger, better educated, and more liberal than the general US population (at least in the period that we analyze), we might be biased toward finding stronger evidence of a liberal backlash against Trump's cues. On the third dimension, we argue that even expressive content is still inherently constrained to be reflective of the user's true beliefs. Yes, there are trolls, but we argue they are of a sufficiently small minority of our users that the expected value of our tweet-based measures nevertheless correlate with the underlying beliefs of the authors.

In spite of these issues, we further defend the value of our study in three ways. First, we surveyed a subset of authors in our Twitter random sample in 2020, where we explicitly asked them about their concern about Covid-19. We show that our tweet-based measure of concern recovers the survey-based measure reasonably well (correlation XXX, see SI Section XXX). Second, even if the reader remains skeptical that our tweet-based measure is a useful proxy for latent attitudes, we believe that understanding the information environment of Twitter is valuable in its own right, especially given its prominence among political and media elites in 2020. Finally, the panel nature of our data means that we are characterizing changes in this measure, providing plausibly causally identified evidence of how high-salience cues influence online discourse.

2.3 Analytic datasets

Our main analyses aggregate these 90 million tweets written by more than 600,000 users down to the ideological group-state-day, where we classify people in one of six ideological groups ranging from extreme liberals to extreme conservatives, and include a missing category for the accounts whose ideologies we were unable to calculate (account ideology scores were calculated using Barberá (2015), although we confirm our conclusions are robust to

alternative approaches, including Eady et al. (2020) and Wu et al. (2023)). Similarly, we geolocated accounts to states, and include a 51st state category for the accounts we were unable to geolocate, relying on the method developed by CITE. Our resulting aggregate dataset is 111,690 rows long, and contains measures of each state's ideological group's net concern about Covid-19. Our main outcome of interest is the difference in the probability that a user's tweet expresses concern minus the probability it expresses skepticism, which is then aggregated up to the state-ideological group-day unit.[8]

Our main predictor of interest is a series of signals which we argue constitute discontinuous changes in the public's information environment surrounding Covid-19. These include the first reported US cases (January 18th, 2020); Donald Trump's announcement to restrict international travel (January 31st, 2020); his initial comments comparing the virus to the seasonal flu (February 26th, 2020); the State of Emergency (March 13th, 2020); Trump's "liberate" tweets which called for residents of Democrat-governed states to protest Covid-related restrictions (April 17th, 2020); Trump's first time wearing a mask in public (July 12th, 2020); and Trump's positive diagnosis for Covid-19 (October 2nd, 2020). Figure 3 provides a descriptive snapshot of these data, including the total tweets (in gray) and Covid-related tweets (in red) in the top panel; the corresponding proportion of tweets written about Covid-19 in the middle panel; and the net difference between the concern and skepticism measures by ideological group in the bottom panel. Our main predictors of interest are highlighted with vertical black lines, excepting the travel ban which is omitted for legibility.

In addition to these signals, we also include two prominent controls to hold constant time-varying confounding from the disease environment and the policy response of state and local governments. Specifically, we merge this dataset with daily measures of new Covid-19 cases by state, as well as a detailed accounting of the relevant local policies that were

[8]We treat each user's expressed concern as zero up until their first tweet appears in our dataset. Subsequent gaps in their posting behavior are imputed as the same level of concern as what was most recently measured.

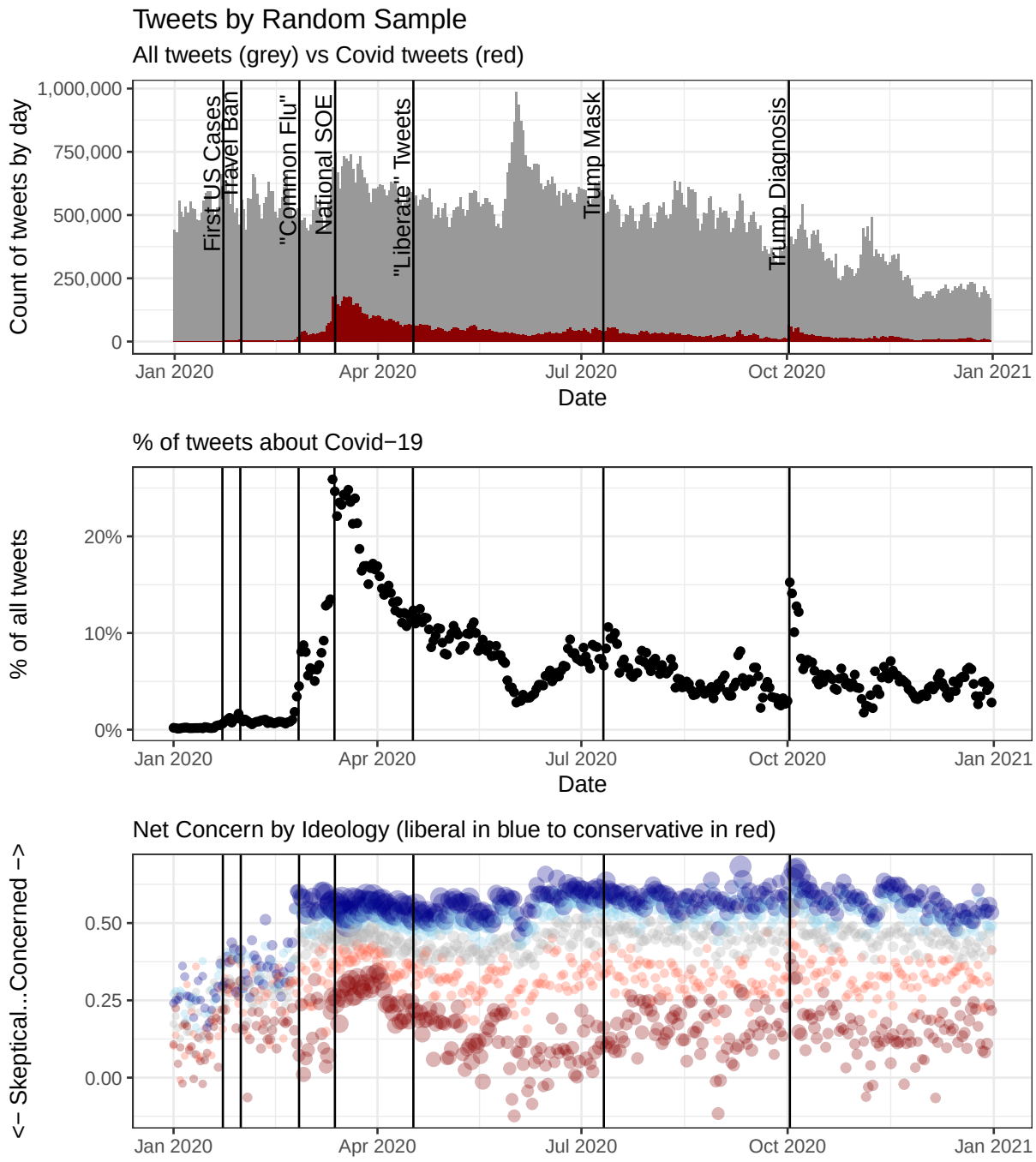

Figure 3: *Top Panel:* Total tweets (gray) and Covid-related tweets (red) by the random sample. *Middle Panel:* Proportion of tweets about Covid-19 by the random sample. *Bottom Panel:* Predicted concern about the pandemic minus predicted skepticism about the pandemic by ideological subset (very liberal in dark blue, liberal in blue, moderate in gray, conservative in red, very conservative in dark red).

implemented on a state-by-state level.

3 Methods

Our estimand of interest is the public's updated net concern about Covid-19 after receiving a signal, conditional on the credibility the public attaches to this signal. To estimate this quantity, we rely on two related statistical techniques which exploit over-time variation in these signals. The simplest strategy is to compare the public's concern prior to, and following, the issuance of a signal. For example, we might compare the degree to which the public was worried about Covid-19's health risks before and after the first recorded cases appeared in the United States. Similarly, we could make the same comparison before and after Trump tweeted that the virus was no worse than the common flu. To capture the credibility component of interest, we can compare the strength of this response among different ideological subsets, assuming that conservatives attach greater credibility to signals from Trump than liberals do.

Interpreting this relationship as a causal quantity is complicated by the richness of the public's information environment, as well as by the possibility of reverse causality. For example, if we find that the public took Covid-19 less seriously following Trump's comparison of the virus to the flu, does this mean that Trump's signal changed the public's views? Or does it reflect a shift in the attitudes of other elites such as the media, to which both the public and Trump reacted? Or might it mean that Trump accurately predicted which direction public sentiment was shifting, and capitalized on this intuition?

One solution to these challenges is to narrow our focus to just prior to, and just following, the signal itself. While many things might change over the course of the month prior to a signal, there are fewer potential confounders over the preceding week, and fewer still over the preceding day. Thus our simplest empirical strategy is to make this before-after

comparison at increasingly narrow windows.

