(Mis)representing Ideology on Twitter: How Social Influence Shapes Online Political Expression*

William Small Schulz†[1], Andrew M. Guess‡[1], Pablo Barberá[2], Simon Munzert[3], JungHwan Yang[4], Alexander Gottlieb[5], Adam Hughes[6], Emma Remy[6], Sono Shah[6], and Aaron Smith[6]

[1]Department of Politics, Princeton University
[2]Department of Political Science and International Relations, University of Southern California
[3]The Hertie School
[4]Department of Communication, University of Illinois at Urbana-Champaign
[5]Department of Geography, Dartmouth College
[6]Pew Research Center

6 September 2020

Abstract

Expressing opinions on social media has become a standard form of participation in the political process, but we know little about the factors that shape it. In this paper, we investigate the role of social context. Decades after the development of the canonical "Spiral of Silence" model, the public sphere has radically shifted toward a networked space mediated by social platforms. We articulate a theory of conformity in social media expression and test it by analyzing unique datasets linking U.S. survey respondents to their public Twitter accounts. To measure political expression, we develop and validate a supervised classifier of tweet-level ideology and apply it to respondents' tweets and the tweets of people they follow. We find that the ideology of Twitter followees' tweets is predictive of respondents' own expressed ideology on Twitter, even after holding constant self-reported ideological predispositions. Our findings offer the first real-world evidence of social conformity effects on social media and demonstrate a powerful methodological approach for studying these dynamics.

*This research was generously funded by a grant from the Volkswagen Foundation Computational Social Science Initiative, reference 92 143. This study was approved by the Princeton University Institutional Review Board (#8327, 10041). We thank Matt Deverna for outstanding research assistance.

†Email: wschulz@princeton.edu

‡Email: aguess@princeton.edu

The individual expression of opinions and attitudes is not strictly an individual affair.

Consider, for example, the consequences of believing that witches do not exist, in the context of a witch-hunting Puritan community.

Katz and Lazarsfeld, *Personal Influence* (1955)

1 Introduction

Social media is a window into people's lives, a place where individuals spontaneously share their thoughts and attachments. As such, it is a space for civically meaningful expression that collectively defines a landscape of perspectives. This development has not evaded politicians and journalists, who use these expressions to construct an understanding of public opinion, however distorted or incomplete (McGregor 2019, 2020).

Citizens' ability to speak their minds publicly, with potentially great reach, is a testament to the empowering potential of social media (Tufekci 2017; Barberá et al. 2019). At the same time, however, growing concerns from the public, policymakers, and academic researchers point to the possibility that the very features enabling this communication might also fuel counterproductive social dynamics. These include the spread of rumors and false news (Friggeri et al. 2014; Vosoughi, Roy and Aral 2018), emotional contagion (Brady et al. 2017), and polarization (Bail et al. 2018).

What these and other pathologies of social media have in common is a pattern in which individuals' online activity is at least partially a function of *context*. Thus for online expression, what social media users say (or don't say), and how they say it, may depend not only on their personal convictions, but also on what *others* in their online social networks are saying. This observation has major implications for how we understand the determinants of political expression, online and offline (e.g., Katz and Lazarsfeld 1955). Further, such social influence dynamics, if confirmed, would call into question a naïve interpretation of social media as an unfiltered record of people's opinions and beliefs, and cast doubt upon its use as a stand-in for survey data (e.g., Barberá 2016; Beauchamp 2017; Posegga and Jungherr 2019; Nguyen et al. 2020).

Influential theories specify how fear of social sanction may cause people to censor their own opinions. Noelle-Neumann (1974) famously proposed a self-reinforcing "Spiral of Silence" in which opinion minorities refrain from expressing their views, creating an illusion of conformity. A complementary mechanism, preference falsification, posits that people actively endorse views that they do not privately hold when they perceive themselves to be at odds with the majority (Kuran 1995). Hampered by a lack of data, attempts to test these theories have, to date, relied on hypothetical survey self-reports (e.g., Fox and Holt 2018), laboratory simulations (e.g., Carlson and Settle 2016; Levitan and Verhulst 2016), and formal models (e.g., Bicchieri 2005). By focusing on social media, we are for the first time able to test real-world social influence dynamics in a naturalistic setting.

Indeed, while they were developed to explain the pre-digital world, these frameworks parallel contemporary debates about self-censorship and opinion cascades on social media. These debates reflect genuine concern: In a 2016 Pew Research Center survey, for instance, 64% of respondents said they think users "refrain from speaking their minds about political issues on social media out of concern that they will lose friends or get criticized" (Duggan and Smith 2016). Users' own admissions of online self-censorship have been documented across a variety of surveys conducted over the past decade (Hampton et al. 2014; Kim 2016; Hoffmann and Lutz 2017; Neubaum and Krämer 2018), especially amongst those who fear negative social feedback (Vraga et al. 2015; Chan 2018; Chen 2018; Fox and Holt 2018; Bäck et al. 2019).

Taken at face value, such logic leads to a puzzle: increasing self-censorship implies conformity of expressed opinion, which contradicts clear patterns of polarization, both on social media and in American society more broadly (Settle 2018). To resolve this seeming contradiction, we propose a theoretical account that emphasizes the different "imagined audiences" that people hold in their heads when deciding what to say online (Marwick and boyd 2011). We conceptualize these audiences as comprising members of a social group, focusing specifically on the ideological subgroups that increasingly structure Americans' experience of mass politics (Mason 2018).

We then leverage new data and computational methods to study the extent to which the ideological content of people's political expression on Twitter is determined by their own ideological self-placement (as measured in surveys) versus the ideological content of the tweets of people they follow. To estimate the ideology of both subjects' expressed tweet ideology and that of their Twitter followees, we develop a supervised learning model trained on approximately 10,000 hand-labeled tweets and apply it to over 40 million tweets collected from the accounts of respondents to two nationally representative surveys and the accounts they follow.

We find that, even after controlling for "private" (survey self-reported) ideological affiliation, the slant of individuals' online political expression is significantly influenced by the slant of those they follow. While we cannot completely rule out homophily as a potential factor, our findings offer an important empirical regularity with implications for our understanding of public opinion, online political expression, and social influence more broadly.

This paper proceeds as follows. In the next section, we briefly review the literature that provides the building blocks of our argument. We then outline our theory and predictions. Next, we introduce the data used in the study: a panel survey conducted by YouGov; a KnowledgePanel survey collected by the Pew Research Center; Twitter accounts linked to respondents of both surveys; and millions of scraped tweets from both those respondents and the accounts they follow. We then outline our supervised tweet ideology classifier and how we aggregate its predictions to individual-level measures of tweet (expressed) ideology and followee ideology. Finally, we present regression results testing our main predictions and conclude with a discussion.

2 The "Imagined Audience" on Social Media

Studying social influence over political expression on Twitter requires a clear understanding of how Twitter users experience the platform as a social environment. Since public tweets are potentially visible to anybody, traditional notions of "audience" do not directly translate to Twitter. Although one's "followers" constitute the most literal definition of a bounded Twitter audience, users are not automatically aware of their followers' views because of the directed nature of the platform's network ties: users must *follow back* their followers if they want to receive their tweets. As a result, while followers may constitute a literal audience, they are not necessarily a *salient* one.

Given an arbitrarily large potential audience and a hazy view of direct followers, Marwick and boyd (2011) find that Twitter users resort to a variety of strategies for constructing their "imagined audience," often basing their assumptions about this audience on the *accounts they follow*: their followees.[1] We therefore consider followees as the most salient reference point for users' "imagined audience" and in subsequent analyses interpret followees as the proximate source of influence over users' tweeted political expression. Followed accounts are also role models in the sense that they often express perspectives that following users agree with: see, e.g., Eady et al. (2019) for evidence of homophily in following patterns of political elites (although less so for the accounts of media outlets).

When, as is often the case, followed accounts have more followers and receive more engagement than the user who follows them, they can implicitly model the forms of expression that a following user would want to imitate in order to receive more likes and follows for their own accounts. In a recent experiment demonstrating this logic, Yang, Qureshi and Zaman (2020) find that when followed by a like-minded bot, Twitter users who followed back — thus placing the bot squarely within their hypothesized "imagined audience" — gradually appeared to change their expressed views related to immigration in the direction of the bot's own gradually shifting positions on the topic.

Next we describe how ideological groups serve as particularly salient imagined audiences on Twitter for individuals expressing themselves politically.

3 Identity, Ideology, and Expression on Social Media

We build on prior work, extending the conception of partisanship as an *expressive* identity to ideological groups (e.g., Huddy, Mason and Aarøe 2015). It is now commonly accepted that partisanship acts as an identity, not only in the sense of group attachment (Campbell et al. 1960) but also as a mode of self-categorization determined by self-perceived match between self-image and party image (Green, Palmquist and Schickler 2002). It is well-established that partisans desire

[1] As one of that study's participants said, "I think I write to the people I follow and have twittered something recently" (Marwick and boyd 2011, p. 118).

consistency with their party (e.g., Groenendyk 2013) and adopt the positions espoused by their party on most topics, largely irrespective of whatever their own prior views might have been (e.g. Lenz 2012).

Ideology, by contrast, has traditionally been interpreted as a core value system, reflecting fundamental predispositions, possibly constrained by a knowledge of "what goes with what" (Converse 2006; Luskin 1990; Haidt, Graham and Joseph 2009; Jost, Federico and Napier 2013). The concept of ideological identity, however, is not new. Conover and Feldman (1981) conceptualized "liberal" and "conservative" as labels with symbolic content independent of specific issue positions. Levitin and Miller (1979) found that the mass public was able to apply ideological labels to parties and candidates, despite their established fuzziness regarding their issue content (Converse 2006). Recent scholarship has extended this tradition, finding that identity-based ideology is far more predictive of affective polarization than its issue-based counterpart (Mason 2018), somewhat resolving the puzzle of how an ideologically innocent public could become so polarized (Iyengar and Westwood 2015).

Recognizing how ideology, too, can function as an identity raises questions about the origins of its issue content. Accumulating evidence suggests that identity precedes content: Malka and Lelkes (2010) define ideological identity as a "readiness to adopt beliefs and attitudes about newly politicized issues that one is told are consistent with the socially-prescribed meaning of liberal-conservatism," and report evidence that this identity motivates cue-taking independent of an individual's substantive political values (see also Groenendyk, Kimbrough and Pickup 2020). This mirrors recent research finding that partisan cue-taking is most prevalent among people with strong partisan identities (Bakker, Lelkes and Malka 2020). This motivates an expressive view of ideology (e.g., Kahan 2013) in which people follow liberal and conservative norms because they desire consistency with their in-group, and not necessarily because it is a low-effort heuristic (e.g. Lupia and McCubbins 1998).

Far beyond passively *following* cues, moreover, there is reason to think that engaging in expression consistent with an identity can itself be constitutive of that identity (cf. Butler 1999). The idea that interpersonal expression plays an important role in defining political views can be traced at least as far back as the Columbia scholars, who studied it in great depth even as they lamented the difficulty of collecting the "systematic inventory" of interpersonal communication necessary for such research (Lazarsfeld, Berelson and Gaudet 1948, p. 13). More recently, Cramer Walsh (2004) offers a detailed qualitative study of how small-group discussions produce the contents of political identity: "Casual exchanges allow people to collectively give meaning to their social identities" (p. 42); "information conveyed in the group context [is] used to update or clarify one's sense of self" provided that "the other participants are people like oneself and thus ... knowledgeable sources of information for how 'one of us' ought to view the world" (pp. 47-48). In other words, political expression in groups is central to the development, maintenance, revision, and sharpening of political identities with attitudinal content. In this view, the in-group naturally serves as the primary point of reference for this process.

Historically, the construction of ideology has been undertaken by elites. Noel (2012) argues that ideological norms have been constructed through expression by public intellectuals (or "coalition merchants") who, through their writings, arrive at new ways of packaging issues together. Today, social media has granted broadcasting power to any citizen who chooses to open an account. It has put the mass public in the position of editorializing on the day's issues, castigating and praising public officials, and defining in their own terms what is right and what is wrong. Although no individual now holds the power to create and disseminate ideological norms, the aggregate public does so on a daily basis.

The advent of social media also creates new opportunities for researchers to study public expression, documented more comprehensively than the Columbia scholars could have ever dreamed. Yet to our knowledge, this is rarely how social media data are used in contemporary studies: Only a handful of studies (e.g., King, Schneer and White 2017; Munger 2017*a*) have considered users' own original expression as an important outcome in its own right. Where expression is used, it is frequently interpreted as a direct expression of attitudes (e.g., Barberá 2016; Beauchamp 2017; Posegga and Jungherr 2019; Nguyen et al. 2020; Yang, Qureshi and Zaman 2020) rather than a distinct phenomenon.

4 Private Attitudes versus Public Expression on Social Media

Indeed there is good reason to believe that public expression does *not* mirror private attitudes, and that the disjuncture between the two has important consequences. Noelle-Neumann's (1974; 1991) concept of "the Spiral of Silence" posits that people have a "quasi-statistical sense" (1991, p. 256) of the distribution of public opinion in their society, and that people who perceive that they hold a minority view on a particular topic tend to avoid expressing that view to others, out of a fear of social sanction. Noelle-Neumann argues that when people self-censor a minority view, it makes that view less visible to others, who therefore perceive it as an even smaller minority than it actually is. Other people who hold that view, then, become even more hesitant to express it publicly, which becomes a self-reinforcing cycle that, over time, makes minority opinions seem to vanish.

Kuran (1995) considers the Spiral of Silence an inadequate description of minority opinion-holders' behavior: "In actual contexts people reluctant to publicize their disenchantments do not just slip into silence. ... To make their efforts at preference falsification convincing, they tend to take steps to affirm their support for the status quo" (p. 113). Kuran's theory of "preference falsification" supposes that people derive three kinds of utility from expressing a political preference in public: the intrinsic utility one would experience from the enactment of an expressed policy preference and gaining material benefits, the reputational utility of expressing that preference, and the expressive utility or sense of integrity one experiences from expressing a preference that is close to one's true preference.

Since expressing a preference has a vanishingly small impact on the probability of that preference being enacted as policy, intrinsic utility has little influence over one's public expression of views, which is instead governed mostly by reputational and expressive utility. So, to the extent they can stomach misrepresenting themselves, individuals find it prudent to publicly align themselves with what they perceive to be the prevailing norms, creating a "spiral of prudence" (Kuran 1995, p. 113) which leads to pluralistic ignorance — a false appearance of consensus that may obscure a large number of dissenters, afraid to speak out. As Bicchieri (2005) argues, pluralistic ignorance occurs most often when individuals engage in social comparison with others whose public signals are clear but whose private motivations are obscure — conditions which characterize social media well.

There is accumulating evidence that these dynamics play out in some form on social media. First, there is survey evidence in favor of an online "quasi-statistical sense." For instance, Hampton et al. (2014) find that social media users' willingness to post their views about the Edward Snowden leaks depended on the extent to which they thought their followers agreed with them. Fox and Holt (2018) find that U.S. Facebook users' willingness to post their views on racial discrimination by police was powerfully inhibited by self-reported fear of social judgment and loss of friendships.

Alongside this evidence, a number of experiments have attempted to manipulate participants' perceptions of online opinion environments. Those that have attempted to do so using textual vignettes (e.g., Gearhart and Zhang 2014, 2018) have generally failed to find an effect. However, manipulations using mock-ups find that perceptions matter: Several studies find a relationship between the text of comments on mocked-up social media posts with participants' perceptions of others' opinions or willingness to express their own views (Neubaum and Krämer 2017; Wu and Atkin 2018; Woong Yun and Park 2011).

These findings suggest that social conformity may exert profound influence over what users choose to say (and not say) on social media, yet to our knowledge, no studies have sought to measure and describe this phenomenon in a real social media environment.

