Follow Your Ideology: Measuring Media Ideology on Social Networks

August 18, 2016

Pablo Barberá[*] and Gaurav Sood[†]

[*] Pablo Barberá is a Moore Fellow at NYU. Pablo can be reached at pablo.barbera@nyu.edu

[†] Gaurav can be reached at gsood07@gmail.com

Abstract

Analyses of causes and consequences of what is on the news require accurate measures of various features of news content. In this paper, we propose a new technique for measuring variation on the politically salient dimension of ideology. We exploit the fact that politically interested people on online social networks tend to follow ideologically proximate news sources. Using the method, we estimate the ideological positions of over 2,300 media sources. We validate the method by comparing estimates from our method with measures obtained using two established content analytic methods. Finally, we illustrate the utility of the new measures by using them to study two important issues. First, we investigate whether journalists' personal ideological beliefs, measured using campaign contributions, affect the ideological slant of the content they produce; the correlation is positive and sizable. Second, we study intra-media ideological heterogeneity; we find that 'moderate' outlets carry ideologically diverse content.

What explains the content of news? What impact do the news media have on people and politicians? These are important questions in the study of politics. Objective measures of politically salient dimensions of a large number of news media sources promise to greatly aid the study of these questions. In this paper, we develop a new method for measuring variation on the politically salient dimension of ideology.

Hitherto, scholars have primarily relied on content analysis to estimate the ideological location of news media sources (Groseclose and Milyo 2005; Gentzkow and Shapiro 2010; Puglisi and Snyder 2011). Here we present a new way of quantifying ideology. We exploit the fact that politically interested users on online social networks tend to *follow* ideologically proximate news sources on these networks to infer the ideological location of news media sources from the list of their *followers* on one such social media site. The technique allows us to simultaneously infer the ideology of news media sources, citizens, and political actors on the same scale.

We assess the validity of the new measures in two different ways. First, we check how well our estimates correlate with available scores from an established technique. Second, we estimate ideological location of nearly 2,000 news sources using another established content analytic method, and assess the extent to which it covaries with our estimates. Both sets of results suggest that measures obtained using our method are valid. We also show that the estimates obtained using our approach are more reliable than those obtained using a popular content analytic method.

To illustrate potential uses of our new finer-grained estimates of media ideology, we present two applications. A great deal of survey evidence shows that only a few journalists identify as conservatives (Kohut 2004, 2008; Weaver et al. 2009). These figures are regularly touted on the right as sufficient evidence of bias. However, journalists reject such inferences, arguing that their political beliefs need not (and generally do not) affect how they cover the news. In our first application, we investigate how (private) ideology of journalists, estimated using political donation records, correlates with the ideology of the content they produce. We find a sizable positive correlation.

In our second application, we assess the extent to which content produced by 'moderate' outlets is non-ideological, as opposed to ideologically diverse. Contrary to popular view, we find that 'moderate' outlets carry a fair bit of ideological content, albeit the content tends to be diverse. More generally, we find that intra-media heterogeneity is inversely correlated with slant; more moderate outlets tend to carry more ideologically diverse content. The finding has important implications for the study of selective exposure.

Till now, most studies of selective exposure (see for example, Arceneaux, Johnson and Murphy 2012; Iyengar and Hahn 2009; Stroud 2008; Gentzkow and Shapiro 2010) have imputed preferences based on ideology and visitation measures at the media outlet level. Our results suggest that such ecological inference carries grave risks. For instance, a person visiting *The New York Times* may only read content that is more liberal than the mean ideology of the outlet. Any such behavior is liable to bias inferences about preferences that either assume that content within an outlet is homogeneous or that people do not self-select within outlets.

1. Learning About Ideology Using Social Networks

Online social networks provide a rich set of informational cues from which to infer traits such as ideology about the individuals who have an active presence on these websites. For one, we have access to what they say publicly on social networks, including what they *like*. For another, we get to observe which users' generated content is shared by which other user; we can also observe what is shared. And lastly, we know whom the users *follow* or are *friends* with. These behaviors are most easily observed when most communication is public. And most richly observed, when the network is successful. Thus, we focus our efforts on the most successful such network, Twitter.

Perhaps the most obvious way to learn about ideological slant of a media 'source' is to analyze its content. But how? If most news media took explicit positions on the issues it were covering, one could scale news content based on the positions it was advocating. Most news stories, however, do not take explicit positions on the issues.[2] Hence, most attempts to scale ideology using

[2] Editorial pages of newspapers are a notable exception. See Ho, Quinn et al. (2008); Habel (2012) who estimate

content analyses rely on more indirect methods, with harder to justify identification assumptions. One popular strategy for ideological scaling using automated content analyses relies on supervised learning —it scales based on the extent to which usage of words in news stories matches usage of words by politicians in Congressional speeches (Groseclose and Milyo 2005; Gentzkow and Shapiro 2010) or books by certain other 'labeled' actors (Sim et al. 2013). In practice, this can be quite risky. For example, extended quotes presented for the exact opposite purpose — for e.g., mocking the politician— can ramp up similarity. And similarities in language can be due to similarities in agendas, not positions on those agendas (Sood and Guess 2015; Shaw and Sparrow 1999).[3] Outside of these concerns, it is not clear why one would make the task of learning about ideology harder by relying on a small set of concise 140-character tweets than a potentially much larger set built from (mostly) public media data.

What (and whose) content users share is another way to learn about ideology. Relying on what users share reduces to content analysis of a small set of tweets. This should outperform content analysis of the larger set of data if retweets are particularly high in ideological content. We have no such expectation. Famously, many Twitter users display the pithy warning "RT $\neq$ Endorsement", which means retweets do not imply endorsement. Expectedly, an analysis of retweeting behavior suggests that retweets are not particularly ideologically motivated (Morgan, Lampe and Shafiq 2013).[4] Thus relying on whose content users share to scale ideology is liable to be error-prone.

Lastly, one may learn about ideology from users' decision to 'follow'; to follow someone on an online social network means subscribing to all their public messages. Due to a variety of reasons,

ideology of editorial pages by tallying explicit stances on issues. Ideology of the editorial page can be, however, an unreliable predictor of ideology of the content of rest of the newspaper. For instance, as we discuss later, we find (like others have before us) that Wall Street Journal's editorial page is notably more conservative than the rest of the paper.

[3] This isn't to say that agendas do not reveal partisanship. Often times they do. But there is no necessary relationship between agendas and positions on those agendas. And threat to inference is considerable. For instance, during the latter stages of Iraq War, military may still have been at the top of the agenda, but Democrats and Republicans espoused very positions on it.

[4] One potential reason for this may be that even when retweets connote ideological intent, it isn't easy to discover the ideology being supported. For instance, many people retweet for ironic purposes, or to highlight inflammatory rhetoric of the 'other side.'

which we expand on below, decisions to *follow* tend to be a particularly rich source of information about ideological positions. A decision to follow entails consideration of at least two kinds of costs —1) opportunity costs: since time (one can devote to consuming political information) is a finite resource, choosing to follow a source often means reducing exposure to other users' messages, and 2) psychic costs: messages that are uncongenial to the person's existing political beliefs cause psychic discomfort (see Festinger 1962).

Following media sources on social media entail significant opportunity costs just because most media sources post frequently. The costs are especially material given the rapid decay of visibility on social media (Oken Hodas and Lerman 2012). Even when opportunity costs are lower, as is the case with media consumption more generally, people tend to (weakly) prefer news that aligns with their existing ideological views (Gentzkow and Shapiro 2005, 2010; Iyengar and Hahn 2009; Stroud 2008; Lazarsfeld, Berelson and Gaudet 1944; Bryant and Miron 2004). At least part of the decision to consume congenial information is driven by the perception that congenial sources are more 'trustworthy' and 'fair' (Arceneaux, Johnson and Murphy 2012). In all, it reasons then that decisions to follow news media accounts are partly ideological.