However, increasingly narrow windows come at the cost of power. 30 days of measures times 6 ideological categories times 50 states yields 9,000 observations with which to calculate the pre- and post-signal attitudes, while a single day only yields 300. This tension is exacerbated by the inherently noisy proxy we rely on for measures of public sentiment. One solution is to implement an interrupted time series (ITS) design in which we model the pre- and post-signal trends and then compare the gap in the predicted attitudes at the moment the signal is issued. We strengthen our assumption that our measure captures the causal effect of the signal on beliefs by narrowing the temporal window, thereby making the alternative stories described above less plausible. Formally:

$$concern_{g,t} = \alpha + \beta_{11} T + \beta_{12} \mathbb{I}(t > t_0) + \beta_{13} T_{t>t_0} + \varepsilon \tag{1}$$

where g indexes the group of the public (ideological group by state) and t indexes days. t_0 is the date of a high-salience cue from Donald Trump, T is a variable counting the number of days as integers from the start of the period of analysis, $T_{t>t_0}$ is a second integer variable counting the number of days from the high-salience cue, and $\mathbb{I}(\cdot)$ is the indicator function that assigns the value 0 for $t \leq t_0$ and 1 for $t > t_0$. We include two subscripts on the coefficients of interest to indicate that these are associated with regression equation 1, and to differentiate these from subsequent specifications that also denote coefficients with β.

Figure 4 plots the results of the estimation strategy for the first recorded U.S. cases on January 20th, 2020, with β_{11} capturing the slope of the line measuring net concern over time prior to this date and β_{13} capturing the slope after this date. β_{12} captures the discontinuous shock to concern associated with the novel signal about Covid-19 embodied in the first recorded U.S. cases, and is the quantity of primary interest for our results. However, we are also interested in the comparison between β_{11} and β_{13} which capture changes in the longer run relationship between concern and time. As illustrated in Figure 4, not only do the first

cases produce an increase in concern ($\beta_{12} = 0.07$, p-value < 0.001), but they also shift the public from a period of static concern ($\beta_{11} = 1e-04$, p-value $= 0.95$) to one of growing concern ($\beta_{13} = 0.012$, p-value < 0.001). Importantly, these two slopes are significantly different from each other ($\beta_{12} = 0.07$, p-value < 0.001), indicating that this signal influenced both the immediate level of concern as well as the long-run dynamics in concern.

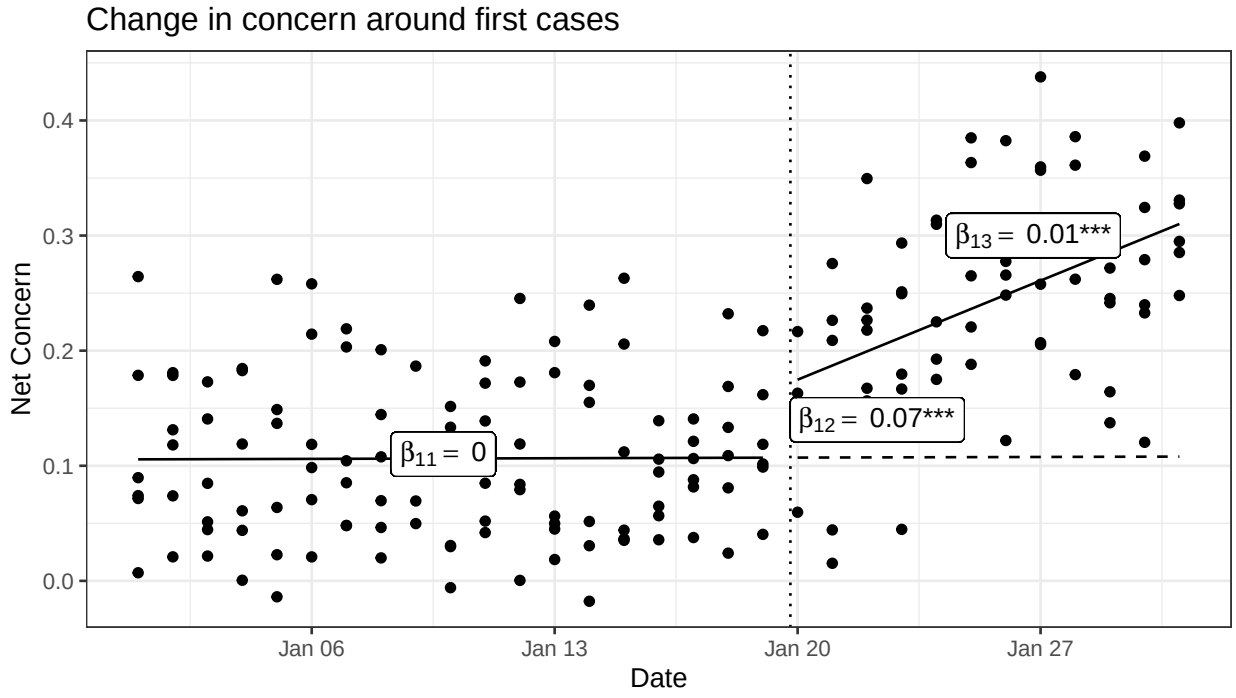

Figure 4: Example of interrupted time series analysis (ITSA) using the first reported U.S. cases on January 20th, 2020. β_{12} captures discontinuous change in concern associated with the signal, while the comparison of β_{11} and β_{13} describes how the signal changed the over-time dynamics of concern, moving from a steady-state to growing concern in the weeks following the first cases.

Note that we are also interested in the heterogeneity of the updating by ideological group, which embodies the assumption about which signals are assigned more or less credibility: we want to know if conservatives will update more than liberals based on a signal from Trump. Here we adopt a difference-in-differences approach, where (for example) conservatives who assign high credibility to Trump are thought of as the treated group, while liberals who assign low credibility to Trump are thought of as the control group. Drawing a causal inference here requires us to make the "parallel trends" assumption: we assume that

the liberal-conservative gap in beliefs would have persisted were it not for the new signal which affected one group more than the other. In other words, we assume there is a mean shift in sentiment between the two groups following the signal, but there is no shift in the relative slopes between the two groups from the signal. Formally:

$$concern_{g,t} = \alpha + \beta_{21} Cons + \beta_{22} \mathbb{I}(t > t_0) + \beta_{23} (\mathbb{I}(t > t_0) * Cons) + \varepsilon_{g,t} \tag{2}$$

where, as above, g indexes groups of the public and t indexes time in days. Again, we provide an example in Figure 5 using our data to make clear the empirical approach. β_1 is the difference between liberals and conservatives prior to the signal ($\hat{\beta}_1 = -0.103$, p-value < 0.001), β_2 is the difference in concern for liberals between the period prior to and the period following the signal ($\hat{\beta}_2 = 0.153$, p-value < 0.001), and β_3 is the difference in the liberal-conservative difference between the periods before and after the signal ($\hat{\beta}_3 = -0.015$, p-value $= 0.51$). In this setting, the β_3 coefficient has a natural interpretation that is relevant for our research question: it is the amount that one group reacts more or less than the other group to the signal. If it is not different from zero, this implies that different ideological groups do not react differently to the signal.[9] So we can see that for "First Cases": there was no observable difference in how liberals and conservatives reacted, although both groups grew substantially more concerned.

[9]Note that this might be because both groups update to a signal in the same manner, or it might be because neither group reacts to a given signal. Adjudicating between these interpretations is facilitated by the β_2 coefficient. If this is significantly different from zero, both groups update in the same manner. If it is not, the signal does not affect the concern of either conservatives or liberals. In Figure 5's example, it is the former interpretation in which the first reported U.S. cases dramatically increased concern among both groups by approximately equivalent magnitudes.