5 Hypotheses

We conduct an analysis of social influence over political expression on Twitter, using a dataset that includes two attitudinal surveys and linked digital trace data from respondents' public tweets as well as a sample of tweets from the accounts they follow (Munzert et al. Forthcoming). To measure the ideological character of these tweets, we develop a supervised learning classifier trained on hand-coded tweets and use this classifier to predict the ideological content embedded in the tweets of survey respondents and the users they follow. We then model respondents' *tweet ideology* as a function of both their private attitudes and their followees' tweet ideology.

Our expectations are as follows. First, we expect to establish that users' public political expres-

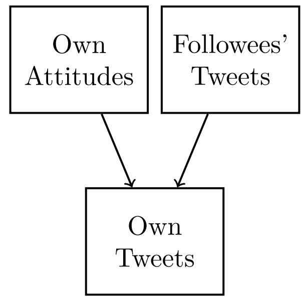


Figure 1: DAG codifying expected relationships.

sion on Twitter is related to their private political predispositions as measured in surveys. This will establish face validity for our tweet ideology classifier and illustrate the extent to which traditional predictors explain variation in online political behavior.

Hypothesis 1 (H1): *More liberal (conservative) individuals tweet more liberally (conservatively), as measured by our classifier.*

Next, we turn to the role of social context. A premise of our theoretical framework is that self-categorization and public expression are distinct manifestations of ideology, both conceptually and empirically. We expect that self-described ideology does *not* fully explain variation in public expression, and that such expression is powerfully shaped by social influence. We expect that on Twitter, the "imagined audience" sets the terms (literally) of political identity expression, and that public expression by this social reference group — which we operationalize as followees' tweets — therefore explains a substantively and statistically significant amount of variation in users' own public expression, independent of their ideological self-categorization as reported on a survey.

Hypothesis 2 (H2): *Individuals whose followees on Twitter tweet more liberally (conservatively) are more likely to tweet liberally (conservatively) themselves, holding constant their own private ideological predispositions.*

Figure 1 summarizes these expected relationships in a directed acyclic graph (DAG). Although this analysis cannot conclusively rule out homophily as a confounder of the relationship between followees' expression and individuals' own expression, we expect to observe that users' public political expression exhibits substantial variation that is not explained by their ideological self-identification.

Finally, we pose a research question. Scholars continue to debate the existence of ideological asymmetries in tendencies toward group conformity (Frimer, Gaucher and Schaefer 2014; Jost et al. 2018; Goldberg et al. 2020). Though the evidence to date is not definitive, we explore whether differences between ideological subgroups can be found in our hypothesized relationships.

Research Question 1 (RQ1): *Are self-described liberals whose followees tweet more liberally likely to tweet more liberally themselves?*

Research Question 2 (RQ2): *Are self-described conservatives whose followees tweet more conservatively likely to tweet more conservatively themselves?*

6 Data and Method

The analysis draws on two datasets: one based on a YouGov panel survey mostly conducted in 2018 (see Munzert et al. Forthcoming), and another based on a 2018 Pew Research Center survey fielded with Ipsos's KnowledgePanel (Wojcik and Hughes 2019). Both surveys collected Twitter account information from respondents who consented to share their public Twitter data, and their tweets (and followees' tweets) were collected for analysis.[2] This follows established practices for linking survey and digital trace data (e.g., Stier et al. 2019; Eady et al. 2019; Osmundsen et al. 2020). One virtue of this approach is that respondents are directly asked for their consent to participate; we make no effort to identify the social media accounts of individuals who do not provide a username. This provides an opportunity for individuals to opt out of the study; evidence suggests that consent is an important consideration for at least some social media users (Fiesler and Proferes 2018). By providing respondents the ability to opt out and by maintaining their anonymity while analyzing data, this study follows best practices in social media data analysis (Williams et al. 2017).

6.1 YouGov Sample

The YouGov dataset combines 7 waves of surveys (covering April 2018–April 2019) with public tweets of respondents (who reported using Twitter and linked their accounts) and their followees.

The subset used in this analysis begins with 328 respondents with valid Twitter account information for scraping. Between October 2018 and December 2019, approximately 1.1 million tweets were collected from respondents' accounts and approximately 22 million tweets were collected from a sample of accounts they followed.[3] Respondents were only included if the dataset contained at least 10 tweets from their account and 10 tweets from accounts they followed. This leaves $N = 287$ users in the final sample.

Appendix A includes detailed plots of the distribution of key summary variables in this sample. The users are predominantly White, with slightly more females than males, and a median age of 55. Users hold a range of educational statuses and family incomes and identify predominantly as Protestant, Roman Catholic, or religiously unaffiliated. The modal employment status is full-time employment, but the sample also includes a substantial number of retirees. Politically, the sample skews left, with substantially more "Very liberal" than "Very conservative" identifiers and more

[2] In the YouGov survey, respondents were asked to authenticate a Twitter app that would automatically collect their profile information. In the Pew survey, respondents could manually provide their Twitter usernames.

[3] First, a random sample of respondents' followees was selected for scraping; then, since the volume of this scraping task proved intractable, followees with 500,000 or more followers were selected for scraping on the basis that high-follower accounts would be especially influential and their tweets more likely to be seen by respondents.

"Strong Democrats" than "Strong Republicans." The majority of users describe themselves as "very interested" in politics.

6.2 KnowledgePanel Sample

The KnowledgePanel dataset consists of a single survey conducted between Nov. 21–Dec. 17, 2018.[4] The survey invitation was sent to 7,850 likely Twitter users (based upon the vendor's previous data collection), of whom 4,829 responded. Of that set, 3,649 confirmed that they used Twitter and 3,293 agreed to provide a Twitter handle.

Before collecting tweets, we evaluated the apparent validity of handles by comparing Twitter profiles with demographic information provided in the survey. All accounts belonging to organizations, institutions, international entities, and public figures are excluded from the analysis. Overall, $N = 2,791$ respondents (76% of those who provided any handle) provided a valid handle, according to these criteria.

However, a much smaller share of users generated enough tweets (10) to reliably estimate their ideological expression. We collected about 1.6 million tweets created at any point in 2019 from these accounts, and about 20 million from a sample of followed accounts.[5] We collected tweets from users on a rolling basis beginning in November 2019. The analysis sample includes $N = 995$ respondents.

6.3 Tweet Ideology Classifier

To measure tweet ideology, we developed a text classifier tailored to this project. (The classifier is discussed in detail in Appendix B.) Importantly, the classifier was trained solely on tweets from the YouGov sample and only subsequently applied to the KnowledgePanel tweets. This "train/test" split was done to mitigate the risks of overfitting and to ensure a robust analysis.

The text classifier was developed with the goal of measuring signals of liberal-conservative identity — that is, ideological identity in the sense of a social category, whose members seek mutual affirmation, and so advertise their membership through public displays of group-prototypical behavior (cf. Turner 1991). Since the premise of our project was that this public identity expression was related to, but meaningfully distinct from, private liberal-conservative affinity, we began by drawing a balanced sample of 10,000 tweets from self-identified liberals, conservatives, and moderates in the YouGov dataset. After excluding spam and duplicates, the annotation sample comprised 9,473 tweets.

[4]Weights for the survey were created by raking to estimated population totals, using both behavioral and demographic targets from a separate survey of U.S. adults (linked with their Twitter account data) fielded as part of Pew Research Center's American Trends Panel.

[5]Informed by our previous experience of scraping YouGov followee tweets, sampling of KnowledgePanel followees selected accounts with 500,000 or more followers.

These tweets were then annotated in triplicate by a group of highly skilled MTurkers, using the annotation interface shown in Figure 2. This interface was specifically designed to capture overall self-presentation in tweets rather than narrowly defined liberal-conservative political attitudes: we instructed coders to "guess whether the person who wrote this tweet is liberal or conservative," and encouraged them to rely on any stereotypes they might consider relevant. Since this meant that coders often applied the "Liberal" and "Conservative" labels with uncertainty, we also instructed them to indicate whether they were "Sure," "Not so sure," or had "No idea at all."

These triplicate ideology and sureness labels were processed (using a modified majority rule procedure) into binary liberal/conservative and sure/not sure labels for model training. We then trained two binomial lasso-regularized regression classifiers to predict the liberal/conservative ($-1/+1$) labels and the sure/not sure (0/1) labels respectively, based on tweet text features, using cross-validation to choose the optimal regularization tuning parameters λ. Tweet text was pre-processed to remove punctuation, numbers, URLs, and standard English stopwords. The remaining terms were stemmed and augmented with bigrams (adjacent terms combined and added as new features) to capture meaning in two-word phrases.

This resulted in a fairly simple bag-of-words text classifier, which identifies certain terms as predictive of liberal slant (e.g., "human," "join," "#imwithh[er]"), and certain terms as predictive of conservative slant (e.g., "god," "media," "#maga"). To illustrate this, we plot term coefficients of the ideology model in Figure 3 (the term coefficients of the sureness classifier are included in Appendix B). The classifier is then able to predict the ideology label that a human coder might give to a new tweet by identifying these meaningful terms, adding up their positive and/or negative coefficients, and assigning a predicted ideology label according to the sign of this total.

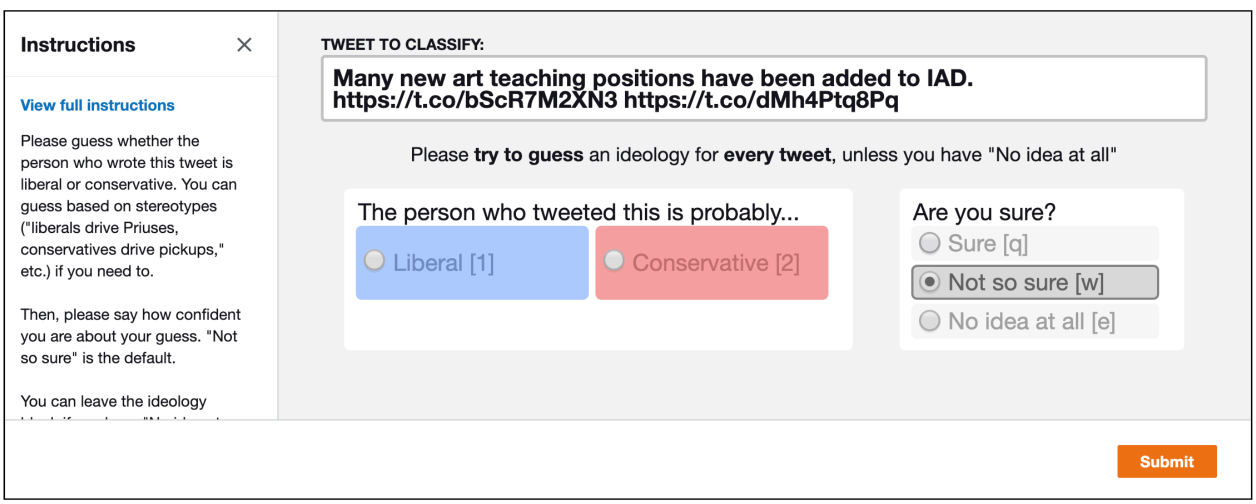


Figure 2: Annotation interface for ground-truth data collection in Amazon Mechanical Turk.

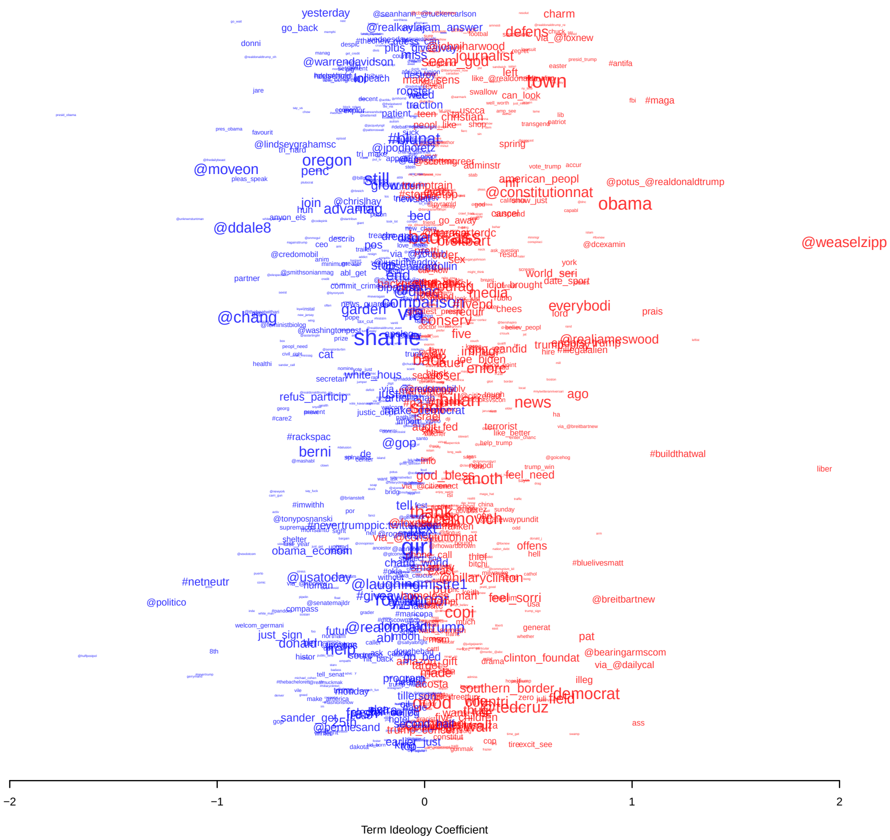


Figure 3: Term coefficients from liberal-conservative binomial lasso model. Terms with negative coefficients (blue) are predictive of liberal slant, whereas terms with positive coefficients (red) are predictive of conservative slant. Text size is proportional to term prevalence in the training corpus.

Despite its simplicity, this method of classification proved reasonably effective: the ideology model demonstrated 66% precision and 56% recall for conservative tweets, and 61% precision and 70% recall for liberal tweets. The sureness model, meanwhile, identified high-sureness tweets with 85% precision and 62% recall. Reassuringly, these performance statistics compare favorably with those of classifiers developed to predict the partisan affiliation of Twitter *users* (see Barberá 2016, Table 3).

6.4 Aggregating Predictions and Face Validity: YouGov Sample

In order to characterize the overall ideological slant of survey respondents' tweets, and the tweets they could have been exposed to from their followees, we first applied both the ideology and sureness models to the entire set of tweets from survey respondents' own accounts, and the sample of accounts they followed, and multiplied the ideology and sureness predictions together. This has the effect of converting a tweet's $-1/+1$ ideology label to 0 when the sureness classifier predicts that a human would be unsure how to classify that tweet. This procedure results in three possible ideology values for each tweet: -1 (Liberal), 0 (Unsure), and $+1$ (Conservative).