Building on these insights, An et al. (2012) scale measures of closeness between media sources based on their common audience onto an ideological dimension. However, they limit their analysis to just 24 news outlets. And some of their results seem to lack external validity. For instance, they write, "*Washington Post* and *Washington Times*, known to have conflicting political preferences, lined up close to each other." Their method also places *The Washington Times* to the right of Fox News. The challenge is therefore twofold —to improve the method, and to extend it to thousands of journalists and media outlets.

We propose two sets of improvements. One is about the data from which to learn. Our conjecture is that both opportunity and psychic costs are particularly high for the politically interested. Ample research supports the conjecture that the politically interested take greater account of partisanship in their decisions to consume political information (see for example, Iyengar and Hahn 2009; Hindman 2008). It is also less likely that other concerns enter into their decisions to follow political sources. Hence, one way to learn better from following decisions is to subset on the most informative following decisions, that is, following decisions of the politically interested (Maestas, Buttice and Stone 2014). Our second improvement comes from a statistical model of following decisions that discounts the popularity of journalists.

2. A Spatial Model of Following Behavior

Suppose that each person $i \in \{1, \ldots, n\}$ is offered a choice to follow or not follow a target user $j \in \{1, \ldots, m\}$, where j is a news media source (program, outlet or a journalist). Let $y_{ij} = 1$ if user i decides to follow news media source j, and $y_{ij} = 0$ otherwise. And let the n by m matrix $\mathbf{Y}$ denote the matrix that aggregates all the individual following decisions (the matrix is called an "adjacency matrix" in social network analysis).

Following research that suggests that preferences of both politicians and the mass public are well-explained by a single dimension (see for example, Poole and Rosenthal 2007; Clinton, Jackman and Rivers 2004; Jessee 2009; Tausanovitch and Warshaw 2013.), we assume ideology to be a uni-dimensional construct. In line with research that we cite above, we also assume that people get greater utility from following an ideologically proximate media outlet than from following a more ideologically distal media outlet. For mathematically tractability, with little consequence for our overall results, we further assume that the utility that the person derives from following an outlet declines as a quadratic function of Euclidean ideological distance between the person and the outlet. Under these assumptions, the decision by user i to follow media outlet j has the following functional form: $-\gamma||\theta_i - \phi_j||^2$, where $\theta_i \in \mathbb{R}$ is the ideal point of Twitter user i, $\phi_j \in \mathbb{R}$ is the ideal point of media outlet j, and γ is a normalizing constant. This core model is equivalent to the core of spatial voting models (Enelow and Hinich 1984; Jessee 2009; Poole and Rosenthal 2007; Clinton, Jackman and Rivers 2004).

We make two additions to this model. To account for baseline differences in popularity of media outlets to do with non ideological factors such as geographic area that the media source is serving (national versus local), we specify a media source specific parameter, α_j, that captures the baseline probability of following media outlet j. To account for idiosyncratic non-ideological differences across users in perceived costs of following an account, such as differences in time available to follow politics, we specify a user-specific parameter, β_i.

In all, we assume that the objective function that user i maximizes when choosing the set of media outlets and journalists to follow is:

$$\underset{y_{1,\ldots,J}}{\arg\max} \left[\sum_{j=1}^{J} \alpha_j(y_j) - \beta_i(y_j) - y_j(\gamma||\theta_i - \phi_j||^2) \right] \tag{1}$$

This final model is similar to the Bayesian model developed by Barberá (2015) to estimate ideal points of legislators and voters based on their Twitter networks, though also see the structural model developed in Gentzkow and Shapiro (2011). Although we adopt a similar model, the size of the "follower" network of media outlets requires a different estimation strategy, capable of scaling a large adjacency matrix. Thus, rather than estimate the spatial model directly, we instead use correspondence analysis (Greenacre 1984, 2010), which approximates the maximum likelihood solution for a one-dimensional spatial model (ter Braak 1985). (See SI 1 for more details about the estimation procedure. As we describe in SI 1, Correspondence Analysis yields estimates that are very highly correlated (over .95) with estimates from a spatial model at a much lower computational cost.)

3. Data

We constructed a list of journalists, news programs, and news outlets with profiles on Twitter using various sources, including websites of major news media outlets, Twitter lists curated by media outlets, compendia of journalists on Twitter, and the list of media websites considered by Gentzkow and Shapiro (2011). To guard against idiosyncratic factors having a sizable impact on our estimates, we only chose journalists, programs, and news outlet profiles with more than 2,000 followers. This yielded a list of m = 2,363 news sources and included, among others, media outlets such as `@cnnbrk`, `@nytimes`, `@TIME`, `@WSJ`, `@CBSNews`, `@washingtonpost`, and

@HuffingtonPost, and prominent journalists such as @AndersonCooper, @maddow, @NickKristof, @BarbaraJWalters, and @seanhannity. The list also included hosts of late night and satirical news shows, such as @ConanOBrien, @StephenAtHome, and @jimmyfallon.[5] Due to reasons highlighted in the previous section, we also added all members of the current US Congress with a Twitter account (obtained using the *New York Times* Congress API), and Mr. Barack Obama to the list of target users. In all, our final list had $m =$ 2,769 accounts.

Next, using the Twitter REST API, we downloaded the entire list of followers for all the m users (as of May 1st, 2014). This gave us a list of $n =$ 72,259,123 users who followed at least one media source. However, an extremely high proportion of these users were either inactive, spam bots or resided outside US. To overcome this problem, we discarded users who 1) followed fewer than 5 journalists or media outlets, and 2) were located outside the US.[6] This reduced our sample to 11,607,284 users. To further zero in on the politically interested, we restricted our sample to users who followed at least three national politicians.[7] The final sample size is $n =$ 4,140,572 users.

4. Estimates of Media Ideology

Figure 1 displays ideology estimates for some of the major media outlets along with those of the median Republican and Democrat in each chamber of the US Congress, and the median follower in our sample. Among these major media outlets, NPR and MSNBC are the most liberal,

[5] While the primary focus of some late night and satirical news shows is apolitical, these shows often interview political figures, and not infrequently comment on political affairs. Accordingly some research suggests that these shows are an important source of political information (in some case the *only* source) for many people, particularly young adults. For example, a Pew Research Center survey on media consumption found that 7% of Americans regularly watched the Daily Show with Jon Stewart in 2010. Among young adults (ages 18 to 29), this percentage was 13% – greater than the 11% who watched NPR, and similar to the 13% who watched CNN, and 14% who watched the network evening newscasts.

[6] We considered anyone who sent a geo-located tweet from outside US between November 1st and November 30th, 2013 as located outside the US. This represents a total of approximately 7 million users, of whom 250,000 were included in our initial sample.

[7] Our list of politicians includes all members of the current US Congress with a Twitter account and President Barack Obama. Their total number of unique followers is 60,130,443, of whom 6,055,779 follow at least three accounts.

and Fox News and The Drudge Report, the most conservative. While most outlets are located near the center of the ideological space, consistent with Groseclose and Milyo (2005), we find that many of the major news outlets are slightly more liberal than the median follower. Given that the average Twitter user is *more liberal* than the average US adult [8], media outlets to the left of the median follower are also to the left of the median US adult. The other notable feature about these major media outlets is that, except for a few outlets, they typically have more moderate positions than the median legislator in each party and chamber.