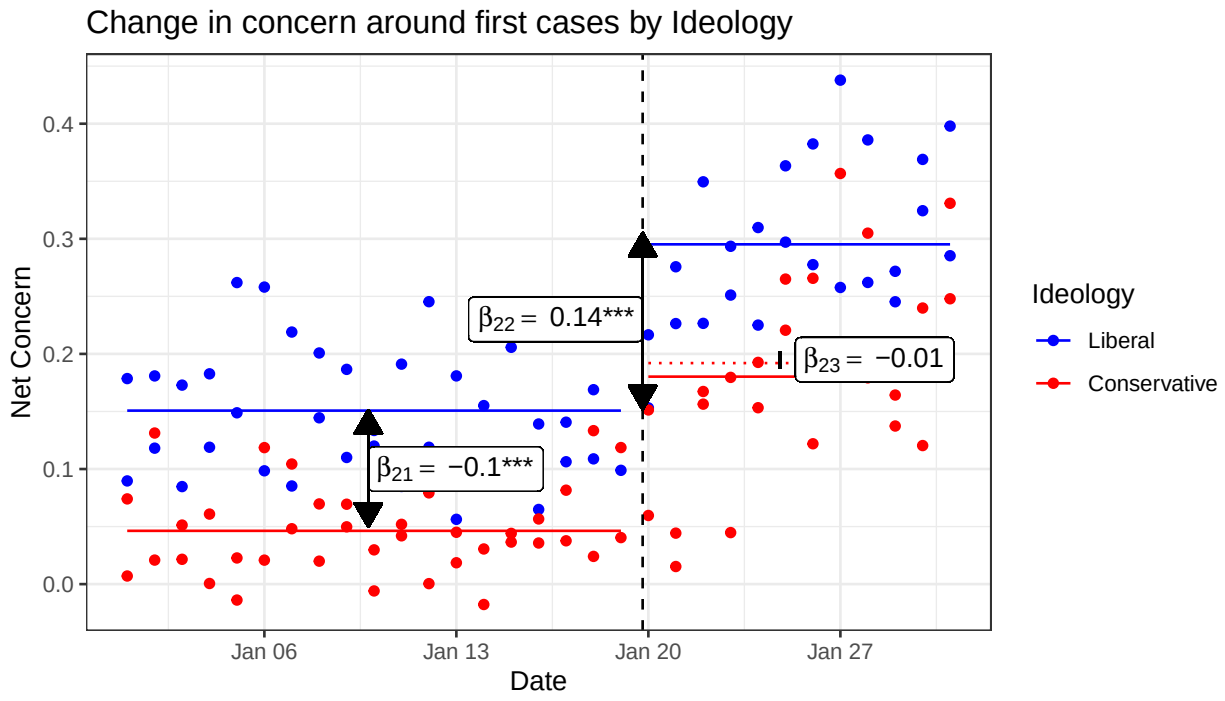

Figure 5: Example of difference-in-differences using the first reported U.S. cases on January 20th, 2020. While conservatives (red) were already less worried than liberals (blue) about the virus, and while both groups more than doubled their concern immediately following the first U.S. cases, there is no evidence that this signal caused them to grow further apart. The dotted red line is the counterfactual prediction implied by the parallel trends assumption, against which the observed outcomes (solid red line) are compared.

3.1 Assumptions required for a causal interpretation

Given the observational nature of our data, we require the following assumptions to interpret the coefficients described above as causally identified estimands. We recognize that none of these events are truly unexpected and, in some instances, are clearly endogenous to the outcomes of interest. For example, the first reported cases in the United States carried a signal (namely that the virus would not be contained in Asia, as MERS was a decade earlier) that had already become clear prior to the specific date of January 21st, 2020. Similarly, the flurry of Covid-skeptic signals sent by Trump in the third week of February had already started in the preceding weeks when asked about the virus. Even truly unprecedented signals, such as a sitting president goading his supporters to "liberate" themselves from state policies, are endogenous to the underlying public sentiment that prompted these protests and rallies in the first place.

Nevertheless, we feel confident in asserting that these events capture discontinuous changes in the level of the each signal in the overarching information environment inhabited by the main actors of interest to our study. The first U.S. cases, while not unexpected, corresponded to a dramatic spike in attention among the media and the public alike. The "liberate" tweets of April 17th crystallized Trump's stance on preventative state policies and was widely reported on as an alarming call to arms that broke with Trump's prior complaints about Democrat-led lockdowns. Even Trump's positive diagnosis on October 5th, while perhaps not unforeseen given his history of flouting recommendations from the medical community, was nevertheless a depth charge in what had grown to be an almost-business-as-usual vein of Covid-19 media coverage that was a diminishing component of the election cycle of that fall. Similar arguments can be made for all the events we identified over the course of 2020.

As such, our causal interpretation relies not on the novelty of a given signal, but rather on the discontinuous shift in its salience, which connects back to the Bayesian model

of preference formation described above. In the context of the Bayesian framework, each of these events represents a stark before-after comparison in which the clarity $\hat{\sigma}^2_{i,x}$ of a signal about the state of the world changes substantially. Importantly, these discontinuities are shocks to the salience component of $\hat{\sigma}^2_{i,x}$ but not to the subjective credibility assigned to the source. As such, we can interpret ideological divergences in the response to the same signal as capturing the credibility component of $\hat{\sigma}^2_{i,x}$, isolating the substantive quantity of interest: the extent of partisan motivated reasoning.[10]

4 Results

We start with an interrupted time series analysis of how each of the 7 signals influenced the overall chatter about Covid-19 on Twitter among our random sample. Specifically, we compare the proportion of all tweets that contained one or more of our Covid-19 keywords just prior to, and just following, each signal. Figure 6 displays the results, indicating the β_2 estimate and statistical significance, along with a test of whether the β_1 and β_3 estimates are equal to each other and the associated p-value. As illustrated, of the seven total signals, only three are statistically significant predictors of the total amount of chatter on Twitter about Covid-19: Trump's comparison of the virus to the seasonal flu, the national state of emergency (SOE) on March 13th, 2020, and Trump's Covid-19 diagnosis in early October, 2020.

While useful to keep in mind as we examine the content of what was written in these tweets, the findings summarized in Figure 6 aggregate over all ideological groups and states. But what about the polarization in concern in response to these events? We investigate this question in two ways. First, we create a new outcome measure that is the absolute differ-

[10]There is also the concern that we have either forgotten or neglected other high-salience cues from 2020. We emphasize that we started work on this project in May of 2020 and, as researchers studying Covid-19, have been attentive to the public discourse surrounding the virus.

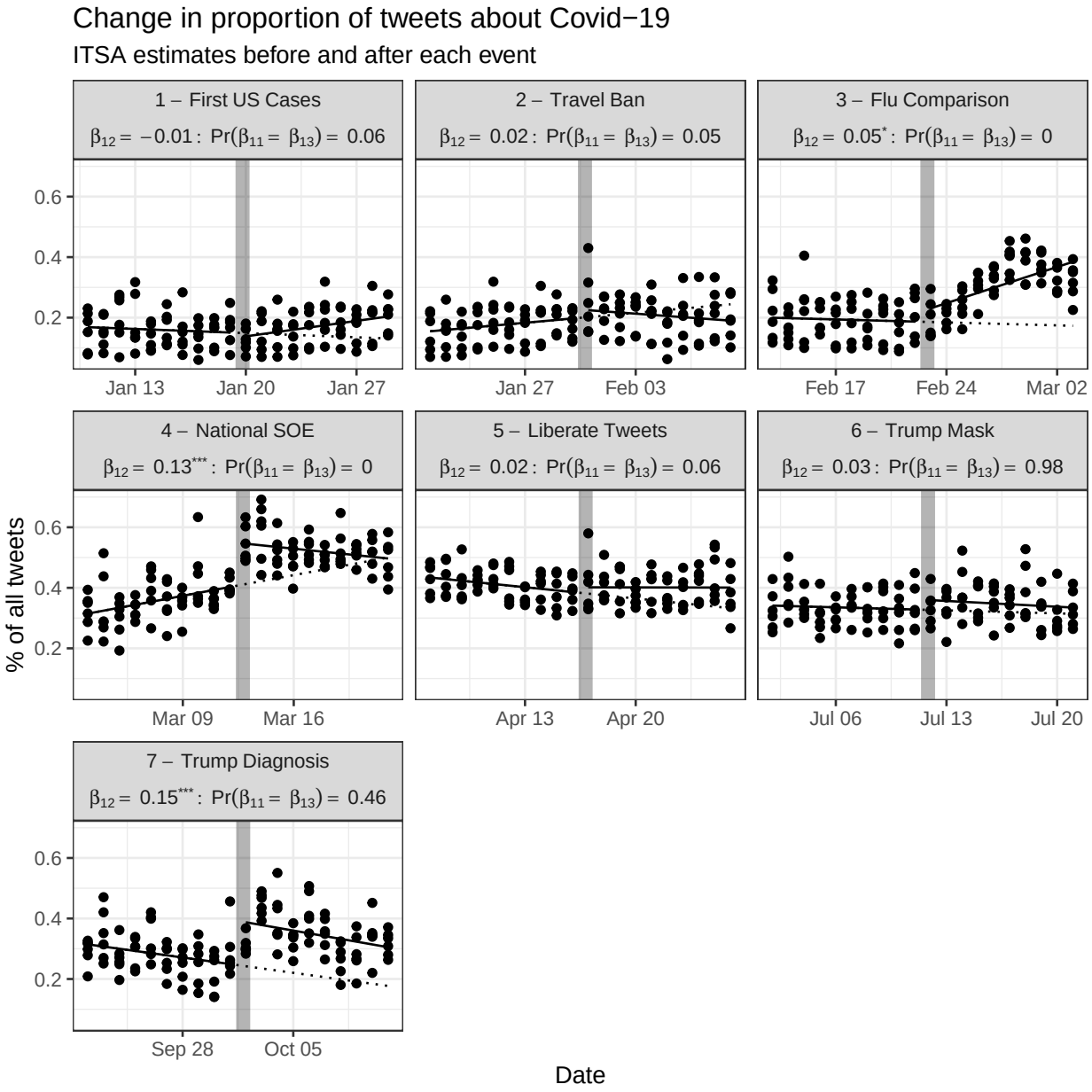

Figure 6: ITS results using the proportion of total tweets that were about Covid-19 as the outcome (averaging by ideological groups of users over all states for each day, indicated by black points). Thick gray vertical lines indicate the event in question, while thin black lines indicate the linear model fit prior to, and following, each event (β_1 and β_3, respectively. Dotted lines indicate the counterfactual of how the outcome would have appeared absent the event. The difference between the pre and post black lines (β_2) captures the immediate "effect" of the event, while the label indicates whether the slope of the pre and post trends differ.

ence in concern between liberals and conservatives and re-run the simple ITS specification. We plot the β_2 coefficients for each event in Figure 7, which captures the difference in the gap in beliefs just prior to, and just following, each new signal, defining the gap between extreme ideologues (red circles), moderate ideologues (blue squares), and both groups combined (green triangles). As illustrated, the flu comparison stands apart in terms of statistical and substantive significance, increasing the gap between liberals and conservatives by over 15 percentage points, and almost 20 percentage points if we compare only ideologically extreme accounts. The only other significant events are the state of emergency and Trump's positive diagnosis, both of which saw a reduction in ideological differences in perception of Covid-19's health risks. There is some evidence that Trump's decision to first wear a mask in public also polarized concern, although this is driven by the ideologically extreme groups.