Then, we created user-level average tweet ideology estimates by taking the average of the individual tweet predictions for all tweets available from each user (including low-sureness tweets on the basis that this best reflects overall self-presentation and more accurately describes apolitical users). To create aggregate tweet ideology estimates for users' followees, we took the average predicted ideology of all tweets that each user could have seen by selecting tweets from accounts they followed in periods when they followed that account.[6]

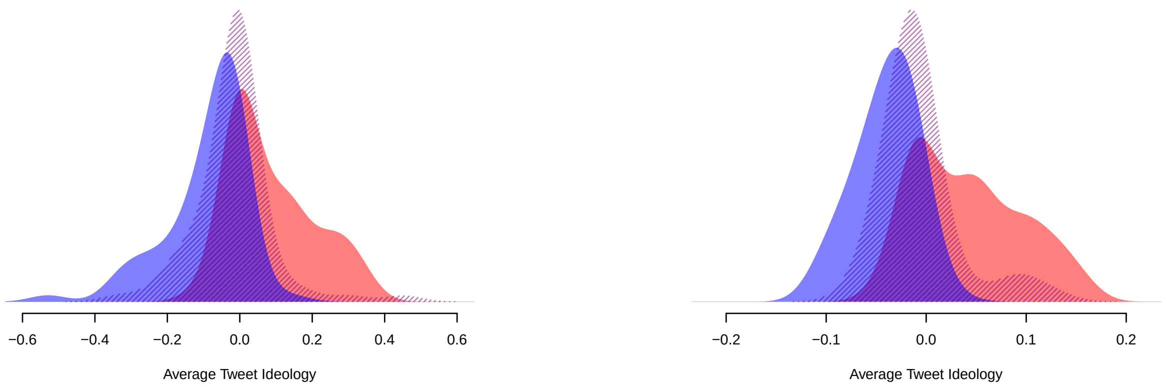
To assess the face validity of these aggregated ideology estimates, Figure 4 plots the distribution of users' own tweet ideology (panel a), and their followees' tweet ideology (panel b), according to their self-described ideology as reported on the survey. This analysis includes all ($\sim$ 1m) tweets from the YouGov respondents, and all ($\sim$ 20m) tweets from accounts they followed. The results are consistent with **H1**: Self-described liberals tweet things (and follow people who tweet things) that the classifier considers more liberal than the corresponding tweets of self-described conservatives, while self-described moderates tweet (and follow people who tweet) close to the center.

7 Main Results

To test our primary hypothesis (**H2**), we estimate an OLS regression of users' tweet ideology on users' survey ideology and their followee tweet ideology. We estimate models using the pooled survey and Twitter data from both the YouGov panel and the much larger KnowledgePanel dataset. The text ideology classifier was trained on tweets sampled from YouGov dataset only, with no contact with the KnowledgePanel dataset prior to this analysis. Likewise, the regression models were specified using the YouGov data and were estimated on the pooled data with no modifications (except where question response format necessitated adjustments in the robustness check discussed in Section 7.1).

In the analyses presented below, the YouGov and KnowledgePanel measures of survey ideology have been processed to aid comparability: The YouGov survey measured self-reported ideology on a 5-point scale, taking repeated measurements over multiple survey waves, which we averaged and rescaled to range from -1 (maximum liberal) to $+1$ (maximum conservative), matching the range of our estimates of tweet ideology. The KnowledgePanel survey measured self-reported ideology only once, on a 10-point scale, which was rescaled to the same -1 to $+1$ interval. This permitted the same model specification to be used to analyze the pooled data.

[6]In the YouGov sample, followee lists were collected 4 times, between November 2018 and May 2019, and followee tweets were included if they were tweeted by accounts on one of these followee lists during a window of time near the date of collection of the relevant followee list (windows being defined by the period since the collection of the previous followee list, or in the case of the first followee list, the previous 5 months). In the KnowledgePanel sample, followee lists were collected only once, and followee tweets were included if they were tweeted by the accounts a user followed after the creation date of the first of the relevant user's own tweets in the dataset.



(a) Own Tweet Ideology

(b) Followee Tweet Ideology

Figure 4: Distribution of estimated tweet ideology for YouGov survey respondents' own tweets (panel a) and their followees' tweets (panel b), grouped by survey self-described ideology: self-described liberals (blue), self-described conservatives (red), self-described moderates (crosshatched).

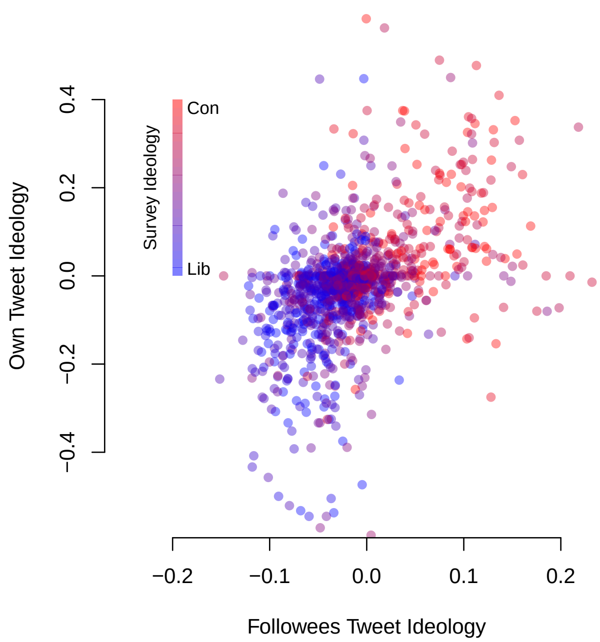


Figure 5: Pooled data comparison of users' own tweet ideology (y-axis), own survey ideology (color), and followees' tweet ideology (x-axis).

Table 1: Pooled Data: Own Tweet Ideology

	Survey	Followees	Survey + Followees	+ Demographics
	(1)	(2)	(3)	(4)
Survey Ideo	0.082***		0.032***	0.035***
	(0.006)		(0.007)	(0.007)
Followee Tweet Ideo		1.158***	0.944***	0.963***
		(0.062)	(0.076)	(0.082)
Age				−0.001***
				(0.0002)
Constant	−0.008**	−0.005	−0.003	0.047
	(0.003)	(0.003)	(0.003)	(0.031)
Observations	1,282	1,282	1,282	1,137
R^2	0.134	0.214	0.228	0.246
Adjusted R^2	0.133	0.213	0.226	0.231

Note: *p<0.1; **p<0.05; ***p<0.01
Coefficients suppressed for non-significant demographic controls: race, gender, education, income.

The results are consistent with **H2**: Figure 5, along with Tables 1 and 2, illustrate the regression of users' own tweet ideology (y-axis) on followees' tweet ideology (x-axis) and own survey ideology (color). Table 1 shows that both survey ideology and followees' tweet ideology are statistically significant predictors of own tweet ideology, that they explain a substantively important proportion (more than 20%) of the variation in own expressed ideology (models 1, 2, and 3), and that this finding is robust to the inclusion of demographic controls (model 4).

Table 2 is informative for **RQ1** and **RQ2**. It shows that, when the sample is split into three[7] categories according to self-described ideology (liberals, moderates, and conservatives), followee tweet ideology explains far more within-subgroup variation in own tweet ideology than does survey ideology. Indeed, when both survey ideology and followee tweet ideology are included in the model, only the latter coefficient remains significant. When survey ideology alone is used to predict own tweet ideology, its coefficient is only significant at standard levels for conservative respondents; for moderates it is borderline significant, and for liberals self-described ideology is not at all significant as a predictor of tweeted ideology.

Apart from differences in coefficient significance, it is noteworthy to observe the differences in

[7] The survey ideology scale (ranging from −1 to +1) was split into equal thirds, by designating as "liberals" all respondents with survey ideology less than or equal to −.3334, designating as "conservatives" all respondents with survey ideology greater than or equal to +.3334, and designating as "moderate" all respondents with survey ideology between −.3334 and +.3334.

the R^2 statistic for models 1, 4, and 7: Followee tweet ideology accounts for approximately 8% of the variance in self-described liberals' own ideological expression, compared to 10% for self-described moderates and 16% for self-described conservatives. If *conformity* is defined as the proportion of one's own expression that is explained by social context, we would conclude that conservatives are nearly twice as likely to exhibit conformity in political expression as liberals. However, this is only one way of operationalizing conformity, and followee tweet ideology is a strongly significant predictor of own expression in all three groups. Further, it is important to note that our subgroup sample sizes are small and that, despite our best efforts to train the classifier on a balanced sample, we cannot completely rule out the possibility of differential measurement error across ideological groups.

What these results show most clearly is that political expression is a distinct phenomenon from political attitudes, even or especially within ideological subgroups of the American ideological spectrum. What makes a liberal or conservative more or less outspoken appears to include social context, and (at least sometimes) includes self-perceived ideological position, but much of the variance in this quantity remains to be explained — an important goal for future research on political expression.

These results are consistent with the hypothesis of social influence in ideological expression on Twitter, although they are also consistent with other explanations. First, it is possible that the similarity between own tweets and followees' tweets is due in part to retweeting. However, retweets constitute a substantively meaningful form of social influence built into the Twitter environment which we argue should be included in this analysis.[8]

Second, the similarity could be due to homophily (McPherson, Smith-Lovin and Cook 2001). However, the fact that the relationship between own tweet ideology and followees' tweet ideology persists when survey ideology is included in the regression indicates that the result is not merely due to ideologues preferring to follow similar ideologues on Twitter. That being said, it remains possible that the followee tweet ideology measure is simply capturing residual variation in survey ideology not captured in survey responses. In future work, we plan to estimate latent ideology based on issue attitudes to address this possibility.

Third, it may be the case that the correlation between own tweet ideology and followees' tweet ideology that remains after accounting for survey ideology is not ideological in nature. That is, although the distribution plots in Figure 4 indicate that tweet ideology correlates with self-described ideology, it may also capture factors other than ideology, such as topic, and it may be the case that covariance in these non-ideological factors of own expression and followees' expression is responsible for the correlation observed after controlling for survey ideology. So these results are not conclusive evidence of social influence over ideological expression on Twitter, but they are fully consistent with that hypothesis.

[8] This result is robust to the exclusion of retweets in the YouGov data, but this robustness check has not yet been conducted for the analysis of the pooled YouGov-KnowledgePanel dataset.

Table 2: Pooled Data: Own Tweet Ideology (Grouped by Private Self-Reported Ideology)

	Liberals			Moderates			Conservatives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Survey Ideo		0.028	0.011		0.064*	−0.004		0.088**	0.050
		(0.022)	(0.022)		(0.036)	(0.036)		(0.034)	(0.032)
Followee Tweet Ideo	1.042***		1.034***	0.877***		0.881***	0.916***		0.884***
	(0.149)		(0.151)	(0.129)		(0.135)	(0.124)		(0.125)
Constant	−0.021***	−0.043***	−0.014	−0.009*	−0.020***	−0.009*	0.026***	−0.00001	−0.008
	(0.008)	(0.016)	(0.016)	(0.005)	(0.005)	(0.005)	(0.009)	(0.025)	(0.023)
Observations	554	554	554	432	432	432	296	296	296
R^2	0.081	0.003	0.081	0.097	0.007	0.097	0.157	0.022	0.164
Adjusted R^2	0.079	0.001	0.078	0.095	0.005	0.093	0.154	0.018	0.158

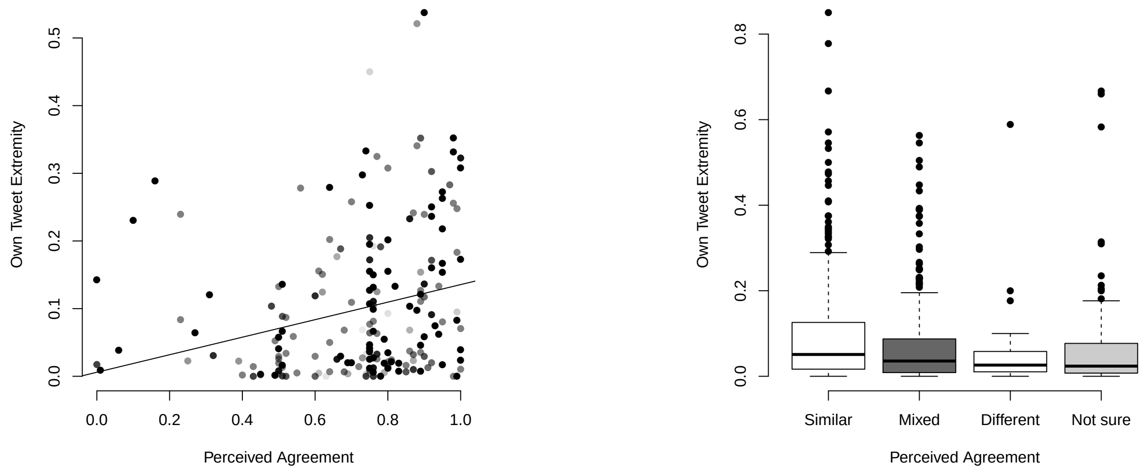
Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

7.1 Robustness Check: Perceived Agreement and Extremity

The third potential concern noted above — that the text classifier might, despite our rigorous training procedure, be measuring substantively non-ideological text features — is a serious one. The textual content of users' tweets is likely to co-vary with that of their followees in any number of dimensions; the theories of social influence discussed above are not limited to the political domain. As such, the residual correlation between own expression and followee expression, which remains significant after accounting for survey-reported ideology, could be interpreted as non-political. In effect, the two tweet ideology measures may exhibit correlated measurement error or apolitical covariance that we would not expect to be accounted for by self-reported ideology. This would severely undermine our interpretation of social influence over public political expression.

To address this concern, we conduct an analysis that does not include followee tweet ideology, and which thereby breaks any potential correlation in measurement error between own tweeted expression and followee tweeted expression. In place of the followee tweet ideology measure derived from our text classifier, we substitute a more proximal measure of users' "imagined audience": their perceived political *similarity* to others in their Twitter networks, as reported on the surveys. We then fold respondents' own tweet ideology at zero, by taking the absolute value of this variable, to construct a new variable which we interpret as "own tweet extremity." We then regress own tweet extremity on the survey-reported Twitter political similarity measures and a folded (absolute value) transformation of survey ideology (as a measure of "self-described extremity"), to replicate the main analyses described above with no risk of correlated measurement error. Since the YouGov and KnowledgePanel questions differed in wording and response format, we present these analyses separately.

The YouGov survey included the question, "Earlier you told us that you are on Twitter. What proportion of the accounts you follow hold political opinions similar to yours (e.g., if you lean left, how many lean left, too)?" and asked respondents to indicate their response on a continuous percentage slider. Figure 6, panel (a) plots the relationship between perceived agreement and own tweet extremity, and the associated regression line estimated in Model 1 of Table 3, showing a significant positive relationship between perceived agreement and own tweet extremity. This coefficient remains strongly significant when survey ideology extremity is included in Model 2, and when political interest is included in Model 3 (although survey ideology extremity is no longer significant when political interest is included).



(a) YouGov Agreement and Extremity

(b) KnowledgePanel Agreement and Extremity

Figure 6: Own tweet extremity and survey-reported perceived agreement with Twitter network, as measured in YouGov (panel a) and KnowledgePanel (panel b) surveys.

As with the main analyses presented in the previous section, models were estimated on the KnowledgePanel dataset only after they were specified and estimated using the YouGov dataset. However, in this case it was necessary to adjust the analysis to accommodate a different question structure. The KnowledgePanel survey included the question, "Now thinking about your friends or people you follow on Twitter... Do most of the people you follow on Twitter have..." and offered the response options, "Similar political beliefs to you," "Different political beliefs from you," "A mix of political beliefs," and "I'm not sure about their political beliefs." Figure 6, panel (b), plots own tweet extremity for users, according to their responses to this question. Model 1 of Table 4 shows that when "Similar political beliefs to you," is designated as the reference category, all other responses ("Different," "Mixed," and "Not Sure") all have statistically significant negative coefficients, reflecting that users who do not perceive their Twitter environment as like-minded are significantly less extreme in their public political expression. When survey ideological extremity (again, calculated as the absolute value of survey ideology) is included in Model 2, all coefficients except "Different from you" (which contains few observations) remain strongly significant. Political interest was not measured in the KnowledgePanel survey and so cannot be included in these models.

These results indicate that the social influence analyses are likely not driven by correlated measurement error between the own- and followee-tweet ideology measures. Moreover, they demonstrate that perceived homogeneity of users' Twitter networks is associated with more extreme public expression on Twitter. So, in addition to a robustness check, this analysis illustrates a significant additional consequence of social influence: when the "imagined audience" is perceived as like-minded, it may encourage the expression of more extreme political sentiments, independent of the strength of users' personal convictions.