Figure 1: Estimates of Media Ideology for Main Outlets and Politicians

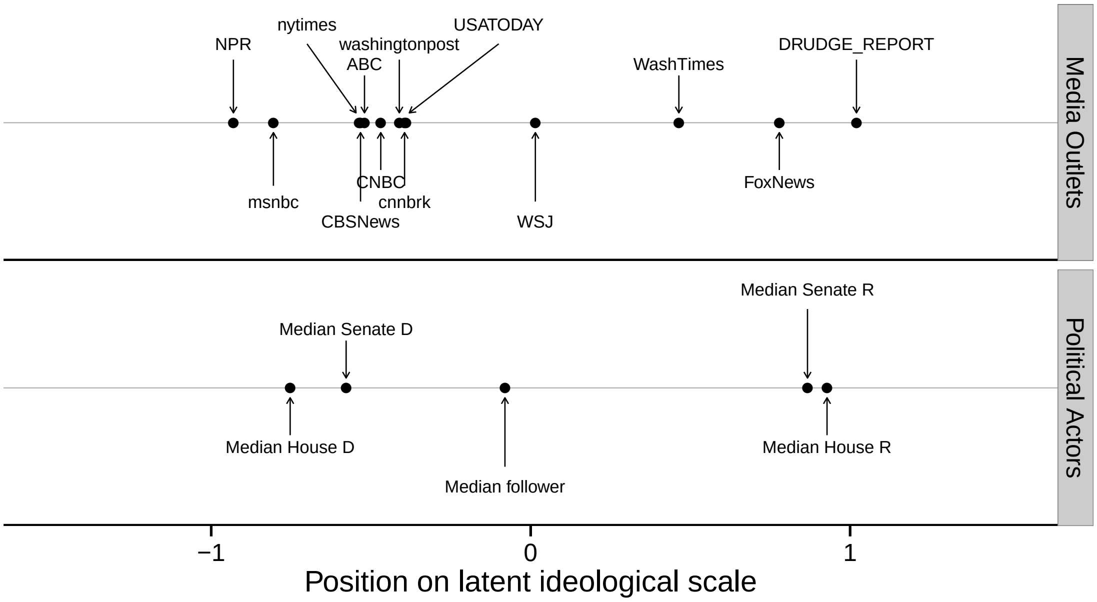
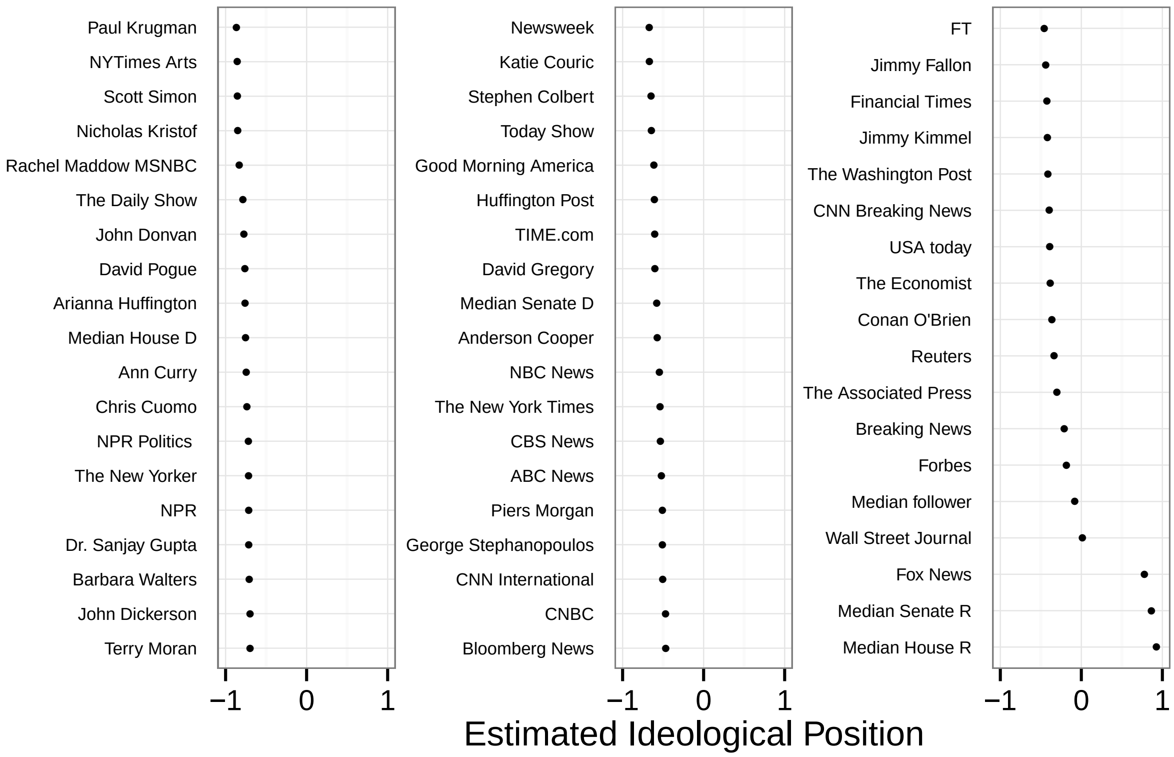


Figure 2 displays estimates for a wider set of journalists and media outlets with a large number of followers (over 1,000,000). Their relative positions are consistent with popular perceptions of their ideology. Paul Krugman and Nicholas Kristof, columnists with the *New York Times*, and Scott Simon, a contributor to NPR were be the most popular liberal journalists in our data, while Fox News was the only conservative media source with over one million followers. The political

[8] According to the Pew Research Center Poll on Biennial Media Consumption (June 2012), the average ideology of politically interested Twitter user on a 5 point semantic conservative (1) to liberal (5) scale is 3.00, while for the general population is 3.16. (We defined politically interested users those who report having followed “very” or “fairly” closely the 2012 Presidential election campaign. The difference is modest but statistically significant ($p < .01$).

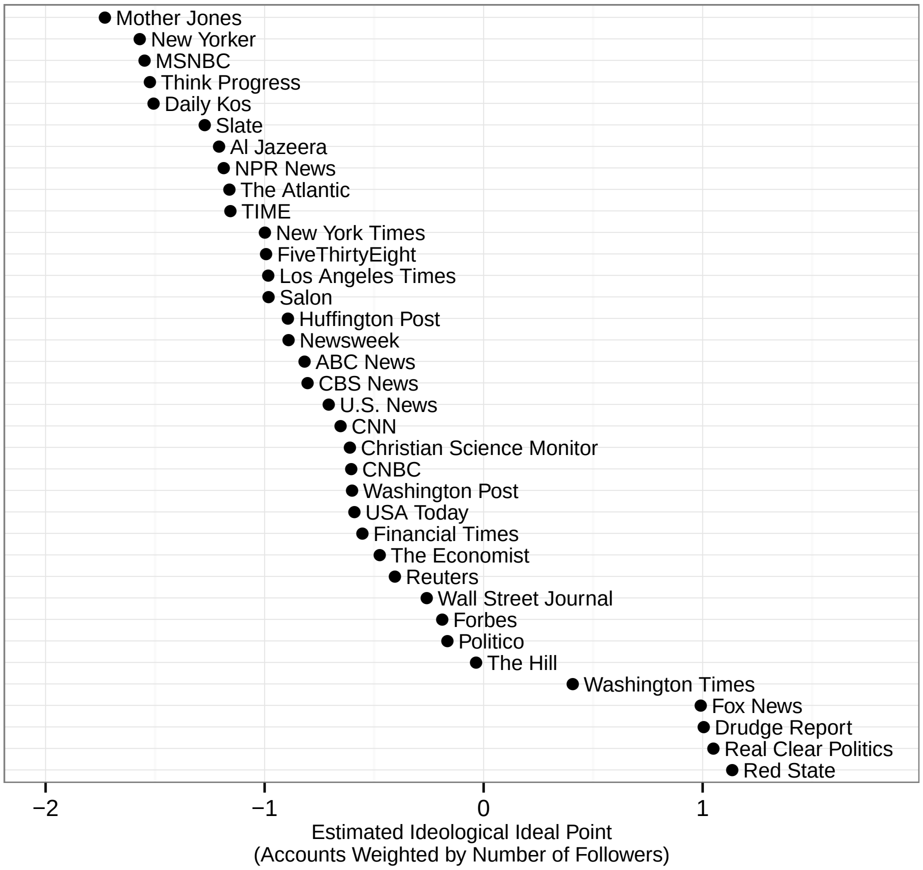
center is occupied by journalists working for the CNN and the three major networks, journalists such as Anderson Cooper, George Stephanopoulos, and David Gregory.