But does this increase in the gap reflect conservatives growing more skeptical, liberals growing more concerned, or a mixture of both? To answer this question, we turn to a more rigorous test of the ITS specification in which we interact each component with indicators for the ideology of the public, control for time-invariant confounders at the state level with fixed effects, and further control for the weekly change in Covid-19 cases and the weekly change in deaths per 100,000 people. Figure 8 summarizes the results. Each point indicates the pre-signal predicted concern, and vertical lines capture the β_2 coefficient from the ITS specification. Thick solid arrows indicate that the shift is statistically significant at the 0.01 level. For example, the first reported U.S. cases occurred on January 20th, 2020. According to our model, all groups became more concerned about the pandemic following the first cases. Substantively, concern about the seriousness of Covid-19 roughly doubled between the two weeks prior to, and the two weeks following, the first reported cases on January 20th.[11] Importantly, all ideological groups increased, with the biggest movers found among liberals and extreme liberals.

[11]The date of the first cases was retrospectively updated to January 18th, 2020. However, at the time the first reports of these cases were January 20th.

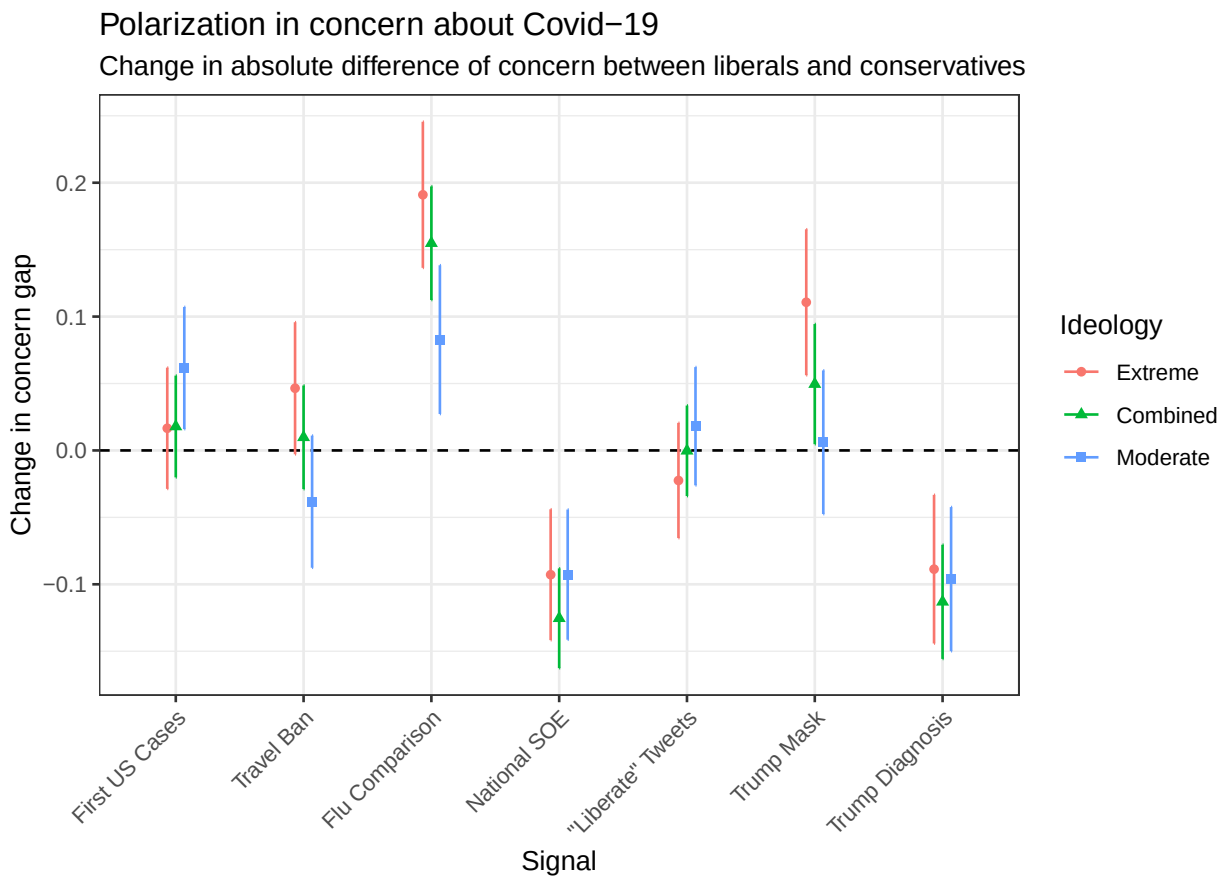

Figure 7: β_{12} coefficient estimates from Equation 1 predicting changes in the liberal-conservative concern gap (y-axis) by signal (x-axis). Two standard errors indicated with bars. Regressions run on two-week window pre and post the signal. Comparison between extreme liberals and extreme conservatives given in red circles, comparison between moderate liberals and moderate conservatives given in blue squares, and overall comparison between all types of liberals and all types of conservatives given in green triangles.

This stands in stark contrast to the before/after comparison for February 23rd, 2020 which is when Donald Trump began sending signals that he thought the health concerns of the virus were overblown with his seasonal flu comparison. For three out of our five ideological groups, concern about Covid-19 experienced a statistically significant discontinuous jump after these tweets. Among conservatives, concern barely changed. But it is among extreme conservatives where the results are most striking, corresponding to a decline in concern of almost 15 percentage points. The result of this two week period is a clear and stark divergence in the public's views on the pandemic that mapped on to ideology, driven by

increasing concern among liberals and moderates, and declining concern among extreme conservatives.

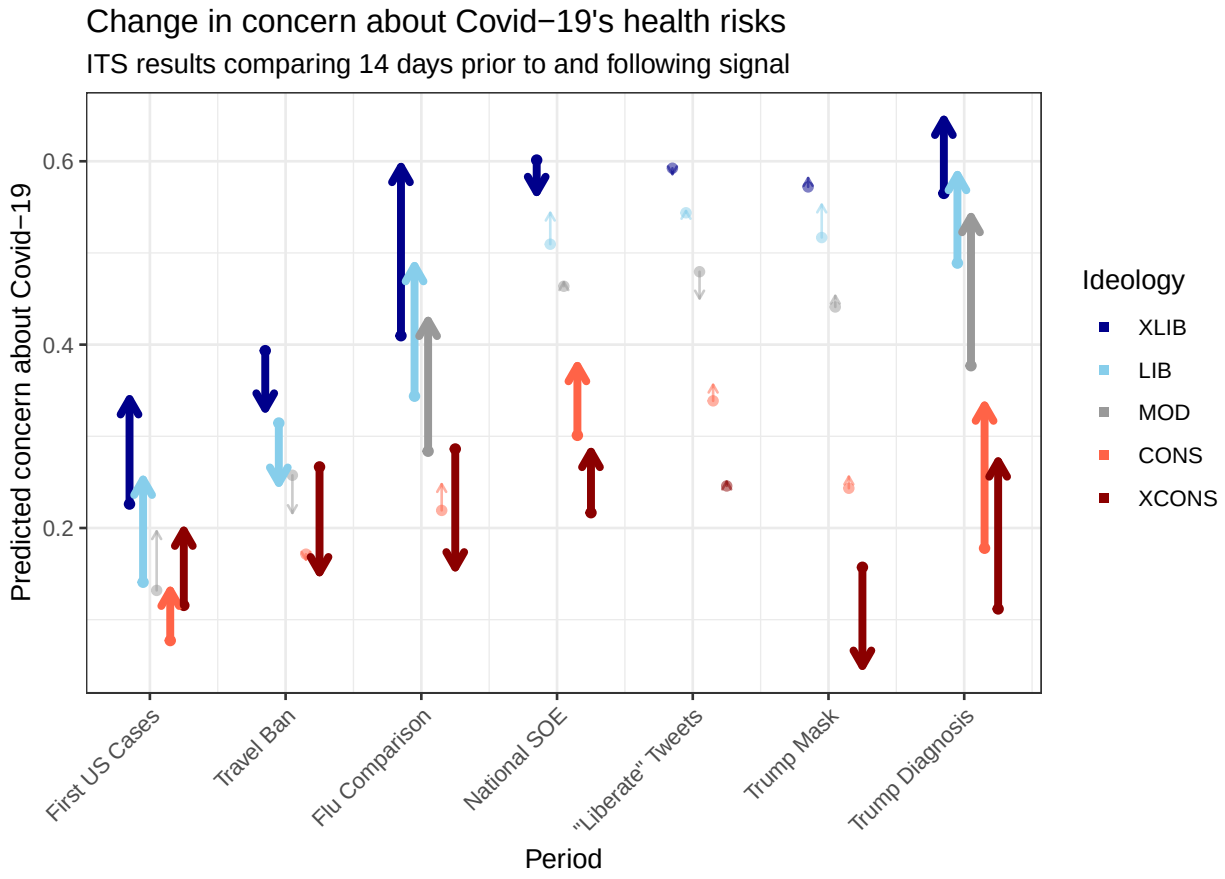

Figure 8: Predicted concern just prior to (circles), and just following (arrowheads), each signal (x-axis), broken out by ideological subgroup. Transparent points and flat arrowheads indicate statistically insignificant estimates of β_2 from Equation 1.