Table 3: YouGov Perceived Agreement and Own Tweet Extremity

	Agreement	+Survey	+Interest
	(1)	(2)	(3)
Perc. Agree	0.130***	0.120***	0.102***
	(0.033)	(0.033)	(0.034)
abs(Survey Ideo)		0.033*	0.024
		(0.019)	(0.020)
Political Interest			0.086**
			(0.033)
Constant	0.006	−0.007	−0.064*
	(0.024)	(0.025)	(0.033)
Observations	226	226	226
R^2	0.064	0.076	0.103
Adjusted R^2	0.060	0.068	0.091
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: KnowledgePanel Perceived Agreement and Own Tweet Extremity

	Agreement	+Survey
	(1)	(2)
Different from You	-0.038^{*}	-0.033
	(0.021)	(0.021)
Mixed	-0.033^{***}	-0.026^{***}
	(0.008)	(0.008)
Not Sure	-0.044^{***}	-0.037^{***}
	(0.010)	(0.010)
abs(Survey Ideo)		0.030^{***}
		(0.010)
Constant	0.099^{***}	0.081^{***}
	(0.006)	(0.009)
Observations	992	992
R^2	0.028	0.036
Adjusted R^2	0.025	0.032
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

8 Discussion

This paper has presented a theory of social influence and a novel demonstration of this dynamic in public political expression on Twitter. Our findings indicate that the political slant of users' public tweets is conceptually and empirically distinct from the affiliations they report in attitudinal surveys. This finding calls into question the common analysis of public tweets as a cheaper or more convenient proxy for survey measures, but it also illustrates that public political expression is a distinct phenomenon, worthy of study in its own right. Far from being a public register of political attitudes, public expression is the product of a complex social process recognized by early political behavior scholars (Katz and Lazarsfeld 1955) but only recently accessible in the form of social media posts — thus far a largely untapped resource for studying expression as a political behavior of significance.

Studying political expression on social media harkens back to earlier conceptions of public opinion: a broader repertoire of actions and expressions that are unconstrained by survey response options (Tilly 1983). While opinion researchers claim to map public priorities in a systematic way, critics charge that this process "domesticates" opinion by failing to capture the intensity of preferences and circumscribing political goals (Ginsberg 1986). Even the mode of public opinion measurement — the survey — is fundamentally private. This results in a measurement program that "treats the formation of public opinion as analogous to the vote choices made at polling places on election day" (Sanders 1999, p. 268). Social media provide an alternative conception. Unlike surveys, which treat all respondents as equals regardless of the degree to which they express political opinions or actually shape the political process (Blumer 1948), platforms like Twitter allow those who wish to convey ideological signals to their audiences unlimited opportunities to do so. In this way, political behavior on social media provides a formally unconstrained mode for engagement. However, as we have shown, powerful social constraints still exist online.

We demonstrate a method for analyzing this new data stream and find that public political expression can be classified in terms of ideological identity categories using relatively simple models. Furthermore, we find evidence consistent with the hypothesis that users' public expression is powerfully shaped by their followees, independent of the political ideology they report identifying with in attitudinal surveys. Finally, we find that users' ideological expression is more extreme when they perceive their Twitter networks to be relatively like-minded. These findings have important implications for our understanding of political expression as a social and political behavior, of the social construction of ideology, of political polarization broadly, and of the role of social media platforms in restructuring basic social patterns into new forms.

The evidence we present in this paper advances our ability to study the social determinants of political expression on social media. But since our data and methods are purely observational, our findings do not support a causal interpretation. Additional research is needed to study the causal mechanisms underlying the patterns that we uncover. For example, it may be possible to take advantage of natural experiments due to the suspension of high-follower political accounts. Randomized experiments, both artificial and in the field, could also be designed to identify the effect of exogenous changes to users' following patterns (e.g., Bail et al. 2018). We hope our findings will be generative for researchers interested in uncovering contextual dynamics on social media.

8.1 The Normative Implications of Social Influence

Social influence over public political expression has broad potential consequences, and therefore our findings merit normative discussion. Indeed, influential theorists have described how public preference falsification can hold entire societies under repressive authoritarian regimes (Kuran 1987) and bind them in social arrangements that violate the true preferences of the majority (Bicchieri 2005). Clearly these are undesirable and anti-democratic outcomes.

We cannot dismiss the possibility that these social dynamics contribute to political polarization. We observe that network homogeneity increases expressed extremism, and other scholars (e.g., Settle 2018) have theorized about how exposure to extreme expression on social media can polarize. Other work (e.g., Munger 2017*a*,*b*) demonstrates that social influence can normalize (and de-normalize) racist and other hateful speech online. If social influence leads people to express extreme political sentiments, there is good reason to expect harmful consequences.

That said, our results do not show conclusive evidence that social media is politically harmful. Even if social media does increase political polarization, it has also been found to increase political knowledge among users (Allcott et al. 2020). To the extent that democracy functions well with high levels of political engagement, increased ideological polarization may induce civic participation (Hetherington et al. 2008).

Moreover, it seems highly unlikely that the specific form of social influence we analyze in this paper is a new phenomenon attributable solely to social media platforms. On the contrary, norm-consistent self-presentation is a prerequisite skill for the development of complex social structures (Leary 1995) of the sort that give rise to political competition in the first place. Our analysis may be modern in its methods and data, but its results likely reflect an ancient empirical regularity. We therefore caution against alarmist interpretations of our findings, as we have yet to demonstrate the circumstances under which these dynamics cause discernible harm. The fact that individuals express sentiments in public that they might not register in private may reflect social solidarity or norm adherence, rather than mindless conformity. Distinguishing these kinds of influence is an important task for future research.

References

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *The American Economic Review* 110(3):48.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.1804840115*

Bakker, Bert N., Yphtach Lelkes and Ariel Malka. 2020. "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives." *The Journal of Politics* 82(3):1061–1077.
URL: *https://www.journals.uchicago.edu/doi/10.1086/707616*

Barberá, Pablo. 2016. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data." *Work Pap NYU* .

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost and Joshua A Tucker. 2019. "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data." *American Political Science Review* 113(4):883–901.

Beauchamp, Nicholas. 2017. "Predicting and interpolating state-level polls using Twitter textual data." *American Journal of Political Science* 61(2):490–503.

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.
URL: *http://ebooks.cambridge.org/ref/id/CBO9780511616037*

Blumer, Herbert. 1948. "Public Opinion and Public Opinion Polling." *American Sociological Review* 13(5):542.
URL: *http://www.jstor.org/stable/2087146?origin=crossref*

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker and Jay J. Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences* 114(28):7313–7318.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.1618923114*

Butler, Judith. 1999. *Gender trouble: Feminism and the subversion of identity.* Routledge.

Bäck, Emma A., Hanna Bäck, Annika Fredén and Nils Gustafsson. 2019. "A social safety net? Rejection sensitivity and political opinion sharing among young people in social media." *New Media and Society* 21(2):298–316.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1960. *The American Voter*. John Wiley & Sons.

Carlson, Taylor N. and Jaime E. Settle. 2016. "Political Chameleons: An Exploration of Conformity in Political Discussions." *Political Behavior* 38(4):817–859. Publisher: Springer US.

Chan, Michael. 2018. "Reluctance to Talk About Politics in Face-to-Face and Facebook Settings: Examining the Impact of Fear of Isolation, Willingness to Self-Censor, and Peer Network Characteristics." *Mass Communication and Society* 21(1):1–23. Publisher: Routledge.
URL: *https://doi.org/10.1080/15205436.2017.1358819*

Chen, Hsuan Ting. 2018. "Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors." *New Media and Society* 20(10):3917–3936.

Conover, Pamela Johnston and Stanley Feldman. 1981. "The Origins and Meaning of Liberal/Conservative Self-Identifications." p. 30.

Converse, Philip E. 2006. "The Nature of Belief Systems in Mass Publics." *Critical Review* 18(1-3):1–74. Place: New York Publisher: The Free Press of Glencoe.

Cramer Walsh, Katherine. 2004. *Talking about politics: Informal groups and social identity in american life*. Chicago: University of Chicago Press.

Duggan, Maeve and Aaron Smith. 2016. The Political Environment on Social Media. Technical report Pew Research Center.
URL: *https://www.pewresearch.org/internet/2016/10/25/the-political-environment-on-social-media/*

Eady, Gregory, Jonathan Nagler, Andy Guess, Jan Zilinsky and Joshua A. Tucker. 2019. "How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data." *SAGE Open* 9(1).

Fiesler, Casey and Nicholas Proferes. 2018. ""Participant" Perceptions of Twitter Research Ethics." *Social Media + Society* 4(1):1–14.

Fox, Jesse and Lanier Frush Holt. 2018. "Fear of Isolation and Perceived Affordances: The Spiral of Silence on Social Networking Sites Regarding Police Discrimination." *Mass Communication and Society* 21(5):533–554. Publisher: Routledge.
URL: *https://doi.org/10.1080/15205436.2018.1442480*

Friggeri, Adrien, Lada Adamic, Dean Eckles and Justin Cheng. 2014. Rumor cascades. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Frimer, Jeremy A, Danielle Gaucher and Nicola K Schaefer. 2014. "Political conservatives' affinity for obedience to authority is loyal, not blind." *Personality and Social Psychology Bulletin* 40(9):1205–1214.

Gearhart, Sherice and Weiwu Zhang. 2014. "Gay Bullying and Online Opinion Expression: Testing Spiral of Silence in the Social Media Environment." *Social Science Computer Review* 32(1):18–36.

Gearhart, Sherice and Weiwu Zhang. 2018. "Same Spiral, Different Day? Testing the Spiral of Silence Across Issue Types." *Communication Research* 45(1):34–54.

Ginsberg, Benjamin. 1986. *The Captive Public: How Mass Opinion Promotes State Power.* New York: Basic Books.

Goldberg, Matthew H., Sander van der Linden, Anthony Leiserowitz and Edward Maibach. 2020. "Perceived Social Consensus Can Reduce Ideological Biases on Climate Change." *Environment and Behavior* 52(5):495–517. Publisher: SAGE Publications Inc.
URL: *https://doi.org/10.1177/0013916519853302*

Green, Donald, Bradley Palmquist and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters.* New Haven: Yale University Press.
URL: *http://poq.oxfordjournals.org/content/67/4/660.full.pdf*

Groenendyk, Eric W. 2013. *Competing Motives in the Partisan Mind.* New York: Oxford University Press.

Groenendyk, Eric W, Erik O. Kimbrough and Mark Pickup. 2020. "How Norms Shape the Nature and Origins of Mass Belief Systems." *SSRN Electronic Journal* .
URL: *https://www.ssrn.com/abstract=3541289*

Haidt, Jonathan, Jesse Graham and Craig Joseph. 2009. "Above and below left–right: Ideological narratives and moral foundations." *Psychological Inquiry* 20(2-3):110–119.

Hampton, Keith, Lee Rainie, Weixu Lu, Maria Dwyer, InYoung Shin and Kristen Purcell. 2014. "Social Media and the Spiral of Silence." *Pew Research Center* pp. 1–44.

Hetherington, Marc J., Deborah Jordan Brooks, John G. Geer and Martin P. Wattenberg. 2008. Turned off or turned on?: How polarization affects political engagement. In *Red and blue nation?: Consequences and correction of america's polarized politics.* Brookings Institution Press pp. 1–54.
URL: *http://www.jstor.org/stable/10.7864/j.ctt6wpdqn.4*

Hoffmann, Christian Pieter and Christoph Lutz. 2017. "Spiral of Silence 2.0: Political Self-Censorship among Young Facebook Users." *Proceedings of the 8th International Conference on Social Media & Society* 8:1–12.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive partisanship: Campaign involvement, political emotion, and partisan identity." *American Political Science Review* 109(1):1–17.

Iyengar, Shanto and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* . tex.ids: Iyengar2015a.

Jost, John T., Christopher M. Federico and Jaime L Jamie L. Napier. 2013. "Political Ideologies and their Social Psychological Functions." *The Oxford Handbook of Political Ideologies* 1(1908):232–250. Publisher: OUP.
URL: *http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199585977.001.0001/oxfordhb-9780199585977-e-024*

Jost, John T, Sander van der Linden, Costas Panagopoulos and Curtis D Hardin. 2018. "Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation." *Current Opinion in Psychology* 23:77–83.
URL: *http://www.sciencedirect.com/science/article/pii/S2352250X17302828*

Kahan, Dan M. 2013. "Ideology, Motivated Reasoning, and Cognitive Reflection." *Judgment and Decision Making* 8(4):407–424.

Katz, Elihu and Paul F Lazarsfeld. 1955. *Personal influence: the part played by people in the flow of mass communications.* New York, NY, US: Free Press. Pages: xx, 400 Publication Title: Personal influence: the part played by people in the flow of mass communications.

Kim, Mihee. 2016. "Facebook's Spiral of Silence and Participation: The Role of Political Expression on Facebook and Partisan Strength in Political Participation." *Cyberpsychology, Behavior, and Social Networking* 19(12):696–702.

King, Gary, Benjamin Schneer and Ariel White. 2017. "How the news media activate public expression and influence national agendas." *Science* 358(6364):776–780.

Kuran, Timur. 1987. "Preference Falsification , Policy Continuity and Collective Conservatism." *The Economic Journal* 97:642–665.

Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification. Cambridge, MA: Harvard University Press.* Harvard University Press.

Lazarsfeld, Paul F., Bernard Berelson and Hazel Gaudet. 1948. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign.* 2nd ed. New York: Columbia University Press.

Leary, Mark R. 1995. *Self-Presentation: Impression Management and Interpersonal behavior.* Social Psychology Series Madison, Wisconsin: Brown & Benchmark Publishers.

Lenz, Gabriel S. 2012. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance.* University of Chicago Press.

Levitan, Lindsey C. and Brad Verhulst. 2016. "Conformity in Groups: The Effects of Others' Views on Expressed Attitudes and Attitude Change." *Political Behavior* 38(2):277–315.

Levitin, Teresa E. and Warren E. Miller. 1979. "Ideological Interpretations of Presidential Elections." *American Political Science Review* 73(3):751–771.

Lupia, Arthur and Matthew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge: Cambridge University Press.

Luskin, Robert C. 1990. "Explaining political sophistication." *Political Behavior* 12(4):331–361.

Malka, Ariel and Yphtach Lelkes. 2010. "More than Ideology: Conservative–Liberal Identity and Receptivity to Political Cues." *Social Justice Research* 23(2-3):156–188.
URL: *http://link.springer.com/10.1007/s11211-010-0114-3*

Marwick, Alice E. and Danah boyd. 2011. "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience." *New Media & Society* 13(1):114–133. tex.eprint: https://doi.org/10.1177/1461444810365313.
URL: *https://doi.org/10.1177/1461444810365313*

Mason, Lilliana. 2018. "Ideologues without Issues: The Polarizing Consequences of Ideological Identities." *Public Opinion Quarterly* 82(S1):866–887.
URL: *https://academic.oup.com/poq/article/82/S1/866/4951269*

McGregor, Shannon C. 2019. "Social media as public opinion: How journalists use social media to represent public opinion." *Journalism* 20(8):1070–1086.

McGregor, Shannon C. 2020. ""Taking the Temperature of the Room" How Political Campaigns Use Social Media to Understand and Represent Public Opinion." *Public Opinion Quarterly* 84(S1):236–256.

McPherson, Miller, Lynn Smith-Lovin and James M Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual review of sociology* 27(1):415–444.