Figure 2: Estimates of Media Ideology for Popular Outlets and Journalists



We find similar estimates regardless of how we compute the ideological position of media outlets. While Figure 1 displayed our results for the main Twitter account of each media outlet, in Figure 3 we computed their ideal points as an average of the ideology estimates of all the journalists affiliated to each outlet, and their shows or sections. Each estimate was weighted by their number of followers, to account for the different popularity or visibility of each Twitter account associated to an outlet. The distribution we observe is essentially identical: on the left, *Mother Jones*, *New Yorker*, MSNBC, *Think Progress*, and *Daily Kos* are the most liberal outlets; on the right, *The Washington Times*, Fox News, *Drudge Report*, *Real Clear Politics*, and *Red State* are the most conservative outlets.

Figure 3: Weighted Estimates of Media Ideology (All Outlets)



5. Assessing the Validity of the Measures

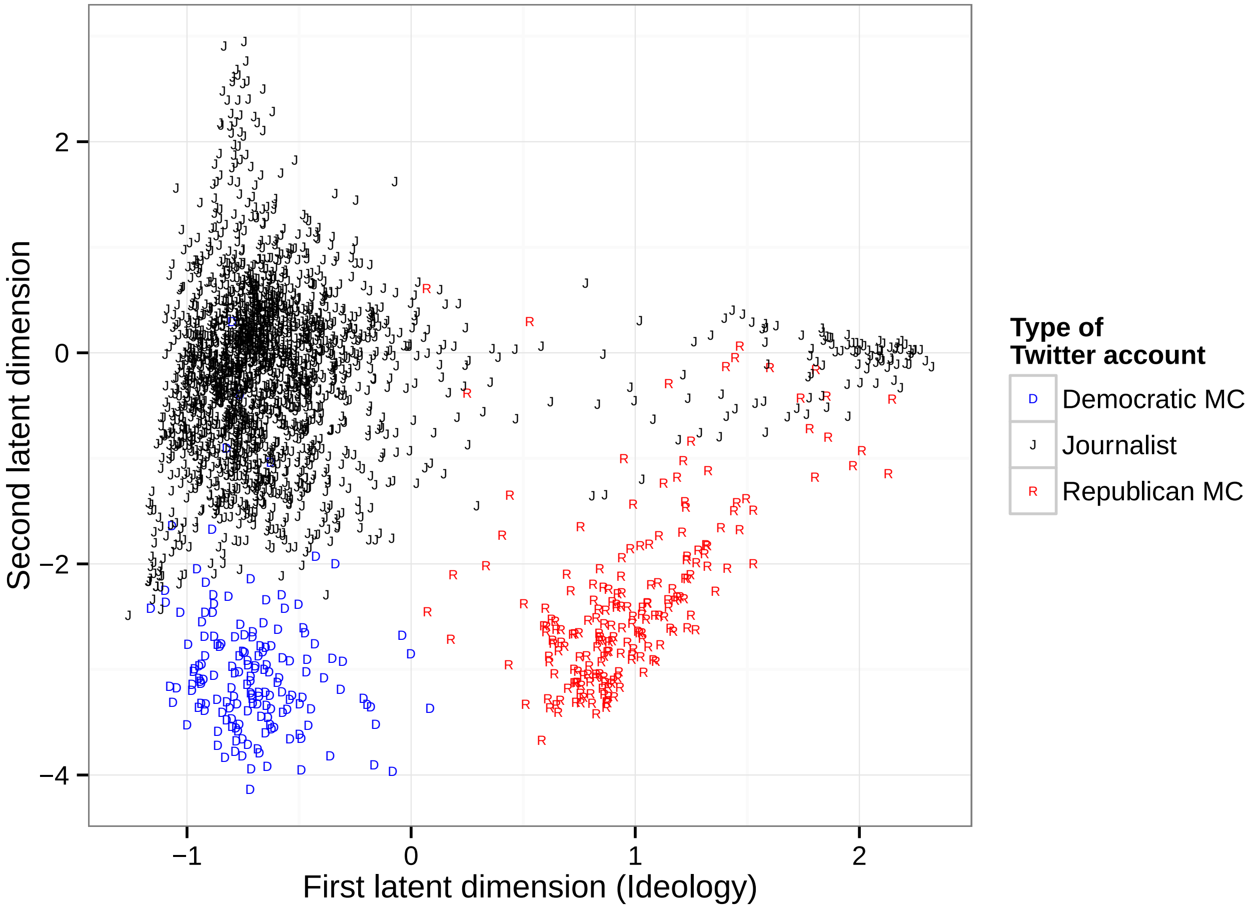
We now turn to examining the validity of our ideology estimates for media sources. First, we discuss the construct validity of our measures. In order to assign a substantive interpretation to the latent dimensions that emerge after correspondence analysis, we examine the locations of journalists and members of Congress on the first two dimensions. Figure 4 clearly suggests that the first dimension correlates heavily with political ideology, as it is usually conceptualized in the American Politics literature. House and Senate Democrats are located on the left of the ideological spectrum, while Republicans are on the right. The correlation between our estimates for members of Congress and an item-response model based on their roll-call votes (Clinton, Jackman and Rivers 2004; Jackman 2014) is $r = .944$.

Figure 4 shows a small group of journalists and legislators that occupy similar positions in the ideological space, on the center right of the panel. This group includes media outlets like Fox News Radio, media personalities like Megyn Kelly, O'Reilly Factor, and Sean Hannity, and legislators like Michele Bachmann, Ted Cruz, Rand Paul, and Mike Lee, as well as other legislators associated with the Tea Party. The fact that this group is estimated to the right of the rest of the Republican Party on this dimension can be seen as additional evidence that the latent dimension is indeed political ideology.

The second latent dimension, on the other hand, seems to capture the extent to which the person or outlet was related to politics. Legislators of both parties score low on this dimension. While journalists who extensively cover politics, journalists like Chuck Todd (-.40), John King (-.26), Fareed Zakaria (-.16), Rachel Maddow (.01), and political satire shows like Jon Stewart's The Daily Show (-.46) have scores in the middle range on the dimension. Journalists who only occasionally cover politics —Late-night show hosts, such as Conan O'Brien (.58), Jimmy Fallon (.66), Jay Leno (.67), Jimmy Kimmel (.78), and journalists such as Barbara Walters (1.48), Jim Cramer (1.51), Frank Bruni (1.62) or Wall Mosberg (1.70) —have high values on this dimension.

Second, we test our measures' convergent validity by measuring correlation with an existing

Figure 4: Distribution of Journalists, Outlets, and Politicians (First Two Dimensions)



measure of media ideology, and with estimates from text scaling techniques applied to tweets. In the following section, we also show that journalists' ideological positions are clustered within each outlet, which further demonstrates their internal validity.

To assess the convergent validity of our estimates, we compare our measures to estimates obtained using alternate methods. In total, we validate our method against five different measures: a) Wordscore method applied to tweets, b) A similar formal model of audience behavior estimated using a very different dataset Gentzkow and Shapiro (2011), and c) measures obtained from analysis based on citations to think tanks (Groseclose and Milyo 2005). In the main text, we limit our focus to comparison with Wordscore, but discuss results from other methods at the end, presenting details of each of the methods in the appendix (see SI 2).