Less than a month later, we show that this ideological divergence was reversed with the national state of emergency (SOE) on March 13th, 2020. Here, concern among conservatives and extreme conservatives increased, while the concern among more liberal groups remained unchanged, even declining among extreme liberals. This increase in concern among conservatives helped compress the partisan gap in concern from approximately 40 percentage points down to roughly 25. Taken together, the evidence from the beginning of the Covid-19 outbreak in the United States suggests a powerful influence of Trump's cues on the severity of the virus, delaying the increasing concern among conservatives, and even reversing the

trajectory among the most conservative members of our random sample of Twitter users.

Turning to the remainder of the year, we find little evidence of shifts in concern that follow subsequent cues from Donald Trump. Neither his "liberate" tweets nor the first time he wore a mask corresponded to significant or substantive changes in concern, despite the abundant media attention each signal garnered at the time.[12] While the partisan gap had again widened by July back to the magnitude found in February with the seasonal flu comparison (roughly 50% separating extreme liberals from extreme conservatives), we do not find evidence that these trends are related to the high-salience cues we identified for Trump. And the decision to wear a mask in public in July had no relationship with the concern of any group with the exception of extreme conservatives, who grew substantially less concerned.[13]

However, Trump's diagnosis on October 2nd saw increasing concern among all subsets of the public. Increasing concern was most pronounced among more conservative Twitter users, but all ideological groups expressed more concern about Covid-19 following the diagnosis. Like the state of emergency before it, Trump's positive diagnosis managed to reduce the ideological gap by increasing concern among conservatives.

4.1 Backlash

Our core finding is that America's polarization on the issue of Covid-19 occurred early on in the year, with the bulk of the evidence accruing to Trump's first comparison of the virus to

[12] See SI for a detailed discussion of each of these events.

[13] If we assume that Trump's decision to wear a mask sent a signal that the virus was dangerous, one might interpret the decline in concern among extreme conservatives as evidence of backlash. However, a careful reading of the event reveals that Trump's behavior was paired with a framing that he was protecting wounded military and at-risk health care workers at Walter Reed Hospital, muddying the clarity of the signal between what the general public should be worried about.

the seasonal flu. Importantly, our results show that the public's divergence associated with this cue was produced by *both* liberals and conservatives moving in opposite directions. As illustrated in Figure 8 above, while extreme conservatives reduced their (expressed) concern for Covid-19's health risks following Trump's comparison, moderates, liberals, and extreme liberals all greatly increased their concern. On its face, this pattern is consistent with a backlash effect in which out-group members move in the opposite direction of a partisan signal.

We posit that there are two plausible explanations for such a reaction. The first is consistent with a rational actor story in which liberals anticipate that Trump's cue will reduce concern among half of the U.S. population, increasing the health risks as conservative individuals take less care in preventing the spread of the virus. The second explanation recognizes that individuals rarely are exposed to a politician's statements in a vacuum. Instead, liberals are likely to learn of Trump's signals downplaying the seriousness of the virus via co-partisan or co-ideological sources. These sources, in turn, are likely to frame Trump's statements in a negative light, and provide additional information about Covid's health risks to support this frame. In the Bayesian formalization presented above, this story rejects the possibility that both liberals and conservatives can receive the same signal $x \sim \mathcal{N}(\mu_x, \hat{\sigma}^2_{i,x})$. Instead, the ubiquity of partisan new sources means that the same objective fact (i.e., Trump made a claim) will be distributed according to different means $\mu_{j,x}$ depending on the news source j. For example, PBS News might report on Trump's flu comparison by mentioning that the health risks of the virus are significantly more dangerous than the seasonal flu, whereas Fox News might report on the same statement without any such countervailing information.

We use both a descriptive visualization of the data along with a revised diff-in-diff analysis, in which we now define the "treated group" to be liberal media whose coverage of

the pandemic responds in the opposite direction from Trump's claims.[14] Before presenting the results from the regression, we start with the visual description of the seriousness of tweets written by all news outlets over the course of 2020 who were followed by our random sample. As illustrated in Figure 9, there is descriptive evidence in line with our theory that Trump's statements are not issued in a vacuum. In particular, we note that the most liberal media were the *least* concerned prior to Trump's comparison of the virus to the common flu. However, within a day of this statement these outlets reversed course, and became the most outspoken proponents of the pandemic's health risks for the rest of the year. Importantly, this difference is substantially attenuated among the more ideologically extreme outlets. Similar dynamics are harder to see visually among ideologically moderate outlets, as illustrated in the middle facet of Figure 9. A diff-in-diff analysis (see Figure 10) reveals that this transition is statistically significant among both ideologically extreme and moderate outlets, although the evidence is substantially larger for the former.

Nevertheless, such a result is only suggestive evidence of how the backlash effect manifests. To more carefully test our claim, we run two tests. First, we test whether the information environment of our random sample diverged following Trump's comparison. To do so, we link each respondent in our random sample to the media accounts they followed at the end of 2019, prior to the notional "treatment" of Covid-19's onset in the United States. We are interested in whether our users were exposed to more or less media concern following Trump's comparison as a function of their own ideology. To evaluate this question, we regress the estimated concern about Covid-19 contained in media posts that our random sample of Twitter users follow using the interrupted time series specification above, except interacted

[14] Our specification is almost identical to that used in Equation 2, except that we compare how liberal media cues responded to Trump's comparison. Formally,

$$media_{g,t} = \alpha + \beta_{31} Lib + \beta_{32} \mathbb{I}(t > t_0) + \beta_{33} \mathbb{I}(t > t_0) * Lib + \varepsilon_{g,t} \quad (3)$$

.