Munger, Kevin. 2017*a*. "Experimentally Reducing Partisan Incivility on Twitter." pp. 1–31.
URL: *http://cess.nyu.edu/wp-content/uploads/2017/02/Dont-@-Me.pdf*

Munger, Kevin. 2017*b*. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39(3):629–649. Publisher: Springer US.

Munzert, Simon, Pablo Barberá, Andrew Guess and JungHwan Yang. Forthcoming. "Do Online Voter Guides Empower Citizens? Evidence from a Field Experiment with Digital Trace Data.".

Neubaum, German and Nicole C. Krämer. 2017. "Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media." *Media Psychology* 20(3):502–531. Publisher: Routledge.
URL: *https://doi.org/10.1080/15213269.2016.1211539*

Neubaum, German and Nicole C. Krämer. 2018. "What Do We Fear? Expected Sanctions for Expressing Minority Opinions in Offline and Online Communication." *Communication Research* 45(2):139–164.

Nguyen, Thu T, Nikki Adams, Dina Huang, M Maria Glymour, Amani M Allen and Quynh C Nguyen. 2020. "The Association Between State-Level Racial Attitudes Assessed From Twitter Data and Adverse Birth Outcomes: Observational Study." *JMIR Public Health and Surveillance* 6(3):e17103.
URL: *https://publichealth.jmir.org/2020/3/e17103*

Noel, Hans. 2012. "The Coalition Merchants: The Ideological Roots of the Civil Rights Realignment." *The Journal of Politics* 74(1):156–173.
URL: *https://www.journals.uchicago.edu/doi/10.1017/S0022381611001186*

Noelle-Neumann, Elisabeth. 1991. The Theory of Public Opinion: The Concept of the Spiral of Silence. In *Communication Yearbook*. International Communication Association pp. 256–287.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence A Theory of Public Opinion." *Journal of Communication* 24(2):43–51.

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2020. "Partisan polarization is the primary psychological motivation behind "fake news" sharing on Twitter.".

Posegga, Oliver and Andreas Jungherr. 2019. Characterizing Political Talk on Twitter: A Comparison Between Public Agenda, Media Agendas, and the Twitter Agenda with Regard to Topics and Dynamics. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Sanders, Lynn M. 1999. "Democratic Politics and Survey Research." *Philosophy of the Social Sciences* 29(2):248–280.
URL: *http://journals.sagepub.com/doi/10.1177/004839319902900205*

Settle, Jamie E. 2018. *Frenemies: How Social Media Polarizes America*. New York: Cambridge University Press.

Stier, Sebastian, Johannes Breuer, Pascal Siegers and Kjerstin Thorson. 2019. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* .

Tilly, Charles. 1983. "Speaking Your Mind Without Elections, Surveys, or Social Movements." *Public Opinion Quarterly* 47(4):461–478.
URL: *https://academic.oup.com/poq/article-lookup/doi/10.1086/268805*

Tufekci, Zeynep. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Turner, John C. 1991. *Social influence.* Bristol, PA: Open University Press.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. “The spread of true and false news online.” *Science* 359(6380):1146–1151.

Vraga, Emily K., Kjerstin Thorson, Neta Kligler-Vilenchik and Emily Gee. 2015. “How individual sensitivities to disagreement shape youth political expression on Facebook.” *Computers in Human Behavior* 45:281–289. Publisher: Elsevier Ltd.
URL: *http://dx.doi.org/10.1016/j.chb.2014.12.025*

Williams, Matthew L, Pete Burnap, Luke Sloan, Curtis Jessop and Hayley Lepps. 2017. Users’ View of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm. In *The Ethics of Online Research.* Vol. 2 of *Advances in Research Ethics and Integrity* Emerald Publishing LImited pp. 27–52.

Wojcik, Stefan and Adam Hughes. 2019. “Sizing Up Twitter Users.” Pew Research Center, April 24, 2019. Downloaded September 5, 2020 from `https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/`.

Woong Yun, Gi and Sung Yeon Park. 2011. “Selective Posting: Willingness to post a message online.” *Journal of Computer-Mediated Communication* 16(2):201–227.

Wu, Tai Yee and David J. Atkin. 2018. “To comment or not to comment: Examining the influences of anonymity and social support on one’s willingness to express in online news discussions.” *New Media and Society* 20(12):4512–4532.

Yang, Qi, Khizar Qureshi and Tauhid Zaman. 2020. “Mitigating the Backfire Effect Using Pacing and Leading.” *arXiv:2008.00049 [cs, stat]* . arXiv: 2008.00049.
URL: *http://arxiv.org/abs/2008.00049*

A YouGov Survey Panel Composition

This section provides summary information regarding the composition of the survey panel, subset to reflect only those respondents with Twitter data used in this analysis.