5.1. Comparison with Tweet WordScores

We collected 3,200 most recent tweets sent by all the journalists in our sample, as well as all Members of the current US Congress with a Twitter account, from the REST API.[9] Then, following a similar approach as in Toff and Kim (2013), we used *Wordscores* (Laver, Benoit and Garry 2003) to scale journalists and media outlets on the same scale as Members of Congress. The tweets sent by this last set of users represented our "reference texts", and were assigned a position based on their ideological position, estimated from roll-call voting records and applying an item-response model (Clinton, Jackman and Rivers 2004; Jackman 2014). Table 1 displays the top scoring words on each each extreme of the ideological scale, demonstrating that the latent dimension does indeed correspond to ideology. Figure 5 compares our network-based estimates with those obtained using the *Wordscores* method applied to tweet text, for our sample of 2,170 journalists and media outlets. The correlation between the two measures is high, $r = .45$.

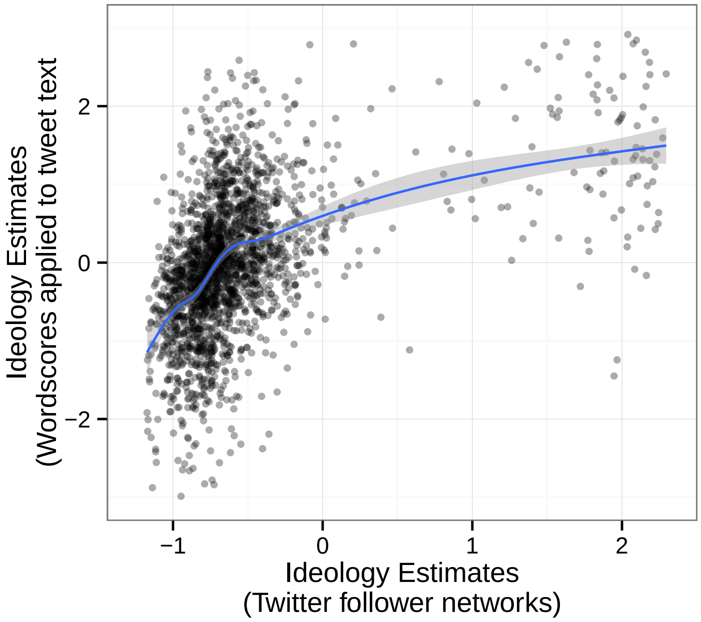
Table 1: Top 30 scoring words predictive of ideology, according to *Wordscores* method

Top 30 liberal words			Top 30 conservative words		
women	ACA	proud	obamacare	tcot	jobs
work	help	must	spending	obama	IRS
families	need	renewUI	budget	house	will
health	end	education	hearing	debt	live
join	raisethewage	rights	watch	video	tune
act	let's	violence	obama's	president	energy
immigration	workers	students	morning	benghazi	washington
community	benefits	women's	listen	tax	read
million	wage	student	GOPleader	president's	HouseCommerce
support	equal	make	taxes	idpol	icymi

To compare the performance of these two methods, we focus on a sample of Twitter users with known ideology: members of the US Congress. First, we split our sample of legislators into two halves; tweets from one half serve as our "reference texts", and are assigned ideal points based on roll-call votes (Jackman 2014), and tweets from the other half serve as our "virgin texts", whose ideology is estimated using *Wordscores*. The correlation between the ideal points and the

[9] Due to the rate limits of the Twitter API, only the 3,200 most recent tweets from any user are available.

Figure 5: Comparing Twitter-Based Ideology Estimates Based on Followers Networks and on Text (Wordscores method)



Wordscores estimates was $r = .898 (N = 255)$. However, the correlation of roll-call based ideal points with social network based estimates was higher still, $r = .944$.

Our analysis here underscores the difficulty in scaling ideological based on text. Among the top scoring words in each dimension, we find "education", "students", "taxes" or "IRS", associated to different important political issues. Mentioning these words can be informative about legislators' ideology, which will tend to emphasize those topics they "own" (Petrocik 1996; Egan 2013) in their effort to influence the political agenda in their benefit. However, that's not necessarily the case for journalists, who may discuss these topics simply because they're covering political news events. In all, we have shown that our approach yields estimates that are valid, and more reliable than those based on content analysis.

To compare the performance of these two methods, we focus on a sample of Twitter users with known ideology: members of the US Congress. The correlation between the ideal points and Text based measures was $r = XXX$. However, the correlation of roll-call based ideal points with social network based estimates was considerably higher, $r = .944$.

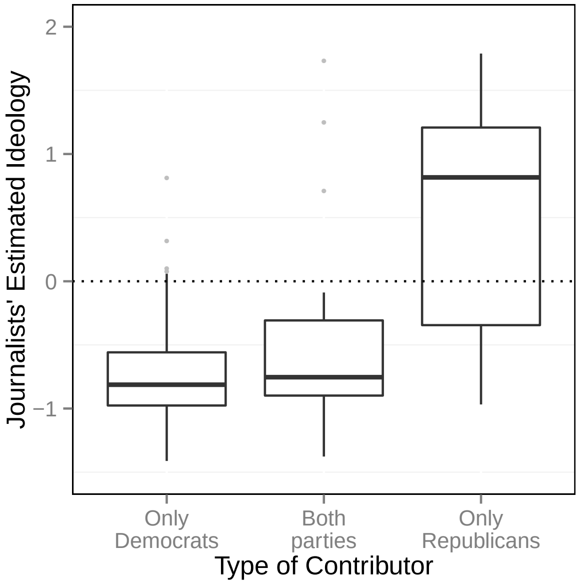
Our analysis here underscores the difficulty in scaling ideological based on text. Among the top scoring words in each dimension, we find “education”, “students”, “taxes” or “IRS”, associated to different important political issues. Mentioning these words can be informative about legislators’ ideology, which will tend to emphasize those topics they “own” (Petrocik 1996; Egan 2013) in their effort to influence the political agenda in their benefit. However, that’s not necessarily the case for journalists, who may discuss these topics simply because they’re covering political news events. In all, we have shown that our approach yields estimates that are valid, and more reliable than those based on content analysis.

6. Private Political Beliefs and Ideological Slant of Content

Our first application examines whether journalists’ campaign contributions are correlated with social-network based estimates of the ideology of the content they produce. To examine the relationship, we first compiled a list of all journalists (and people who reported working in the media) who had contributed to campaigns from the Database on Ideology, Money in Politics and Elections (Bonica 2013*a*) and the Dataset on Media Donations by the Center for Responsive Politics. Next, we searched for the Twitter profiles of journalists who had donated; we were able to match 306 journalists. Since only three of the 306 journalists were in our initial sample — the other 303 journalists either had less than 1,000 followers, or were affiliated with regional media outlets — we downloaded their follower networks on Twitter, and added them as supplementary points to the matrix of following decisions. As before, we projected these supplementary points onto a low-dimensional space to compute their ideology scores. (See sections on Model and Estimation for more details.)

Figure 6 summarizes our results. It displays our ideology estimates for journalists who donated only to Democratic candidates (236), only to Republican candidates (40) or to both (30). As expected, the median ideology score for the first group is more liberal than that of the second group, with journalists giving to both parties in the middle. Our measures of journalists’ ideology are also highly correlated ($r = .73$, $N = 250$) with their Campaign Finance Scores (Bonica 2013*b*),

Figure 6: Comparing Journalists' Ideology Estimates with Campaign Contributions



estimated based on what specific candidates' campaigns they contributed to. Thus, contrary to journalists' claims, our analysis demonstrates that journalists' private political beliefs (measured through their campaign contributions) are correlated with ideology of the content they produce.