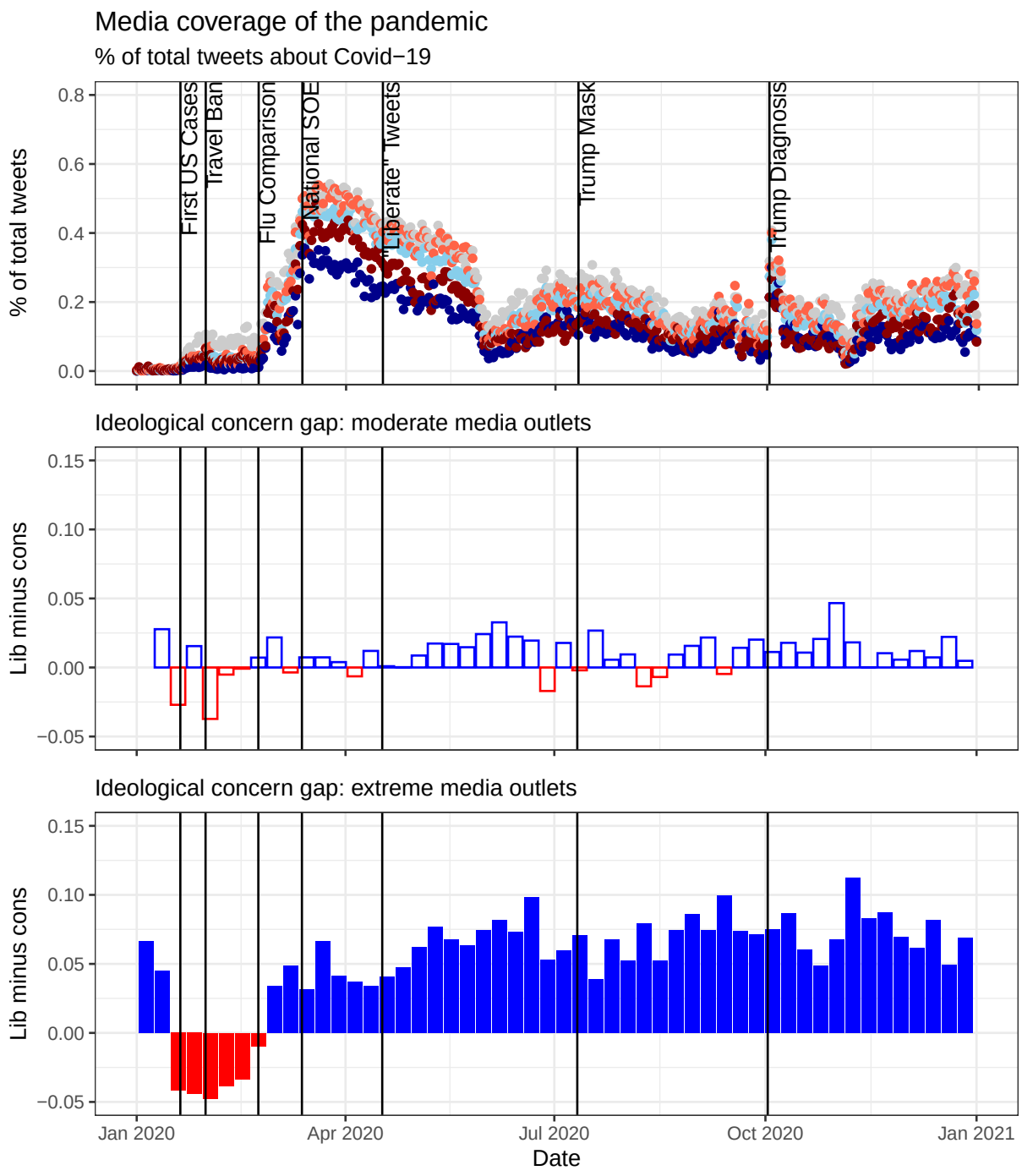

Figure 9: Media coverage of Covid-19 by proportion of total tweets written that contained Covid-19 keywords (top facet) and concern about the health risks expressed in tweets, visualized as the gap in concern between liberal and conservative outlets, broken out by moderate (middle facet) and extreme (bottom facet) ideologies.

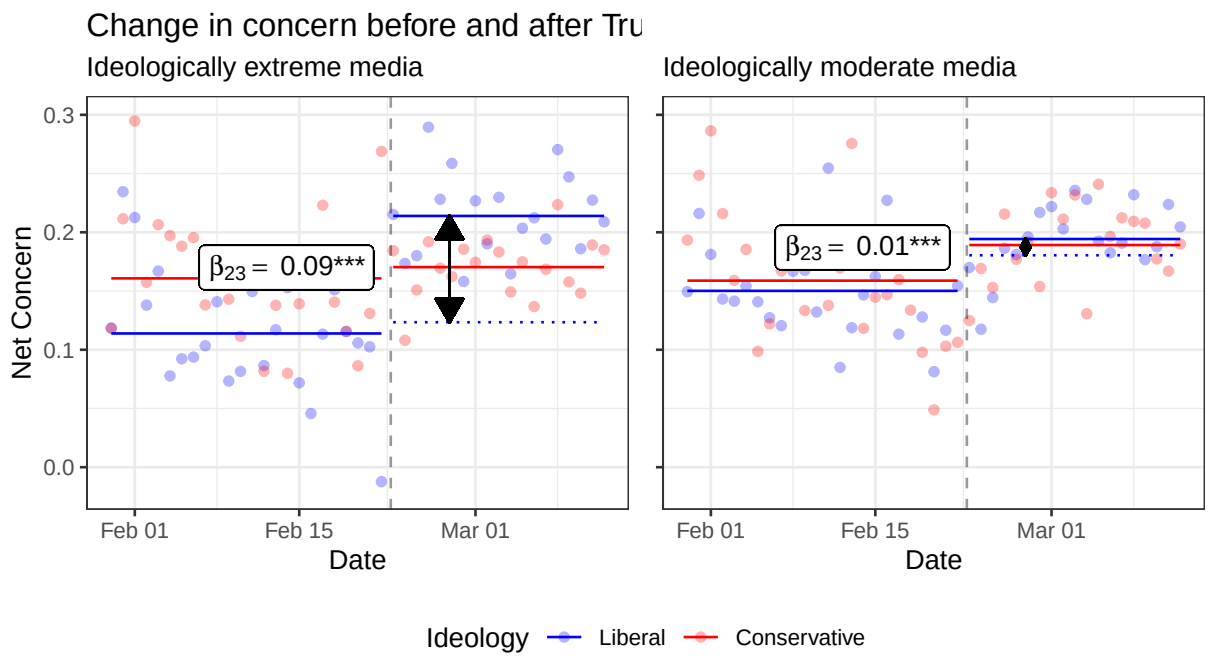

Figure 10: Diff-in-diff results comparing concern about Covid-19 expressed in liberal (blue) and conservative (red) media's tweets before and after Trump's comparison to the seasonal flu, modeled as $media_{g,t} = \alpha + \beta_{31} Lib + \beta_{32}\mathbb{I}(t > t_0) + \beta_{33}\mathbb{I}(t > t_0) * Lib + \varepsilon_{g,t}$. While conservative outlets did not significantly change their cues about Covid-19's health risks (β_{32}), liberal outlets went from being 5 percentage points less concerned than conservative outlets prior to Trump's cue (β_{31}), to 5 percentage points more concerned than conservative outlets following Trump's cue, yielding a diff-in-diff estimate of approximately 9 percentage points (β_{33}).

with the user's ideology.

$$
\begin{aligned}
media_{i,t} = \alpha_i &+ \beta_{41}\, T + \beta_{42}\mathbb{I}(t > t_0) + \beta_{43}\, T_{t>t_0} \\
&+ \sum_{j=4}^{8} \beta_{4j}\, ideo_{j-3,i} \\
&+ \sum_{j=9}^{13} \beta_{4j}\, T * ideo_{j-8,i} \\
&+ \sum_{j=14}^{18} \beta_{4j}\mathbb{I}(t > t_0) * ideo_{j-13,i} \\
&+ \sum_{j=19}^{23} \beta_{4j}\, T_{t>t_0} * ideo_{j-18,i} + \varepsilon
\end{aligned}
\tag{4}
$$

Note that the unit of observation in these data is the random Twitter user-by-day, allowing us to implement user fixed effects, denoted with α_i. As above, T is a variable counting the number of days as integers from the start of the period of analysis and $T_{t>t_0}$ is a second integer variable counting the number of days from the high-salience cue. To save space, we indicate a series of coefficients with a summation, whose indices capture both the coefficient of interest as well as 1 through 5 ideology dummies, corresponding to extreme liberals (1) through moderates (3) to extreme conservatives (5).[15]

We calculate the average level of concern contained in the daily posts of media outlets followed by our users by weighting the average concern expressed by these accounts by the proportion of total accounts that they follow. For example, consider a user who follows two extremely liberal accounts whose net concern measured on a given day is 0.6, and who also follows 8 liberal accounts whose net concern is 0.3. On this day, we would calculate this user's total exposure to concern about Covid-19 on this day as 0.2*0.6 + 0.8*0.3 = 0.36. We can interpret this value to mean that 36% of the tweets written by the media outlets they follow on Twitter expressed concern about Covid-19 or, alternatively, than the probability of reading a tweet expressing concern is 36 out of every 100 tweets on their timeline. Conversely, a different user who follows 5 conservative accounts with a net concern of 0.05, and 5 extremely conservative accounts with a net concern of -0.1 would have an aggregate net concern of -0.025, meaning that they are more likely to be exposed to tweets expressing skepticism about Covid-19's health risks than to tweets expressing concern.

As illustrated in Table 1, regardless of which measure we use, there is a linear and significant difference in the information environment our panel of randomly sampled accounts was exposed to following Trump's flu comparison. Specifically, the most liberal users were exposed to more concern in tweets written by the media accounts that they follow imme-

[15]Thus, for example, the notation $\sum_{j=4}^{8} \beta_{4j} ideo_{j-3,i}$ is shorthand for $\beta_{44} ideo_{1,i} + \beta_{45} ideo_{2,i} + \beta_{46} ideo_{3,i} + \beta_{47} ideo_{4,i} + \beta_{48} ideo_{5,i}$ where the numeric subscripts on $ideo_{j,i}$ indicate extreme liberals through extreme conservatives.

diately following Trump's comparison (roughly a 5 percentage point increase if calculating the aggregate concern using the proportion of total accounts followed). This increased exposure to concern was significantly higher than the increase experienced by liberal, moderate, conservative, or extreme conservative accounts. Importantly, however, we note that the regression models suggest that all media coverage grew more concerned following Trump's comparison – including extremely conservative outlets – in two out of the three models. Only when we define media coverage as cues from the co-ideological accounts followed by our users do we observe a net negative coefficient on extremely conservative outlets. We visualize the predicted probabilities from the ITSA results in Figure 11 to aid interpretation.