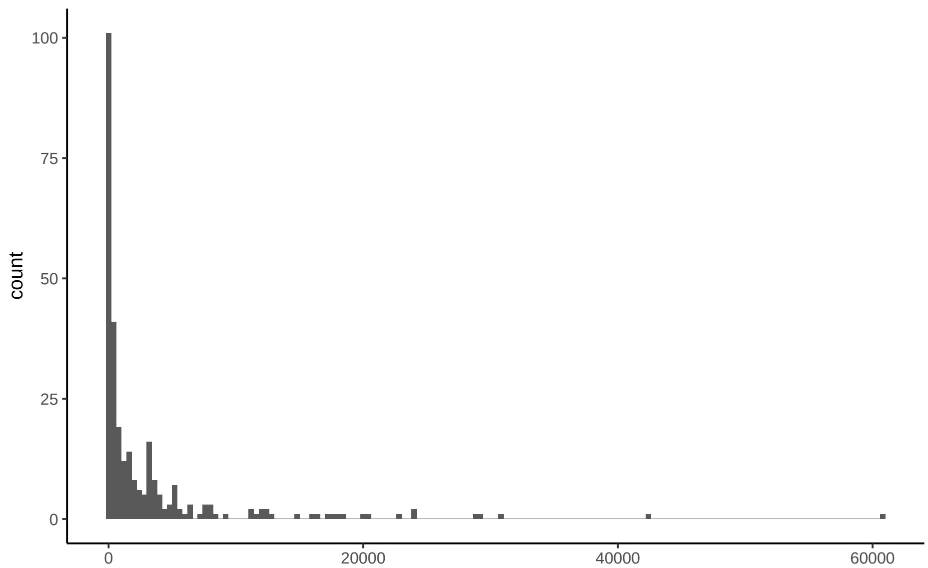


Figure 7: Number of Own Tweets Collected

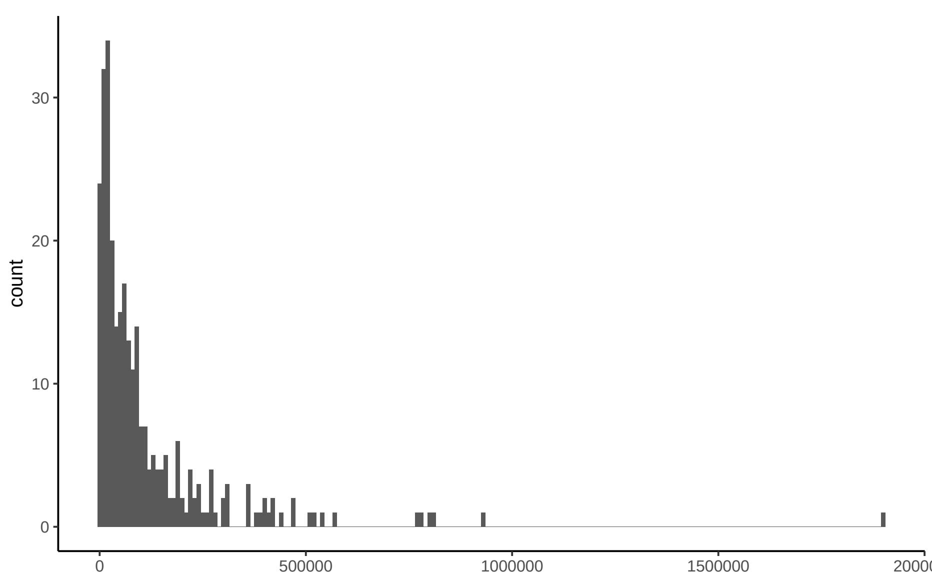


Figure 8: Number of Followee Tweets Collected

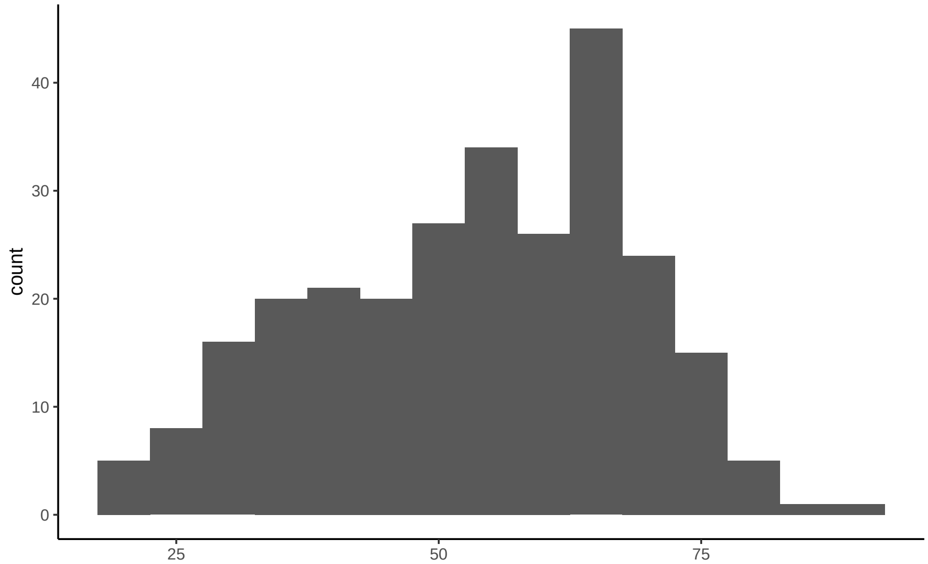


Figure 9: Age

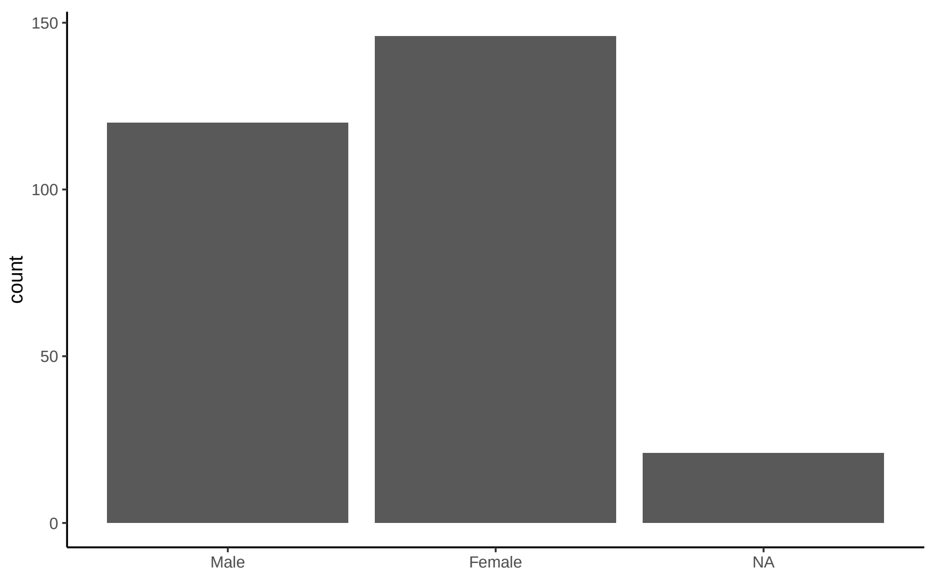


Figure 10: Gender

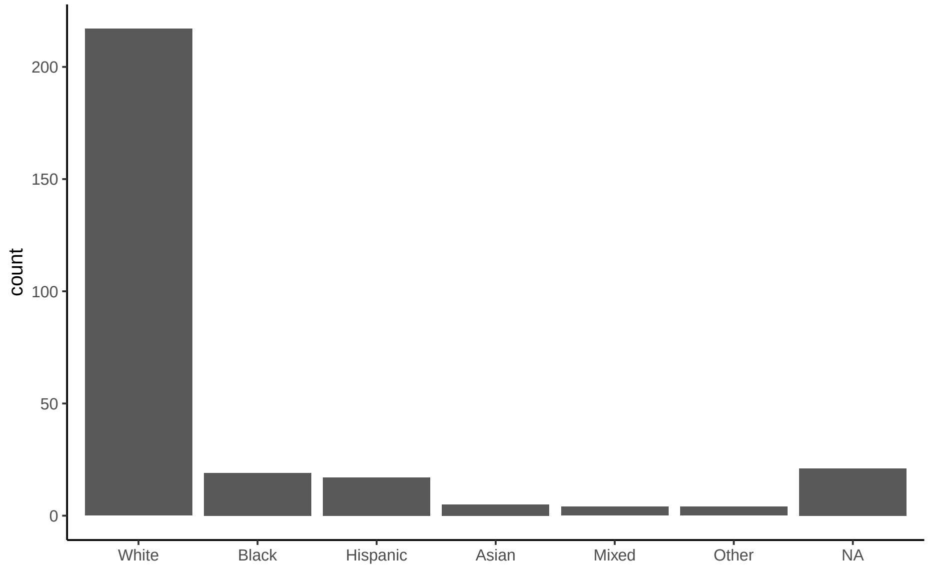


Figure 11: Race/Ethnicity

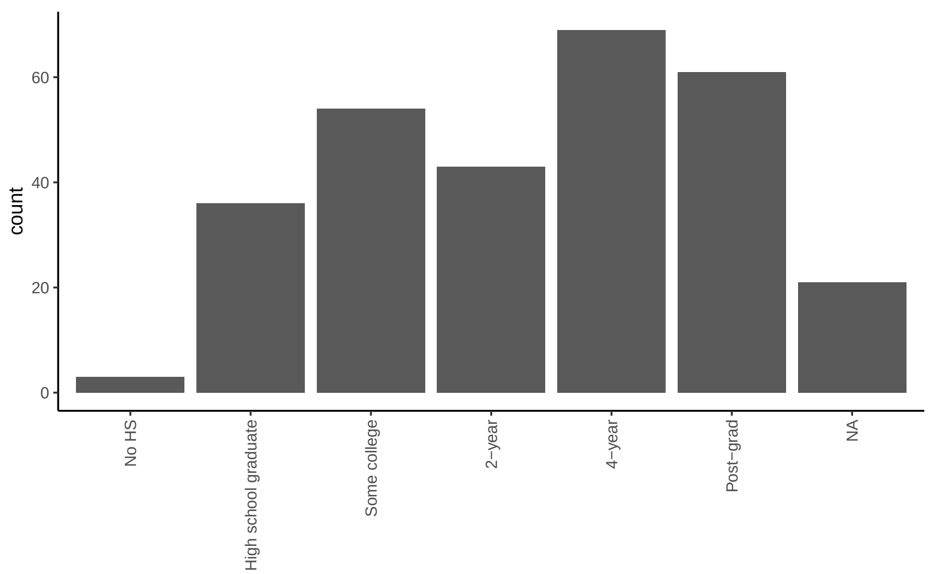


Figure 12: Education

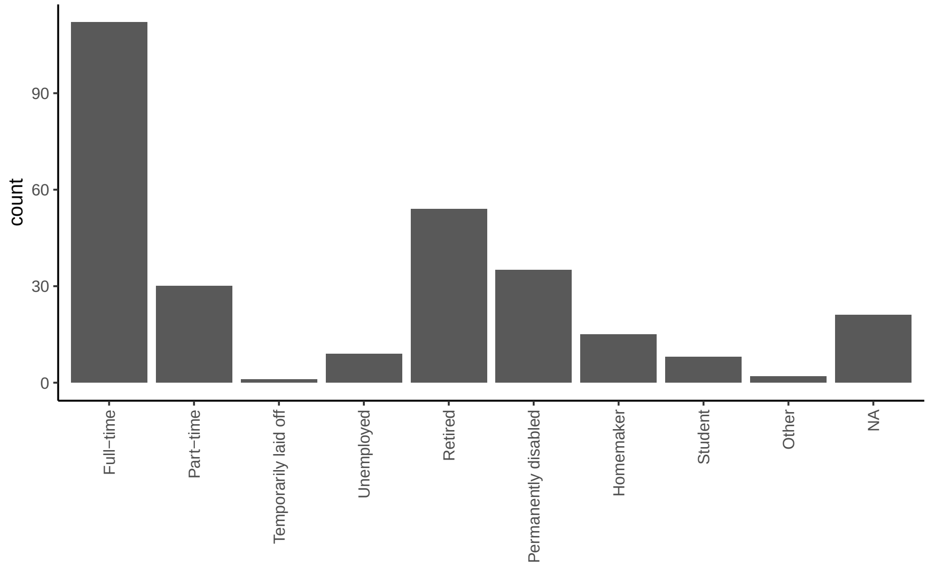


Figure 13: Employment Status

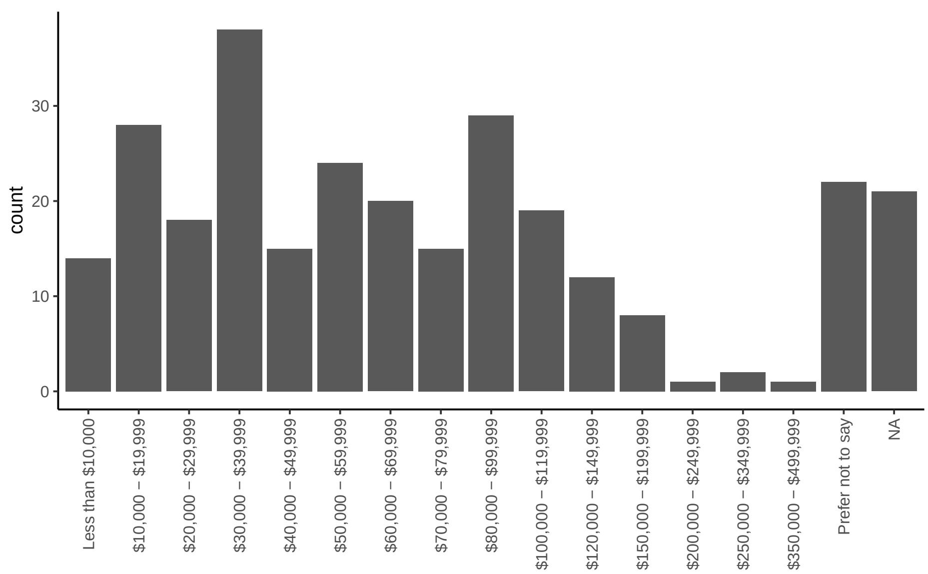


Figure 14: Family Income

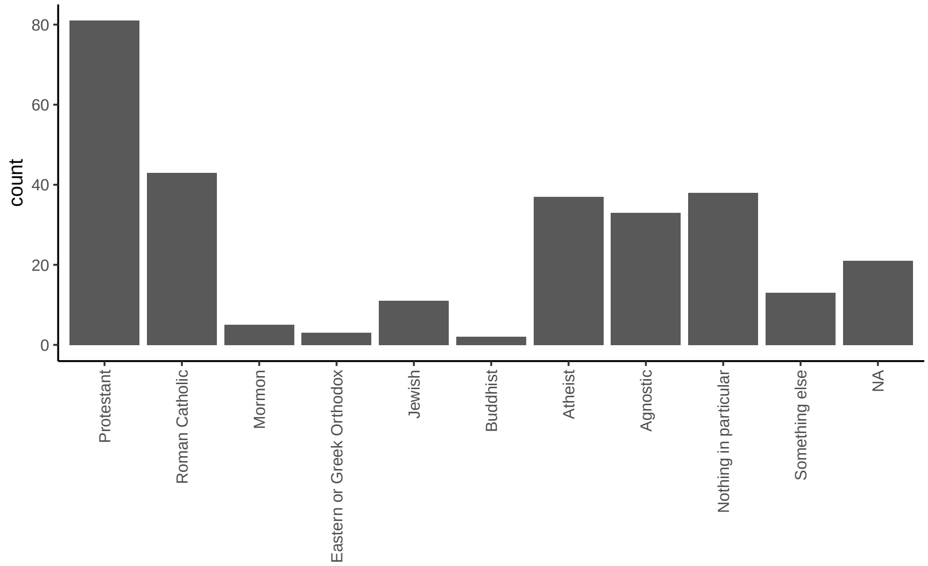


Figure 15: Religion

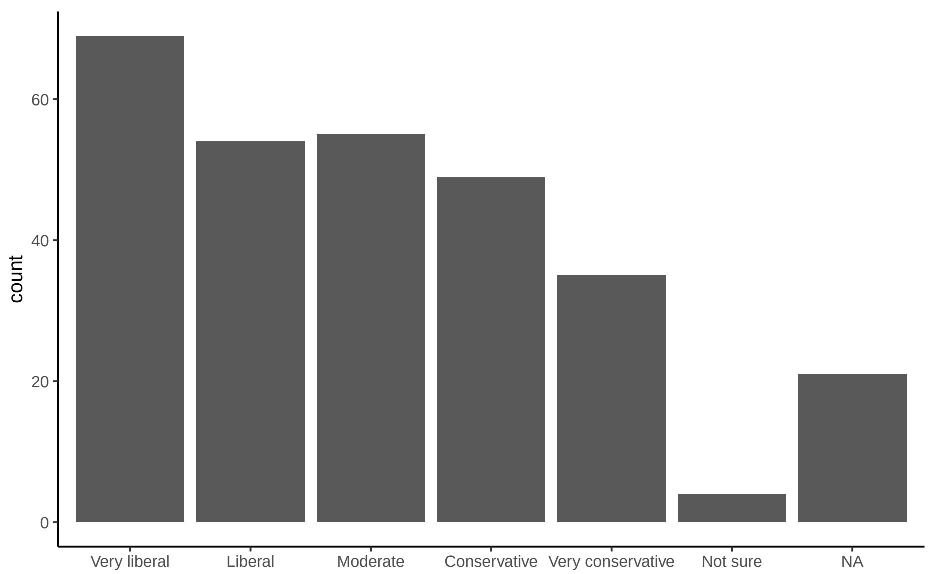


Figure 16: Ideological Identification

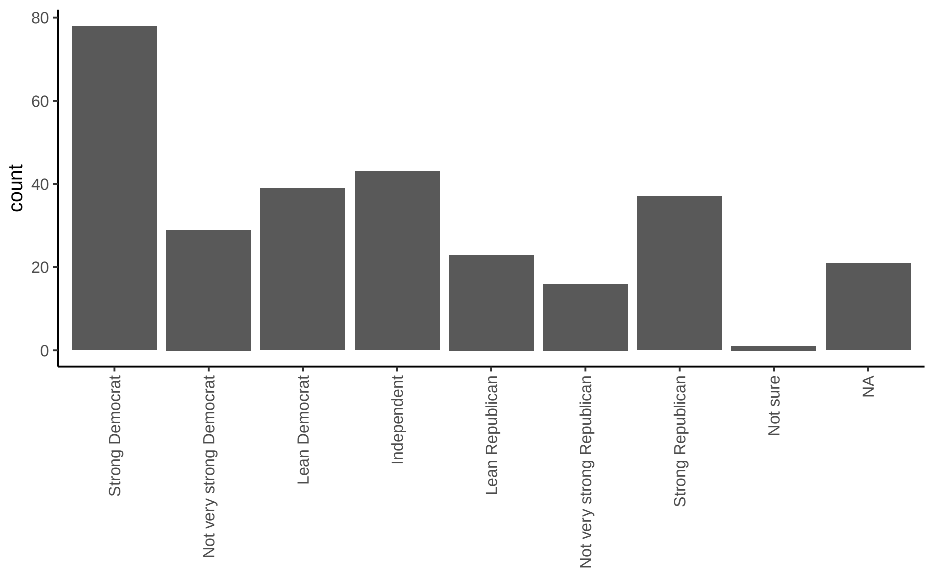


Figure 17: Partisan Identification

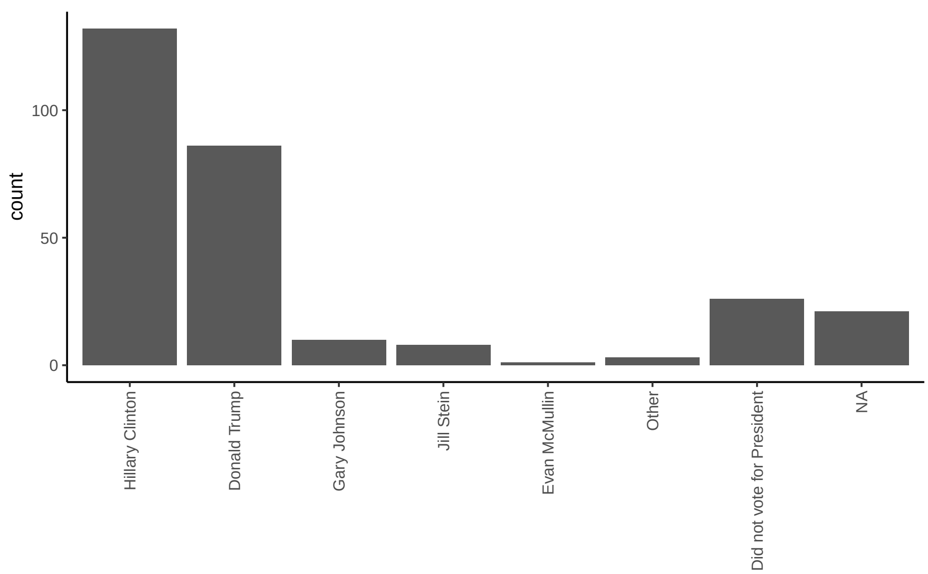


Figure 18: 2016 Presidential Vote

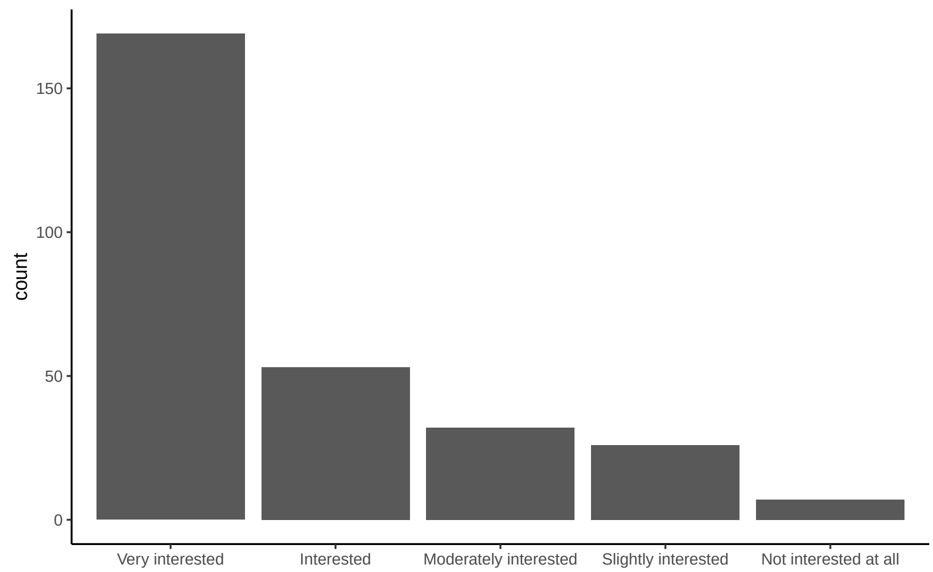


Figure 19: Political Interest

B Classifier Development

B.1 Ground Truth Training Data

To develop the current tweet ideology classifier, we began by developing a tweet annotation interface for use with Amazon Mechanical Turk. The design of this interface reflected our substantive perspective on the ideological information contained in tweets, and our need to make the system as efficient as possible for crowd workers to use.

The initial attempt at designing an annotation interface was clunky and over-complicated (top pane of Figure 20). It has been iteratively revised, resulting in the interface shown in the bottom pane of Figure 20. This interface prompts the annotator to give both an ideology label and a confidence level. Earlier versions of the interface incorporated a "Don't Know" option into the list of possible ideological labels, but this resulted in annotators choosing "Don't Know" whenever they perceived a tweet to be ambiguous, and thus failed to capture valuable information from medium-confidence guesses. Still, it proved necessary to retain a "No idea at all" option, since it was irritating to code truly ambiguous tweets without this option. We decided to set "Not so sure" as the default sureness response, for several reasons: first, we believed this was the modal category, and therefore setting this default would reduce workers' effort and fatigue; second, we wished to normalize the expectation that most labels would be given with some uncertainty, so that workers would not exaggerate or deprecate their confidence; third, setting this default makes guessing (choosing an ideology, and leaving sureness at "Not so sure") and abstaining (not choosing an ideology, and selecting "No idea at all") equally effortful, since both require one click/keystroke, and therefore there is no implicit incentive to say "No idea at all" in order to save effort.