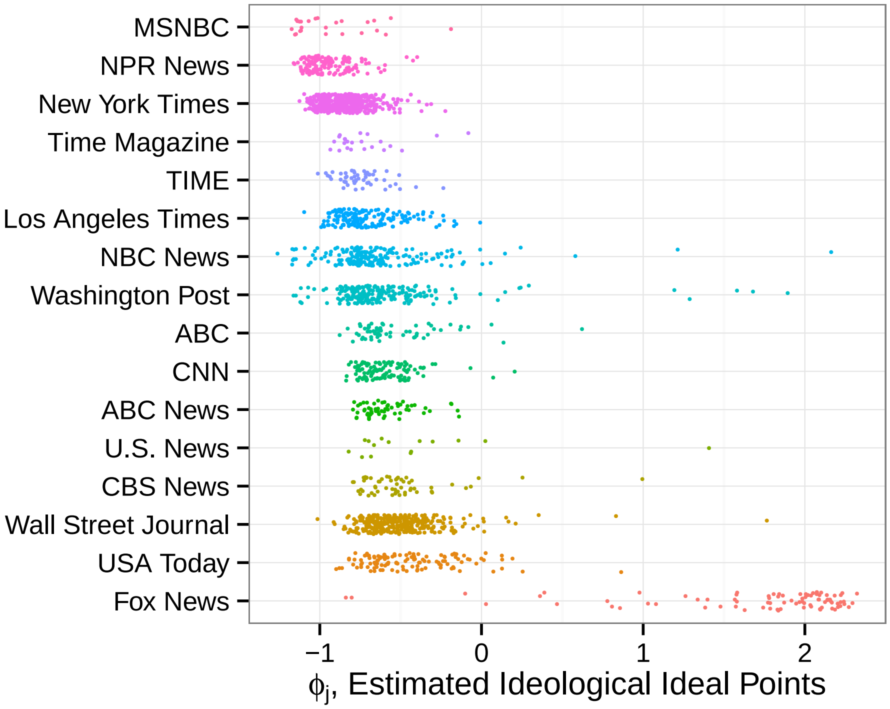
7. Intra-media Ideological Heterogeneity and Selective Exposure

We now examine ideological heterogeneity within outlets. Either due to sorting, or editorial policy (and attendant pressures to comply), we expect journalists affiliated with the same outlet to have ideal points clustered around the ideological location of the outlet. As figure 7 shows, that this is indeed the case.[10] While there is some overlap across media outlets, journalists that belong to MSNBC, NPR, and the *New York Times* are likely to be more liberal than journalists in ABC, CBS, *Wall Street Journal*, *USA Today*, and particularly Fox News. The plot also shows clear outliers; most of them being occasional contributors or former employees now at different outlets. For example, the two most conservative journalists working for NBC News are Dave Briggs (formerly Fox News) and Jenna Bush. Similarly, the most liberal journalists working for

[10] We only show here estimates for media outlets with more than 10 journalists with a Twitter account with more than 1,000 followers.

FOX News are contributors like Sally Kohn (now at CNN) and Mary C. Curtis (recently hired by the *Washington Post*).

Figure 7: Distribution of Ideology Estimates, by Media Outlet

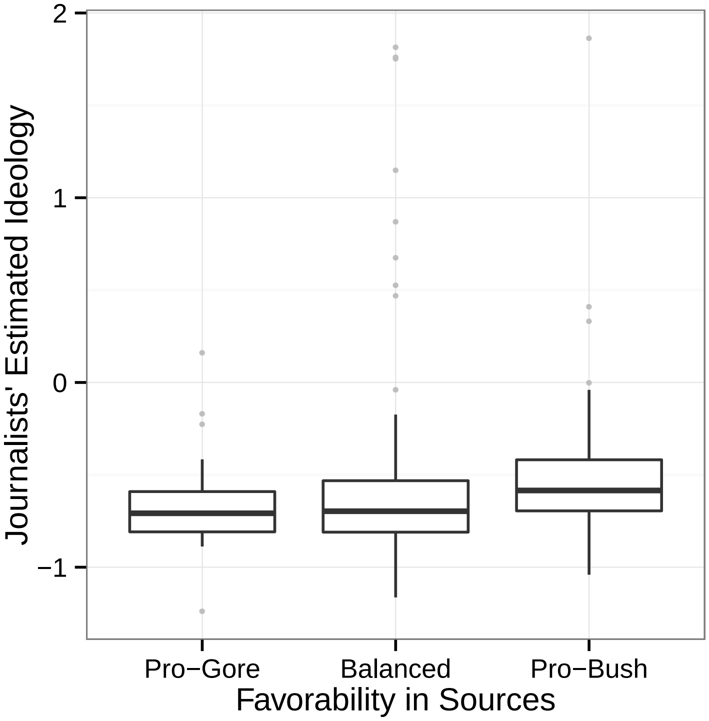


8. Over Time Stability of Ideology of Published Content

As our last application, we assess the over-time stability of ideology of the content published by journalists. To accomplish this, we exploit data from Shaw and Sparrow (1999), which uses measures on political slant of published work from Computer-Aided Research and Media Analysis, International (CARMA). CARMA was contracted by the Republican National Committee (RNC) in 1992 to code newspaper articles over the course of the campaign. As part of the contract, CARMA collected 11,877 articles from 41 newspapers articles during the 27-weeks between May 1 and November 2, 1992 and coded the favorability of the coverage toward H. W. Bush and Bill Clinton on two different 100-point scales.

Of the 305 journalists in the CARMA data, we were able to match 142. Figure 8 presents boxplots of ideological-estimates of three groups of journalists—those who articles were on average pro-Clinton, those whose articles were on average pro-Bush, and those who were balanced. As is evident in the figure, journalists covering George H. W. Bush favorably in 1992 published slightly more more conservative content in 2014 than those who covered Bill Clinton more favorably. The overall differences are modest, and suggest one of many things — a) somewhat more tepid ideological differences between the two parties in 1992, leading to ideology playing a much smaller role in coverage, b) measurement error in ideology stemming from small sample of articles coded per journalist, and c) low over-time stability. Our data cannot disambiguate between the three explanations. But it presents the possibility that one could chart the dynamics of ideology of published content, and assess the extent to which journalists are merely anti-establishment.

Figure 8: Distribution of Ideology Estimates, by Media Outlet



9. Discussion

This paper proposes a way of producing reliable and valid estimates of ideology of a large number of media personalities, television and radio shows, and media outlets on the same scale as politicians and citizens using data from online social networks. The data suggest that the method largely achieves what it promises; the measures obtained using the method correlate heavily with those obtained using other established methods.

The method we propose is one of among many methods that use behavioral measures of audience composition (as opposed to audience measures based on survey self-reports, which are subject to issues like expressive responding (Prior 2013)) to measure ideology of news sources (Gentzkow and Shapiro 2011; Flaxman, Goel and Rao 2014; Tewksbury 2005). Flaxman, Goel and Rao (2014), for instance, use proportion of conservatives following an outlet. Gentzkow and Shapiro (2011), meanwhile, use a structural model of consumer behavior that is similar to the one specified here, except for two important differences. Firstly, unlike the method we propose here, which limits itself to learning from the politically interested, Gentzkow and Shapiro (2011) use data from all the respondents. Furthermore, Gentzkow and Shapiro (2011) assume that the utility function doesn't vary by political interest. And secondly, rather than use social networking data, Gentzkow and Shapiro (2011) relies on domain level visitation data from comScore.