Table 1: Concern about Covid-19 in one's media environment

Model:	# Outlets Weights (1)	# Tweets Weights (2)	Congruent Outlets (3)
Variables			
Post	0.0545***	0.0423**	0.1480***
	(0.0165)	(0.0159)	(0.0281)
Post × Liberal	-0.0133***	-0.0090**	-0.1140***
	(0.0028)	(0.0034)	(0.0342)
Post × Moderate	-0.0245***	-0.0163***	-0.1272***
	(0.0051)	(0.0059)	(0.0341)
Post × Conservative	-0.0291***	-0.0192***	-0.0871**
	(0.0082)	(0.0066)	(0.0327)
Post × Extreme Cons	-0.0439***	-0.0332***	-0.1655***
	(0.0155)	(0.0121)	(0.0417)
Fixed-effects			
Account	Yes	Yes	Yes
Fit statistics			
Observations	3,677,905	3,677,905	2,616,076
R^2	0.41243	0.32138	0.25181
Within R^2	0.19989	0.10610	0.13096

Clustered (Account & date) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The preceding results provide suggestive evidence in support of our explanation for why the liberal public expressed *greater* concern about Covid-19 following Trump's flu comparison: their co-ideological media shifted the tone of their coverage of the virus to emphasize the health risks. However, does this shift in the media information environment explain the apparent backlash to Trump's comparison among the liberal users in our random sample?

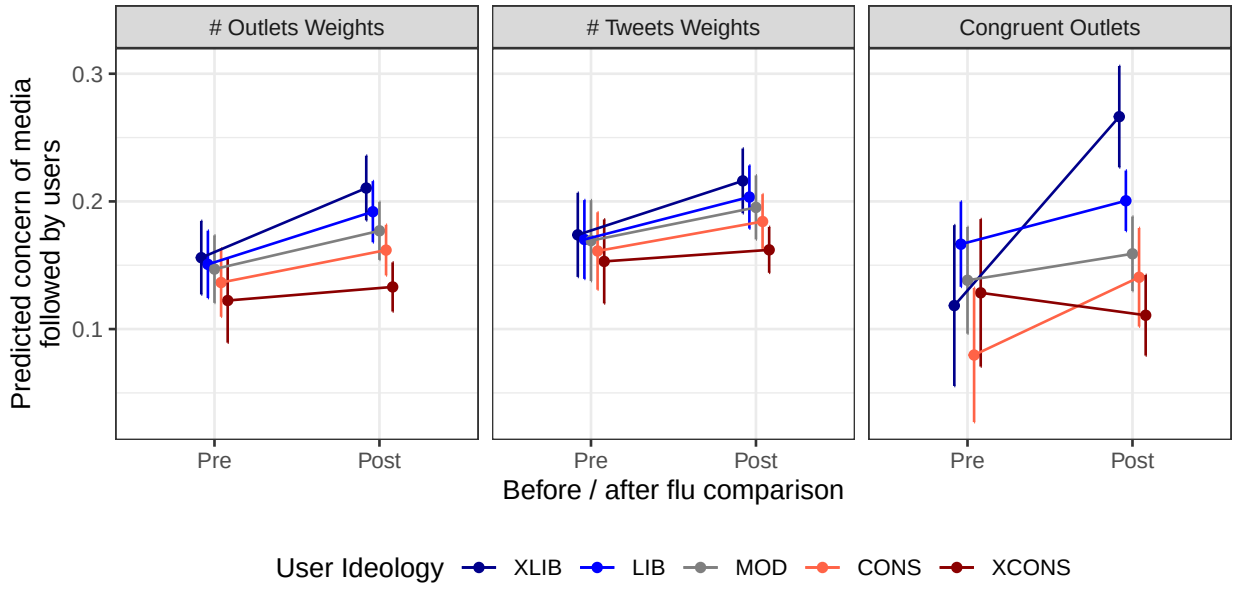

Figure 11: Predicted probabilities of concern (y-axes) expressed by media accounts followed by our random sample of Twitter users (user ideology indicated by colors) just before and just after Trump's comparison to the seasonal flu (x-axes). Left facet uses the number of outlets each user follows to calculate the net exposure to media concern, center facet uses the number of tweets, right facet uses only tweets written by ideologically congruent outlets.

Or in the language of causality, to what extent does the change in media coverage *mediate* the relationship between Trump's cue and the public's attitudes? To evaluate this question, we refer to the directed acyclic graph (DAG) visualized in Figure 12. Writ large, we know that the total "effect" of Trump's comparison produced a net increase in concern among all groups except extreme conservatives, based on the interrupted time series analyses described above in Equation 1. To evaluate the extent to which these relationship is mediated by the media coverage of Trump's cue, we unpack this total relationship into the association between the cue and media coverage (pathway a); the association between media coverage and public concern, holding the cue constant (pathway b); the association between the cue and public concern, holding the media coverage constant (pathway ADE, or the "average direct effect"); and the total association (pathway tot). Pathway a is estimated using Equation 5, while pathways b and c are estimated using Equation 6. The total mediated effect – also known as the "average causal mediation effect" or ACME – is the total relationship less

the ADE, the ratio of which (the "proportion mediated") captures the extent to which the mediation pathway explains the empirical phenomenon.

$$a: \quad media_{i,t} = \alpha_i + \beta_{51}\mathbb{I}(t > t_0) + \beta_{52}T + \beta_{53}T_{t>t_0} + \varepsilon_{i,t} \tag{5}$$

$$b, c: \quad concern_{i,t} = \alpha_i + \beta_{61}\mathbb{I}(t > t_0) + \beta_{62}media_{i,t} + \beta_{63}T + \beta_{64}T_{t>t_0} + \varepsilon_{i,t} \tag{6}$$

$$tot: \quad concern_{i,t} = \alpha_i + \beta_{71}\mathbb{I}(t > t_0) + \beta_{72}T + \beta_{73}T_{t>t_0} + \varepsilon_{i,t} \tag{7}$$

We add the coefficients of interest to the DAG in Figure 12, ignoring the ideologies of our users for the moment. Writ large, we confirm that there is a statistically significant positive association between Trump's cue and public concern (pathway *tot*), which we disaggregate into the positive association between Trump's cue and the average news media concern (pathway a), a statistically significant positive association between media concern and public concern (pathway b), and a statistically significant positive association between Trump's cue and public concern (pathway ADE). The proportion of the total relationship that is mediated by the media coverage is 12.1%. For inference, we rely on the `mediate` package for `R` developed by Tingley et al. (2014), which allows us to cluster the quasi-Bayesian standard errors by individual.

However, we suspect that these results might vary by the ideology of the users, as well as by whether they are measured prior to, or following Trump's cue. As such, we re-run the preceding analysis, but subset the data to calculate the mediation results by each ideological group of users separately, and adjust Equation 6 to include an interaction between the flu comparison indicator and the media coverage measure. We visualize the results in Figure 13 by plotting the ADE (white bars), ACME (gray bars), and total relationships (combined) together. Two meaningful conclusions present themselves from this analysis. First, there is systematically stronger evidence of a mediated pathway being active among more liberal

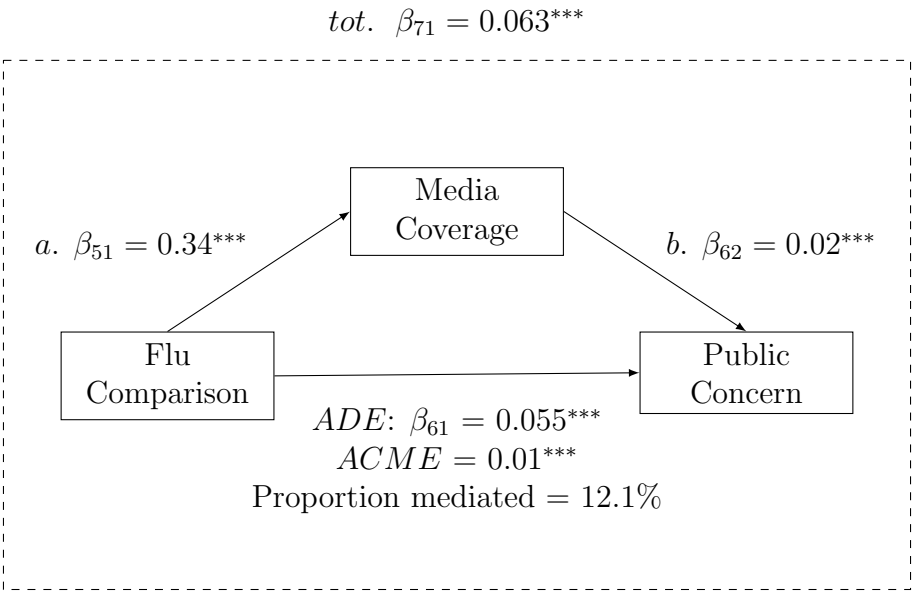

Figure 12: Directed acyclic graph (DAG) illustrating relationship between Trump's signal ("Flu Comparison") and concern about Covid-19 expressed by media outlets ("Media Coverage"); relationship between media concern and public concern ("Public Concern"); and direct relationship between Trump's signal and public concern.

users, consistent with our expectations. Second, Trump's cue is what activates the mediation pathway, as illustrated by the very weak and statistically insignificant associations prior to his flu comparison, compared to the substantially larger associations following.