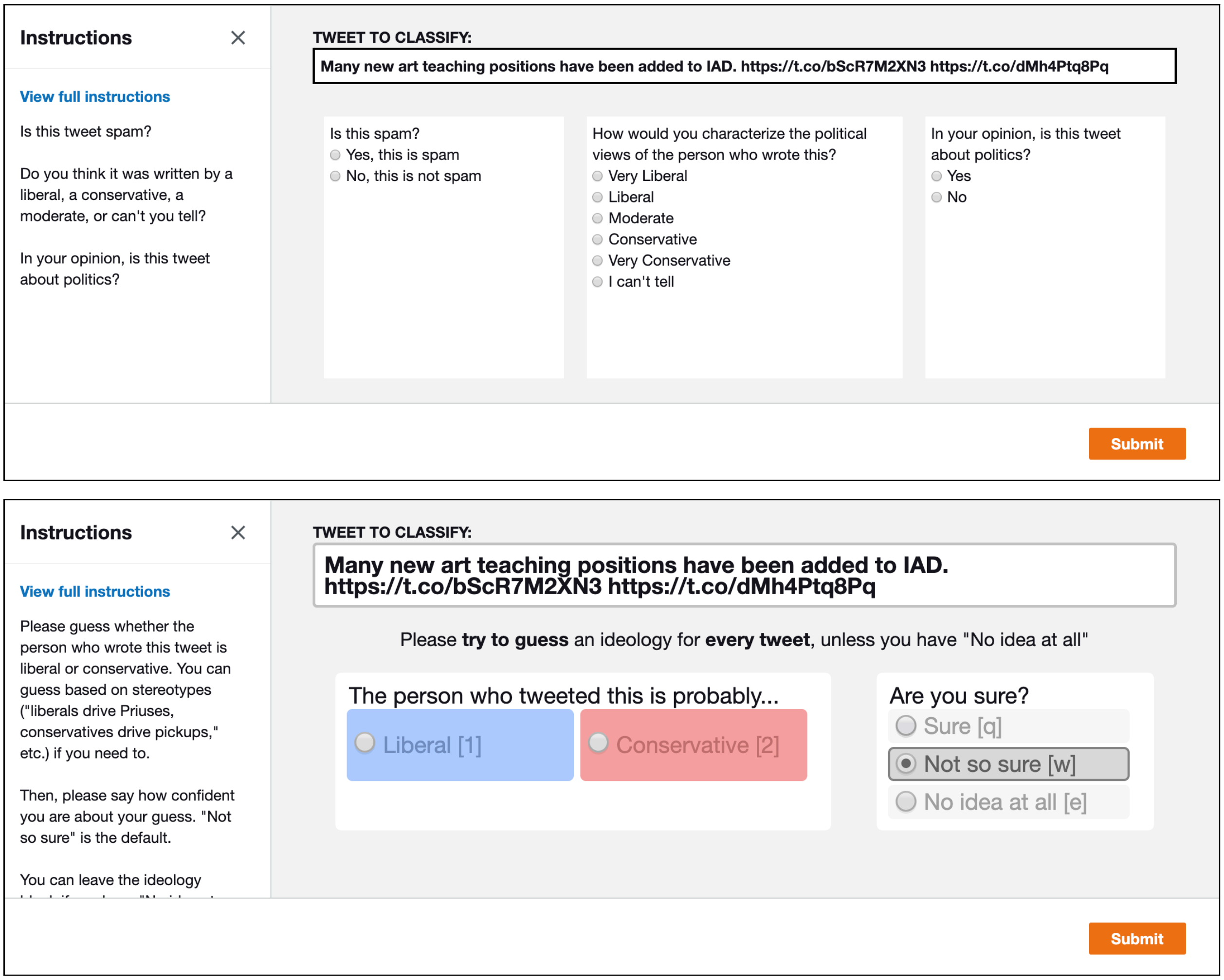


Figure 20: Annotation interface for ground truth data collection in MTurk. First draft shown on top; final draft shown on bottom. Spam question was dropped, because non-ML methods of spam removal proved highly effective. Ideology label was simplified to a lib-con forced choice, because in earlier versions too many workers indicated "moderate" or "don't know" for tweets where an ideological guess was possible – including these options implied that the task was to report ideology only when it could be guessed with high confidence. Instead, a separate question, "Are you sure?" was added, to measure confidence, and "Not so sure" is checked by default, since otherwise the easiest (and thus the incentivized) response is to say "No idea at all" for each tweet. The question, "In your opinion, is this tweet about politics?" was removed, since in this theoretical context, responses to this question were deemed informationally redundant with "Are you sure?" responses.

Next, we needed to identify skilled workers to annotate future large batches of tweets. In order to do this, we recruited from a set of 84 workers to annotate 50 tweets that we had already coded for ground-truth ideology. We estimated worker ability using a two-parameter item response model. We recruited the top-performing workers to annotate future ground truth tweets.

Third, we needed a valid sample of tweets to code for ground truth. To do this, we drew a sample of 10,000 tweets from the set of tweets scraped from the survey respondents' accounts. In order to prevent the most prolific users from dominating this set, sampling was weighted by the inverse of the total tweets available for each user in the set. Since we abandoned the "Is this spam?" element of the annotation interface, it was desirable to remove as much spam as possible from the ground truth set. We achieved this by a combination of methods: prior to sampling, we excluded several users who tweeted almost exclusively promotional content, and after sampling, we dropped tweets containing any features that indicated promotional content (for example, "#sponsored"). Dropping these tweets resulted in a set of 9,473 tweets, annotated by skill-qualified MTurkers using the interface described above. This annotation set was published in batches of approximately 1,000 tweets during April, May, and June 2020.

B.2 Label Quality Assessment

In this section we analyze inter-rater agreement with respect to tweet ideology. Each of 2000 tweets was assigned to three individual annotators recruited from a pool of 18 qualified workers (only 17 participated in the batches published thus far), so agreement is calculated as the proportion of tweets for which all annotators who gave an ideology label (not NA) gave the same label. We omit individual NA ideology labels cell-wise rather than row-wise. The latter, which is more standard practice, would omit any tweets that had at least one NA label, but since two other annotators labels are available for comparison in these cases, row-wise omission would discard valuable data. Instead, we define agreement the number of tweets with multiple nondiscordant labels, divided by the number of tweets with multiple labels. So, this includes cases where all three annotators gave identical labels, and cases where one annotator gave no label (NA) and the other two gave identical labels. Cases of disagreement are those in which at least one "Liberal" label and at least one "Conservative" label exist for the same tweet. Cases where more than one annotator gave no ideology label are excluded, since agreement is undefined. Agreement is calculated solely from the ideology labels. Sureness is then used to subset the dataset to explore how agreement covaries with sureness, shown in tabular form below.

It is standard practice to report agreement rates relative to the expected agreement rate if labels were completely random. This would be relative to 50% when two workers give independent labels, or 25% when three workers give independent labels. Since the present data includes a mixture of 2-fold coding and 3-fold coding, and this mixture will vary with each batch, we adopt a simulation approach to illustrate the expected null distribution of agreeement. We draw simulated datasets by replacing non-NA cells of the observed dataset with Bernoulli noise, and calculate agreement in each of 100,000 simulated datasets.

In Figure 21, below, we plot the distribution of agreement under these null simulations, and plot the observed agreement as a red line. Since agreement covaries with annotator sureness, we present the agreement results by sureness subsets. For example, the first row corresponds to tweets for which all three annotators said they were "sure" of the ideology label they ascribed. The plot shows 92% agreement, which is substantively high, as we would expect given the high level of sureness expressed by all three annotators. Agreement in this set is statistically distinct from the simulated null distribution, which is centered at .25, because all tweets in this ”all sure” subset received three ideology labels. This high-agreement subset accounts for 16% of the tweets in this set of 2000 labelled tweets, as illustrated in the pie chart at center. Examples of tweets matching this criterion are shown at right, truncated at 60 characters, and in most cases it is clear why annotators felt sure of the ideology labels they ascribed.

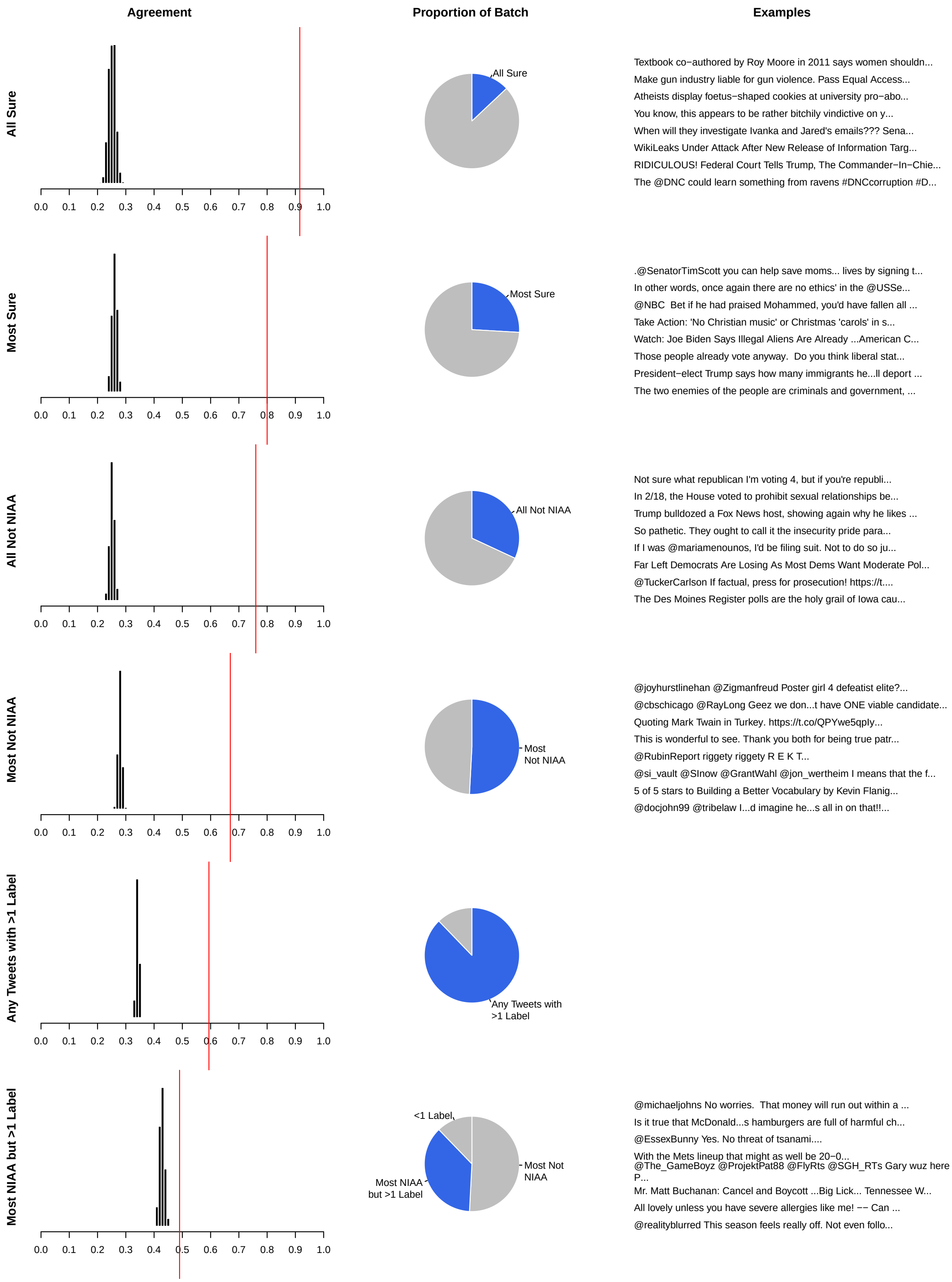


Figure 21: Inter-rater agreement on tweet ideology, subset by sureness

Subsequent rows reflect less restrictive sureness criteria: "Most sure" requires only that at least 2 of 3 annotators said they were "sure" of the ideology label they ascribed. This includes "all sure" tweets as well (hence it accounts for a larger portion of the total dataset), although the examples shown at right are chosen to exclude "all sure" tweets and therefore illustrate the somewhat greater ambiguity of tweets in this subset, reflected in the slightly lower agreement rate of 81%. "All not NIAA" includes tweets where all three annotators reported sureness greater than "No idea at all," and therefore reflects a generally looser standard of sureness: any combination of "sure" and "not so sure" is permitted. Note that "most sure" is not a strict subset of "All not NIAA" since the former admits at least one NIAA, however this set still shows less agreement overall. "Most not NIAA" represents a further loosening of this criterion, which includes 61% of the dataset, in which we observe 66% agreement. This agreement is still very distinct from the simulated null distribution, but observe that this distribution is no longer centered at .25: annotators are instructed to ascribe ideology labels, "unless you have no idea at all," and so the inclusion of tweets with at least one NIAA sureness implies the inclusion of many tweets with only two ideology labels (the third ideology label being NA), and this pulls the expected null agreement towards .5, which is the expected null agreement for two-fold binary annotation. Note also that as a larger proportion of the dataset is included in the analysis, the spread of the null distribution becomes tighter.

The bottom two rows of Figure 21 illustrate an unexpected annotation behavior. The annotation interface instructs annotators to select NIAA when they cannot guess an ideology label, and yet there are a substantial number of tweets that have >1 NIAA responses, and also >1 ideology labels, indicating that some workers are giving labels even when they say they have "no idea at all". Are these labels meaningful? The last row of the figure plots agreement for this subset of tweets: those with more than one ideology label that also have more than one NIAA response. The results indicate that there is a signal in these ideology labels, albeit a very weak one. An important topic for subsequent analysis is whether annotators are using the NIAA response differently - it is possible that some annotators' ideology labels under NIAA are more informative than others'. In the present analysis, we include ideology labels with NIAA sureness in the training set for the classifier, but down-weight them relative to "not so sure" and "sure" labels. The next section describes this procedure for processing labels into training data.

B.3 Processing Labels into Training Data

In order to transform triplicate labels of (binary -1/+1) ideology and (ordinal 0/1/2) sureness into ground truth labels with which to train a tweet classification model, we use a modified majority rule procedure.

To create ideology training data, we first multiply each worker's ideology response by a binary indicator for whether they indicated sureness greater than "No Idea At All" (NIAA). We do this because some workers gave ideology labels with sureness of NIAA, and we consider these labels too noisy to use to train a classifier. This procedure has the effect of converting low-confidence (NIAA) ideology labels to zeroes. Then, we reshape the column of triplicate ideology labels to a 9473-by-3 matrix, in which rows represent annotated tweets, and columns represent the three ideology codes for each tweet. We then use a majority-rule procedure to create training labels, by calculating row means for the matrix, and transforming negative means to -1 (liberal) and positive means to +1 (conservative). For tweets whose ideology labels average to zero (because workers indicated NIAA and/or nonzero ideology labels cancel each other out), this procedure results in an ideology label of 0 (neither liberal nor conservative).

To create sureness training data, a different modified majority rule is used: for each tweet, we count the number of sureness labels greater than 0 (NIAA), and if this sum equals 3, we designate this tweet as "Not all NIAA," and give it a binary sureness value of 1. If the sum is less than 3, this means that at least one worker said that they had "No idea at all" how to label ideology, and we give that tweet a binary sureness value of 0.

This procedure processes triplicate labels for ideology and sureness into single-valued ideology and sureness labels for each tweet, to be used in training a text classifier.

B.4 Classifier Training and Evaluation

In this document, we train and evaluate binomial Lasso models to classify tweets as liberal or conservative, and predict how sure a human coder would be about this ideology classification.

The training data for these models come from hand-coded tweets sampled from the set of all survey respondent tweets scraped, subset to exclude accounts that tweet (almost) exclusively promotional content, with sampling weights defined for each user inversely proportional to total tweets available per user (to avoid having the training set dominated by tweets from the most prolific users). A total of 9,473 tweets were annotated (threefold, by 3 different coders drawn from a set of 18 qualified workers) in a custom-designed interface, which asked coders to assign a liberal/conservative label to each tweet, and indicate a level of "sureness" regarding this label ("Sure", "Not so sure", "No idea at all").

First, we load in ground truth training data, previously processed from triplicate human coding for ideology and sureness. Next, we preprocess the text of tweets in the ground truth set, and save the feature names to allow corresponding featurization of tweets to be classified using these models. The preprocessing regime implemented below removes punctuation, numbers, and URLS. Bigrams are created based on stemmed features, and the feature names are saved to facilitate corresponding featurization of the classification set in subsequent analyses.

B.4.1 Ideology Classifier (Binomial Lasso)

To predict ideology expressed in tweets, we train a binomial Lasso model to predict whether a tweet author is expected to be Liberal (-1) or Conservative (+1). First, we select the binary liberal/conservative labels as the outcome of interest, drop excess liberal-labelled tweets to balance the training set, and drop non-ideological tweets (which were designated the value 0 during preprocessing). We subset the DFM to correspond to this selection procedure. We then train a cross-validated binomial Lasso model to predict these labels, with no intercept. To evaluate the model, we plot term coefficients in Figure 22. Text size is proportional to feature prevalence in the training corpus.