Our method potentially has three advantages over other audience based methods. First, the audience data that Gentzkow and Shapiro (2011); Flaxman, Goel and Rao (2014) use are proprietary, and considerably smaller than data we use. For instance, comScore panel includes just 100,000 households, as opposed to the more than 4 million users that we can rely on. We exploit data that is big, and that is freely available for a large number of sources. Second, our method allows us to produce estimates at variety of different levels of aggregation: journalist, program, newspaper section, newspaper, and television channel. Our method also allows us to estimate ideology of television, radio, and online media. Most common extant proprietary audience data do not provide audience measurement at a fine-grained level, or for multiple media channels. For instance, comScore generally only allows access to online data at the domain level. And crude domain level measures of ideology can bias inferences about the utility function, and hence, ordinal location on the ideology scale. Third, by limiting whom we learn from—the politically interested, our method capitalizes on prior political science research that suggests that the politically interested are the most sensitive to ideology, possibly because they are almost the most aware of it. Exploiting data from the uninformed carries the risk of adding noise.

Our method may also be potentially better than some methods that scale ideology by 'directly' analyzing the content, for e.g, exploiting mentions to think tanks and policy groups (Groseclose and Milyo 2005), and similarity to legislators' speech (Gentzkow and Shapiro 2010), etc. As other scholars have noted, inferring ideology from content is tough, complicated by norms of objectivity in journalism, and the subtlety with which coverage is 'slanted'. Groseclose and Milyo (2005) give the example of the difference between a headline stating 'GDP increased by 5%' versus 'GDP increase less than expected'. It isn't impossible to discern slant directly from automated analyses of content —just hard and likely error-prone (a point our analyses corroborate). And likely, as a result, some of these measures suffer from some critical issues (see for instance, Gasper 2011). Prominently, within sample correlation is low, and some of the estimates have low external validity. For instance, Groseclose and Milyo (2005) estimate *Wall Street Journal* as more left-wing than the *New York Times*.

In all, it may well be 'better' to rely on the judgments about content made by direct manual analysis of the content by millions of people. In this paper, we propose model of how behavior maps on to these judgments, and using the model, we aggregate judgments across people to estimate the ideology of both people and media sources. One concern, however, that is regularly touted is that people are choosing content based on ideological reputations rather than the actual ideology of the content. And this is problematic because reputations can be skewed. The concerns about choosing content based on ideological reputations or for 'expressive reasons' is pertinent where there is limited data. In case of Twitter, where there is a steady stream of information from the sources, it seems implausible that the active politically interested followers are unaware about the ideology of the sources they follow and continue to follow them despite learning that they are far away from their ideal points.

The other major potential concern that is sometimes highlighted is that people follow media and politicians for different reasons. And that validation across politicians doesn't say much about the validity of the method itself. We take the concern seriously but of reflection think that there are three good reasons that suggest that the concern may not be merited. First, the proof is simply in the pudding. As we show in a slew of our validation checks, measures based on the proposed method correlate well with other established measures. Second, we contend that on Twitter, everyone is a news provider, and that little difference exists between a politician and a political news provider. Furthermore, we are subsetting on the judgments of the politically interested about political content. So we see little reason to think that these judgments would be less ideological. And lastly, our method's identifying assumption isn't that ideology be the only variable in decision but that other factors be orthogonal. And as our validation shows, accounting for other factors that influence media consumption like co-location, has little bearing on the overall results. In all, the data suggest that the measures are valid. And in some crucial ways, for e.g., greater granularity and greater reliability, the method also improves upon existing methods.

As we show, having more granular measures allows for discovery of new insights. We find, consistent with Shaw and Sparrow (1999) (see Table 6, pg. 342), that there is considerable ideological heterogeneity within media outlets. And that variation is inversely correlated with outlet extremity — moderate sources carry more diverse ideological content than more extreme sources. Separately, we find that the private beliefs of journalists are correlated with what they publish.

Besides yielding new insights, and providing additional evidence for some existing insights, the new measures also promise to shed light on some long standing issues. For instance, economists have for long speculated about the extent to which journalists' tastes matter above and beyond owner tastes, and market pressures exerted through editorial control. While we find considerable heterogeneity in journalists' ideology employed by the same media outlet, this may be because of conscious editorial policy. To parse out the influence, one can easily build over time measures. Such dynamic measures can us to observe the extent to which journalists sort into ideologically congenial media outlets, either due to self-selection or conscious policy of media companies to recruit journalists of a particular ideological orientation.

A more general formulation of our method can also allow us to discover the extent to which consumption of news media in the population at large is driven by ideology, vis-à-vis say its coverage of sports. In doing so, it can provide more general insights into the structure of news media markets, an important area of investigation.

Lastly, our method may also bring greater precision and clarity in understanding of the 'treatment' offered by partisan media. A variety of identification strategies have been developed to measure the impact of partisan media. Notably, DellaVigna and Kaplan 2007 use idiosyncratic introduction of Fox News Channel to estimate the impact of Fox News on citizen's preferences (see also, Hopkins and Ladd 2013). Meanwhile (Martin and Yurukoglu 2014) use position of the channel on the menu. But little is known with any precision about the ideological 'treatment' that Fox News provides. Precise estimates of salient dimensions of the 'treatment' can allow us to generalize and build better models of persuasion.

Lastly, the success of our method also provides important new methodological directions for future work. Till now, most efforts in supervised text classification —scaling news media is but one example —have relied on direct analysis of structure of text and the extent to which the content 'overlaps.' How the text is cited by (particular) others and the overlap in those citations provides a hitherto underutilized resource in learning about the location of the text on a particular dimension. In some other ongoing work, we merge two sources of information and show that we can achieve considerably better success rates at classifying textual data (though the method need not be limited to text) (Sood and Habel 2016).

References

An, Jisun, Meeyoung Cha, Krishna P Gummadi, Jon Crowcroft and Daniele Quercia. 2012. "Visualizing media bias through Twitter." *Association for the Advancement of Artificial Intelligence (AAAI), Technical WS-12-11* .

Arceneaux, Kevin, Martin Johnson and Chad Murphy. 2012. "Polarized political communication, oppositional media hostility, and selective exposure." *The Journal of Politics* 74(01):174–186.

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

Bonica, Adam. 2013*a*. "Database on Ideology, Money in Politics, and Elections: Public version 1.0 [Computer file]." Stanford, CA: Stanford University Libraries.
URL: *http://data.stanford.edu/dime*

Bonica, Adam. 2013*b*. "Mapping the Ideological Marketplace." *American Journal of Political Science (forthcoming)* .

Bryant, J. and D. Miron. 2004. "Theory and research in mass communication." *Journal of communication* 54(4):662–704.

Clinton, J., S. Jackman and D. Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(2):355–370.

DellaVigna, Stefano and Ethan Kaplan. 2007. "The Fox News effect: Media bias and voting." *The Quarterly Journal of Economics* 122(3):1187–1234.

Egan, Patrick J. 2013. *Partisan Priorities: How Issue Ownership Drives and Distorts American Politics.* Cambridge University Press.

Enelow, J.M. and M.J. Hinich. 1984. *The spatial theory of voting: An introduction.* Cambridge Univ Pr.

Festinger, Leon. 1962. *A theory of cognitive dissonance.* Stanford university press.

Flaxman, Seth, Sharad Goel and Justin M Rao. 2014. "Ideological segregation and the effects of social media on news consumption." *Available at SSRN 2363701* .

Gasper, John T. 2011. "Shifting ideologies? Re-examining media bias." *International Quarterly Journal of Political Science* 6(1):85–102.

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

Gentzkow, Matthew and Jesse M Shapiro. 2011. "Ideological segregation online and offline." *The Quarterly Journal of Economics* 126(4):1799–1839.

Gentzkow, Matthew and Jesse Shapiro. 2005. Media bias and reputation. Technical report National Bureau of Economic Research.