5 Discussion

In this paper, we have used rich data to characterize the response of different actors in American democracy to Covid-19 over the course of 2020. Our daily tweet-level data, combined with a powerful machine learning classifier trained on over 42,000 human labeled tweets, allows us to describe the evolution of this public health crisis with unprecedented detail. We show that former President Donald Trump was a strong opinion leader of conservatives, particularly extreme conservatives. His influence on the degree to which the public took Covid-19 seriously was especially pronounced in February when he compared the virus to

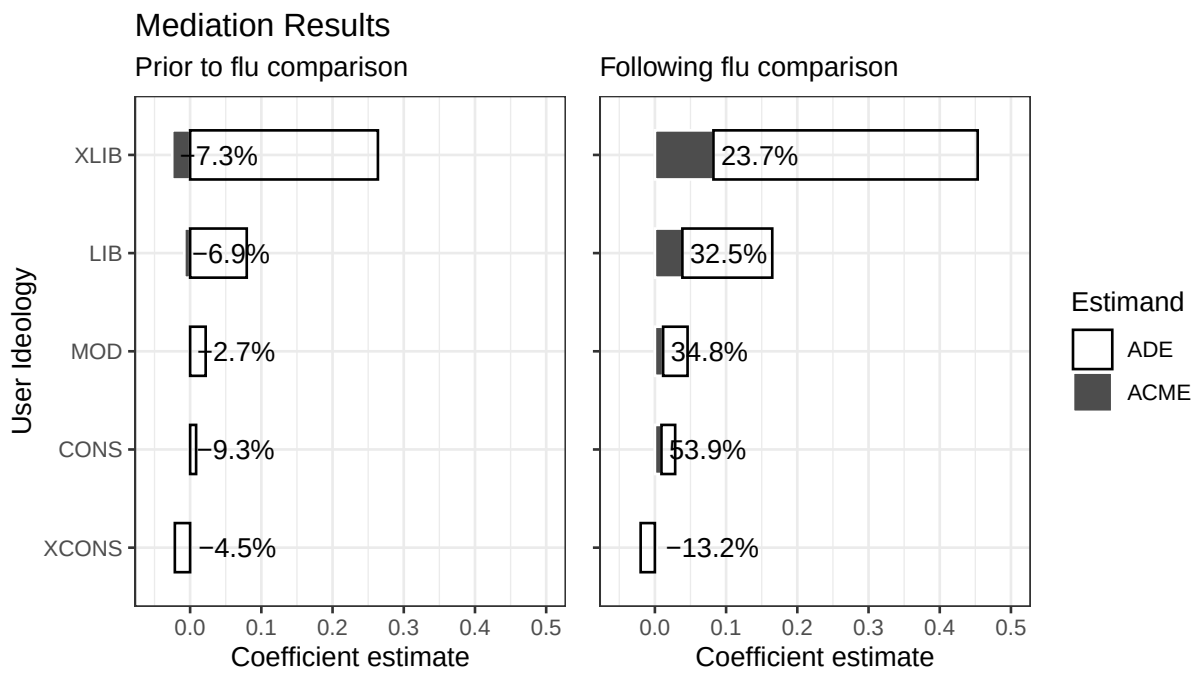

Figure 13: Descriptive mediation results. Average direct effect (x-axes) in white bars and average "causal" mediation effect in dark gray bars, broken out by ideology of users (y-axes). The two bars are stacked, meaning that their sum is the total effect, the proportion of which is mediated by the media's reaction to Trump's flu comparison is given in percentages on the plots.

the flu. This cue dramatically separated liberals from conservatives in terms of their perceptions of Covid-19's health risks. While his statement reduced concern among the most extreme conservatives in our sample, Trump also appears to have influenced liberals in the opposite direction. In particular, we document evidence of a seeming backlash to his statements comparing Covid-19 to the seasonal flu, and suggest that this may reflect the reality that the American public rarely confronts Trump's statements in a vacuum. Instead, the liberal backlash is stronger among those who follow more liberal accounts, which themselves grew more concerned with Covid-19's health risks in response to Trump's attempts to downplay them. We argue that this finding complicates the scholarly consensus that a "backlash" is rare. While this conclusion holds in survey experiments where the researcher has tight control over the information environment of their subjects, in the "real world" of Twitter – or the online information environment more broadly – individuals rarely, if ever, encounter

a single elite cue in isolation. As we demonstrate, Trump's comparison between Covid-19 and the seasonal flu was immediately broadcast by media and other elites, whose coverage was naturally framed according to their own ideological disposition.

However, across several other high-salience events, evidence of similar polarization is lacking. Objective facts (such as the first reported cases in the United States) as well as non-partisan signals (the national State of Emergency) moved the public in concert, with the SoE doing much to mitigate the divergence generated by Trump's flu comparison only three weeks earlier. Importantly, subsequent cues from Trump (the "liberate" tweets in April and the first time he wore a mask) produced negligible differences in the public's perception of Covid's health risks. Even among extreme conservatives for whom these cues produced statistically significant shifts in concern in the regression discontinuity model, the magnitude of the change was only a fraction of the influence of Trump's flu comparison.

These patterns may reflect the gradual ossification of public perceptions. The Bayesian model of belief formation assumes that today's posterior becomes tomorrow's prior. Since each posterior must be (weakly) less variable than the prior that preceded it, it must therefore be the case that opinions grow increasingly difficult to change over time.

The one exception to the pattern of diminishing marginal influence is Trump's Covid diagnosis in the beginning of October. This event corresponded to a uniform shift among all members of the public in the direction of greater concern. At face value, this would seem to indicate that this event was both extraordinarily high salience and also non-partisan. However, more detailed analysis of the minute-to-minute changes in sentiment suggest that there were nuances even within this episode (see SI Section XX). Specifically, while the initial report of the positive test was met with systematic increases in concern among all members of the public, subsequent events – most notably the announcement of Trump's return to the White House – undid much of the bipartisan concern.

These conclusions should be caveated with the knowledge that our analysis relies on the opinions expressed by Twitter users on that platform. While we argue that this source is insulated from some of the challenges associated with survey self-reports (i.e., partisan motivated responding), we also acknowledge that the positions taken online (particularly where users have anonymous accounts) might capture a complicated combination of true beliefs and self-presentation. Future work that marries rich datasources such as what we use in this paper with objective measures of behavior could help identify the degree to which our findings map onto real world behavior. Nevertheless, we argue that our results provide an important contribution to a growing consensus on the American response to Covid-19 by highlighting the degree to which Donald Trump influenced public perceptions, particularly in the earliest days of the pandemic.

References

Achen, Christopher H. 1992. "Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research." *Political behavior* 14:195–211.

Achen, Christopher, Larry Bartels, Christopher H Achen and Larry M Bartels. 2017. *Democracy for realists*. Princeton University Press.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political analysis* 23(1):76–91.

Bartels, Larry M. 1993. "Messages received: The political impact of media exposure." *American political science review* 87(2):267–285.

Bullock, John G. 2009. "Partisan bias and the Bayesian ideal in the study of public opinion." *The Journal of Politics* 71(3):1109–1124.

Coppock, Alexander. 2023. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

Druckman, James N and Mary C McGrath. 2019. "The evidence for motivated reasoning in climate change preference formation." *Nature Climate Change* 9(2):111–119.

Eady, Gregory, Richard Bonneau, Joshua A Tucker and Jonathan Nagler. 2020. "News sharing on social media: Mapping the ideology of news media content, citizens, and politicians.".

Fowler, Anthony and Andrew B Hall. 2018. "Do shark attacks influence presidential elections? Reassessing a prominent finding on voter competence." *The Journal of Politics* 80(4):1423–1437.

Freeder, Sean, Gabriel S Lenz and Shad Turney. 2019. "The importance of knowing "what goes with what": Reinterpreting the evidence on policy attitude stability." *The Journal of Politics* 81(1):274–290.

Guess, Andrew and Alexander Coppock. 2020. "Does counter-attitudinal information cause backlash? Results from three large survey experiments." *British Journal of Political Science* 50(4):1497–1515.

Kahan, Dan M, Donald Braman, John Gastil, Paul Slovic and CK Mertz. 2007. "Culture and identity-protective cognition: Explaining the white-male effect in risk perception." *Journal of Empirical Legal Studies* 4(3):465–505.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108(3):480.

Lenz, Gabriel S. 2013. *Follow the leader?: how voters respond to politicians' policies and performance*. University of Chicago Press.

List, Christian, Robert C Luskin, James S Fishkin and Iain McLean. 2013. "Deliberation, single-peakedness, and the possibility of meaningful democracy: evidence from deliberative polls." *The journal of politics* 75(1):80–95.

Lodge, Milton and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

Porter, Ethan and Thomas J Wood. 2022. "Political misinformation and factual corrections on the Facebook news feed: Experimental evidence." *The Journal of Politics* 84(3):1812–1817.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele and Kosuke Imai. 2014. "Mediation: R package for causal mediation analysis.".

Wood, Thomas and Ethan Porter. 2019. "The elusive backfire effect: Mass attitudes' steadfast factual adherence." *Political Behavior* 41:135–163.

Wu, Patrick Y, Joshua A Tucker, Jonathan Nagler and Solomon Messing. 2023. "Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting." *arXiv preprint arXiv:2303.12057* .

Zaller, John R. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Zechman, Martin J. 1979. "Dynamic models of the voter's decision calculus: Incorporating retrospective considerations into rational-choice models of individual voting behavior." *Public Choice* 34(3-4):297–315.