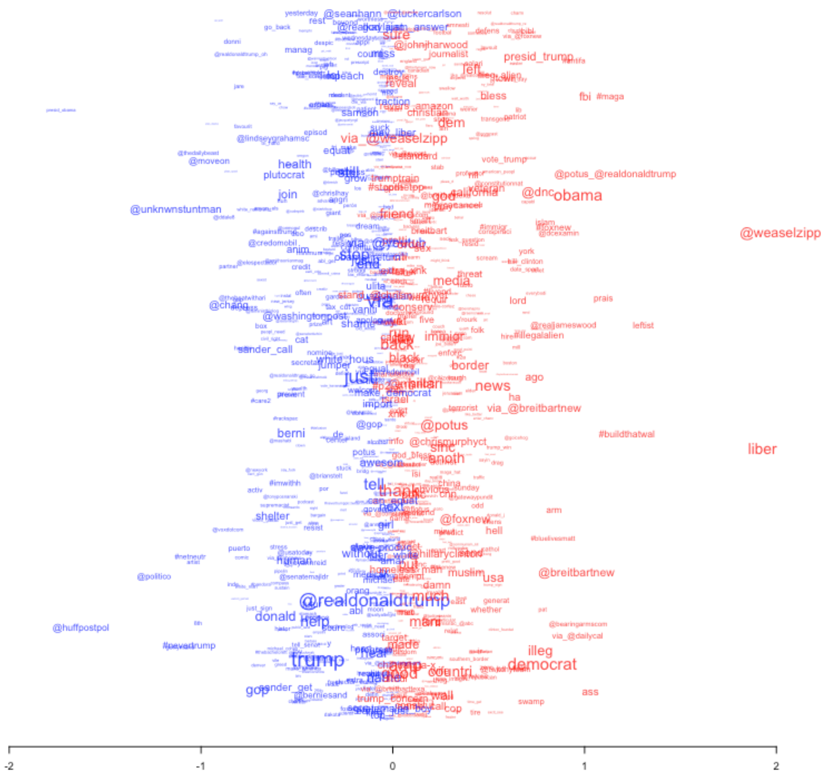


Figure 22: Term coefficients from liberal-conservative binomial lasso model

Since `cv.glmnet()` doesn't provide specificity or recall performance metrics, we retrain the Lasso on 3/4 of the data, and test performance on the remaining 1/4. Note that since the Lambda was chosen by CV using data included in the test set created below, this is not a *true* train-test evaluation. We create a test set, and generate predictions. The classifier shows a leftward misclassification bias:

- Conservative Precision: 66%
- Conservative Recall: 56%
- Liberal Precision: 61%
- Liberal Recall: 70%

		True	
		L	C
Predicted	L	35%	22%
	C	15%	28%

Table 5: Liberal-Conservative Confusion Matrix

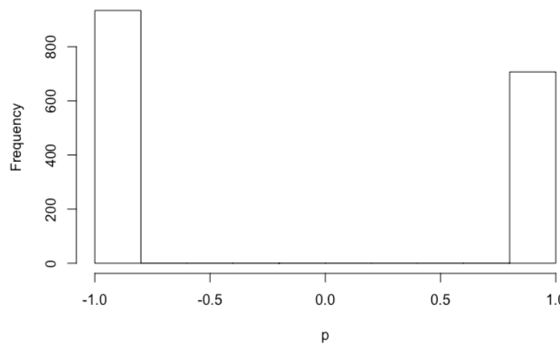


Figure 23: Histogram of test set predicted classes from ideology classifier.

B.4.2 Sureness Classifier (Binomial Lasso)

We train a binomial lasso model to predict "sureness", based on a training set of binary indicators for whether the majority of 3 human coders said they had "no idea at all" (NIAA) about the ideology of the author of a tweet (0), or whether a majority indicated more than "No Idea" level of sureness (1)

Term coefficients and full-data predicted response values are plotted. Note spammy text features with negative coefficients, political features with positive coefficients. Note that this model does include an intercept, and predicts "not sure" for most tweets. When applied to analyze user tweeting behavior, this tendency to classify as "not sure" pulls users' overall estimated expressed ideology towards the center. This may be relaxed in revisions.

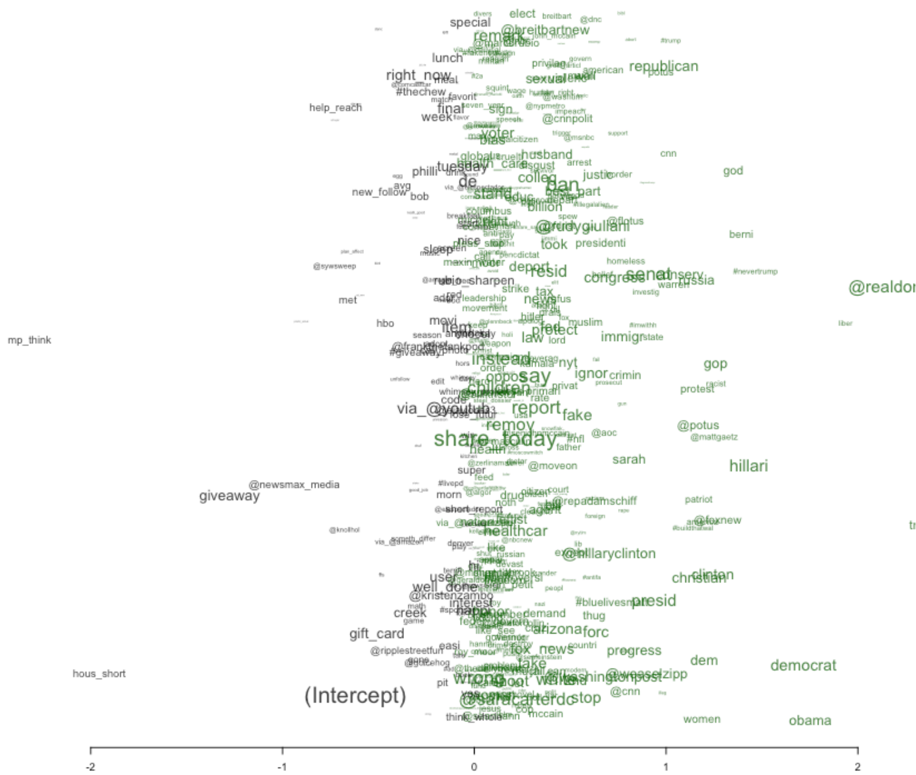


Figure 24: Term coefficients from sureness binomial lasso model

- Most-Not-NIAA Precision: 85%
- Most-Not-NIAA Recall: 62%

		True	
		Sure	Not Sure
Predicted	Sure	44%	19%
	Not Sure	5%	31%

Table 6: Sureness Confusion Matrix

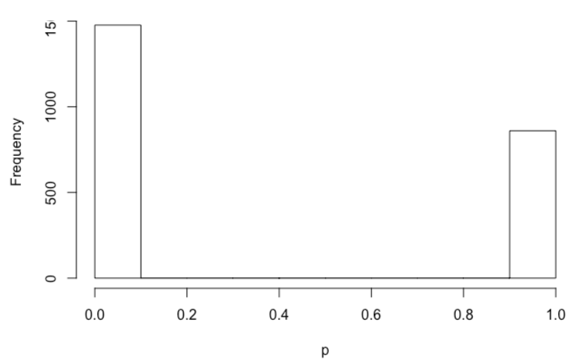
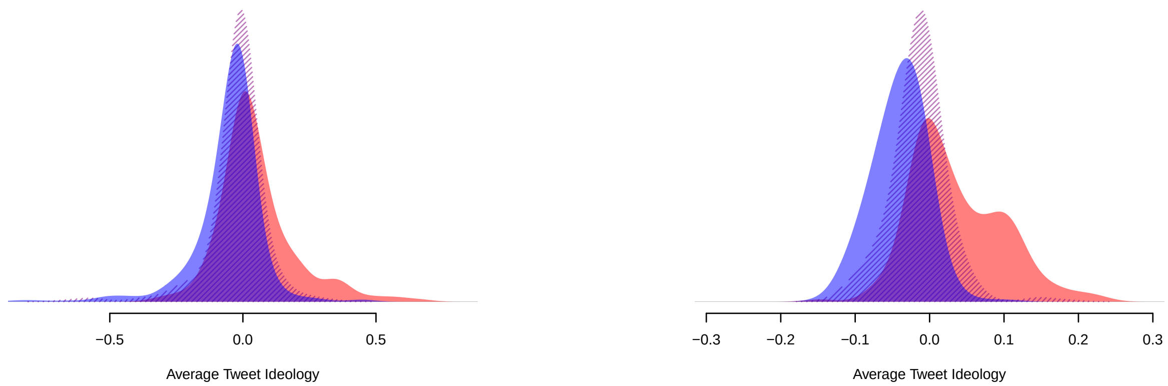


Figure 25: Histogram of test set predicted classes from sureness classifier.

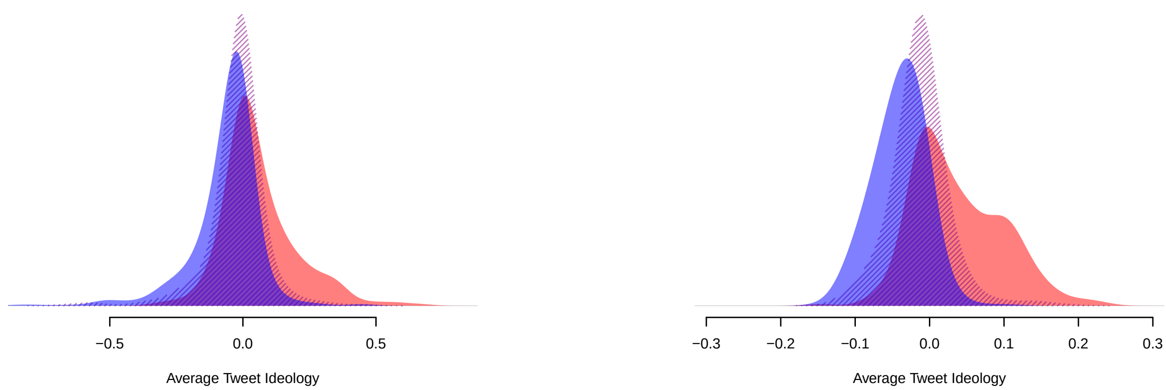
B.5 Face Validity of Measure



(a) Own Tweet Ideology

(b) Followee Tweet Ideology

Figure 26: Distribution of estimated tweet ideology for **KnowledgePanel** survey respondents' own tweets (panel a) and their followees' tweets (panel b), grouped by survey self-described ideology: Self-described liberals (blue), self-described conservatives (red), self-described moderates (crosshatched). Compare to YouGov-only distributions in Figure 4.

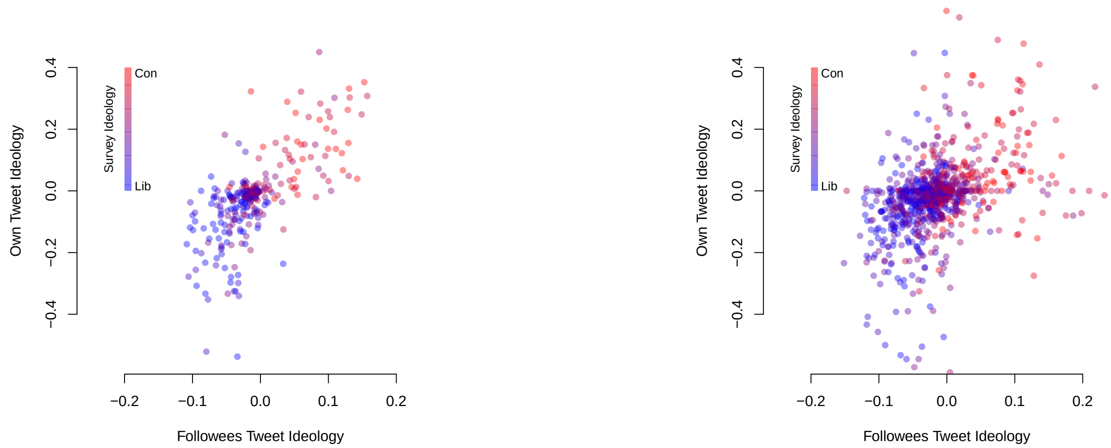


(a) Own Tweet Ideology

(b) Followee Tweet Ideology

Figure 27: Distribution of estimated tweet ideology for **pooled YouGov and KnowledgePanel** survey respondents' own tweets (panel a) and their followees' tweets (panel b), grouped by survey self-described ideology: Self-described liberals (blue), self-described conservatives (red), self-described moderates (crosshatched). Compare to YouGov-only distributions in Figure 4 and KnowledgePanel-only distributions in Figure 26.

C Non-Pooled YouGov and KnowledgePanel Analyses



(a) YouGov-only analysis

(b) KnowledgePanel-only analysis

Figure 28: Non-pooled YouGov (panel a) and KnowledgePanel (panel b) data comparisons of users' own tweet ideology (y axis), own survey ideology (color), and followees' tweet ideology (x axis)

Table 7: YouGov-only analysis of own tweet ideology

	Survey	Followees	Survey and Followees	Demographics
	(1)	(2)	(3)	(4)
Survey Ideo	0.112***		0.037***	0.044***
	(0.010)		(0.011)	(0.013)
Followee Tweet Ideo		1.761***	1.458***	1.447***
		(0.108)	(0.141)	(0.151)
Constant	−0.004	−0.007	−0.004	0.093
	(0.007)	(0.006)	(0.006)	(0.063)
Observations	287	287	287	268
R^2	0.312	0.482	0.501	0.540
Adjusted R^2	0.310	0.480	0.497	0.508
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: KnowledgePanel-only analysis of own tweet ideology

	Survey	Followees	Survey and Followees	Demographics
	(1)	(2)	(3)	(4)
Survey Ideo	0.069***		0.026***	0.028***
	(0.007)		(0.008)	(0.009)
Followee Tweet Ideo		0.983***	0.824***	0.830***
		(0.073)	(0.088)	(0.096)
Constant	−0.010**	−0.005	−0.004	0.026
	(0.004)	(0.004)	(0.004)	(0.035)
Observations	995	995	995	890
R^2	0.089	0.155	0.163	0.185
Adjusted R^2	0.089	0.154	0.162	0.164
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: YouGov-only analysis of own tweet ideology (survey ideology subgroups)

	Liberals			Moderates			Conservatives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Followee Tweet Ideo	1.378***		1.353***	1.673***		1.679***	1.434***		1.399***
	(0.309)		(0.312)	(0.252)		(0.265)	(0.185)		(0.188)
Survey Ideo		0.047	0.030		0.131	−0.006		0.106*	0.044
		(0.041)	(0.038)		(0.089)	(0.075)		(0.057)	(0.044)
Constant	−0.035**	−0.053	−0.013	−0.005	−0.015	−0.005	0.025**	0.010	−0.005
	(0.015)	(0.033)	(0.032)	(0.010)	(0.012)	(0.010)	(0.013)	(0.043)	(0.033)
Observations	133	133	133	74	74	74	80	80	80
R^2	0.132	0.010	0.136	0.379	0.029	0.379	0.436	0.043	0.443
Adjusted R^2	0.125	0.003	0.122	0.370	0.015	0.361	0.429	0.030	0.429

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10: KnowledgePanel-only analysis of own tweet ideology (survey ideology subgroups)

	Liberals			Moderates			Conservatives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Followee Tweet Ideo	0.969***		0.979***	0.698***		0.701***	0.761***		0.735***
	(0.169)		(0.171)	(0.146)		(0.153)	(0.151)		(0.153)
Survey Ideo		0.006	−0.013		0.052	−0.002		0.076*	0.045
		(0.027)	(0.026)		(0.039)	(0.040)		(0.042)	(0.041)
Constant	−0.015	−0.050***	−0.023	−0.012**	−0.021***	−0.012**	0.023**	−0.00003	−0.006
	(0.009)	(0.019)	(0.019)	(0.006)	(0.005)	(0.006)	(0.011)	(0.030)	(0.029)
Observations	421	421	421	358	358	358	216	216	216
R^2	0.073	0.0001	0.073	0.060	0.005	0.060	0.106	0.015	0.111
Adjusted R^2	0.070	−0.002	0.069	0.058	0.002	0.055	0.102	0.010	0.103

Note: *p<0.1; **p<0.05; ***p<0.01