Greenacre, Michael. 2010. *Correspondence analysis in practice.* CRC Press.

Greenacre, Michael J. 1984. *Theory and applications of correspondence analysis.*

Groseclose, Tim and Jeffrey Milyo. 2005. "A measure of media bias." *The Quarterly Journal of Economics* 120(4):1191–1237.

Habel, Philip D. 2012. "Following the Opinion Leaders? The Dynamics of Influence Among Media Opinion, the Public, and Politicians." *Political Communication* 29(3):257–277.

Hindman, Matthew. 2008. *The myth of digital democracy.* Princeton University Press.

Ho, Daniel E, Kevin M Quinn et al. 2008. "Measuring explicit political positions of media." *Quarterly Journal of Political Science* 3(4):353–377.

Hopkins, Daniel J and Jonathan M Ladd. 2013. "The consequences of broader media choice: evidence from the expansion of Fox News." *Quarterly Journal of Political Science* .

Iyengar, Shanto and Kyu S Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of Communication* 59(1):19–39.

Jackman, S. 2014. "Estimates of Members' Preferences, 113th U.S. House and Senate." Retrieved on March 11th, 2014.

Jessee, Stephen A. 2009. "Spatial voting in the 2004 presidential election." *American Political Science Review* 103(01):59–81.

Kohut, Andrew. 2004. How Journalists See Journalists. Technical report Pew Research Center for The People and The Press.

Kohut, Andrew. 2008. State of the News Media. Technical report Pew Research Center for The People and The Press.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(02):311–331.

Lazarsfeld, P.F., B. Berelson and H. Gaudet. 1944. *The peoples choice: How the voter makes up his mind in a presidential election.* New York: Duell, Sloan and Pearce.

Maestas, Cherie D, Matthew K Buttice and Walter J Stone. 2014. "Extracting Wisdom from Experts and Small Crowds: Strategies for Improving Informant-based Measures of Political Concepts." *Political Analysis, forthcoming* .

Martin, Gregory and Ali Yurukoglu. 2014. "Bias in Cable News: Real Effects and Polarization." *Unpublished Manuscript* .

Morgan, Jonathan Scott, Cliff Lampe and Muhammad Zubair Shafiq. 2013. Is news sharing on Twitter ideologically biased? In *Proceedings of the 2013 conference on Computer supported cooperative work.* ACM pp. 887–896.

Nenadic, O. and M. Greenacre. 2007. "Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package." *Journal of Statistical Software* 20(3):1–13.
URL: *http://www.jstatsoft.org*

Oken Hodas, N. and K. Lerman. 2012. How Visibility and Divided Attention Constrain Social Contagion. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).*

Petrocik, John R. 1996. "Issue ownership in presidential elections, with a 1980 case study." *American Journal of Political Science* pp. 825–850.

Poole, K.T. and H. Rosenthal. 2007. *Ideology and Congress.* 2nd edition ed. Transaction Pub.

Prior, Markus. 2013. "Media and political polarization." *Annual Review of Political Science* 16:101–127.

Puglisi, Riccardo and James M Snyder. 2011. The Balanced US Press. Technical report National Bureau of Economic Research.

Shaw, Daron R and Bartholomew H Sparrow. 1999. "From the inner ring out: News congruence, cue-taking, and campaign coverage." *Political Research Quarterly* 52(2):323–351.

Sim, Yanchuan, Brice DL Acree, Justin H Gross and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP.*

Sood, Gaurav and Andy Guess. 2015. Measures of Ideology: Agendas, and Positions on Agendas. In *New Directions in Text as Data Analysis.*

Sood, Gaurav and Philip Habel. 2016. Strength in Numbers: Using Multiple Measures to Esimate Media Ideology. In *The Annual Meeting of the Midwest Political Science Association.*

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior* 30(3):341–366.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics* 75(02):330–342.

ter Braak, Cajo JF. 1985. "Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model." *Biometrics* pp. 859–873.

Tewksbury, David. 2005. "The seeds of audience fragmentation: Specialization in the use of online news sites." *Journal of broadcasting & electronic media* 49(3):332–348.

Toff, Benjamin J and Young Mie Kim. 2013. "Words That Matter: Twitter and Partisan Polarization." Unpublished manuscript.

Weaver, David H, Randal A Beam, Bonnie J Brownlee, Paul S Voakes and G Cleveland Wilhoit. 2009. *The American journalist in the 21st century: US news people at the dawn of a new millennium.* Routledge.

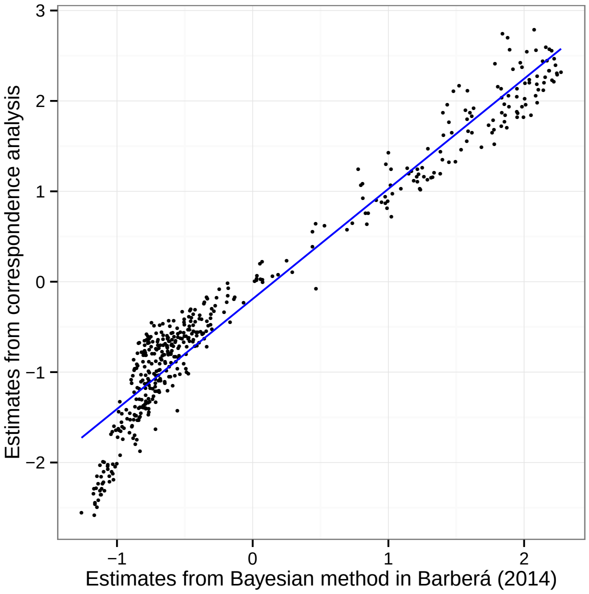
SI 1. Estimation

In order to estimate ideological positions on the latent ideological dimension based on $\mathbf{Y}$, the observed matrix of following decisions on Twitter, we use correspondence analysis (Greenacre 1984, 2010), implemented in the `ca` package for R (Nenadic and Greenacre 2007). The method considers $\mathbf{Y}$ as a representation of a set of points in multidimensional space where distance is measured using a weighted Euclidean metric. A low-dimensional solution is obtained by finding the closest plane to the points in terms of weighted least-squares. Points are then projected onto this plane, and their coordinates are equivalent to the ideal points of each actor on the latent dimensions. ter Braak (1985) shows that correspondence analysis is mathematically close to a log-linear ideal point model.

Correspondence analysis has several advantages that make it particularly useful for our purposes. First, it yields estimates that are extremely highly correlated with estimates from more complex statistical models at a much lower computational cost. In order to demonstrate this, we estimated the MCMC-based method used by Barberá (2015) with a sample of our adjacency matrix, with 500 journalists and 10,000 users. As we show in Figure SI 1.1, we obtained ideal points that are very highly correlated with those estimated using correspondence analysis, but at a computational cost that was dramatically higher, with a running time around 50 times longer.

Second, as Bonica (2013*b*) notes, one of the steps of correspondence analysis consists on normalizing the adjacency matrix by re-weighting rows and columns that are more populated than others, which is equivalent to including user and journalist fixed effects. This is particularly important in our case, given that some journalists are more likely to be followed than others because of their popularity. Finally, it is possible to “project” supplementary observations onto the same low-dimensional space as the main observations, which further reduces computational cost by optionally computing the model only with those observations that contain more information about the latent dimension of interest (ideology, in our case), and then generating estimates for the rest of observations based on the estimated weights of that first subset.

Figure SI 1.1: Comparing MCMC and CA as Estimation Methods



However, our approach also involves at least two drawbacks. First, to the best of our knowledge, there is no easy way to compute standard errors for the estimated ideal points. Second, as Bonica (2013*b*) notes as well, it is not possible to directly add non-spatial covariates into the computation of correspondence analysis. This can be a problem if following decisions are influenced by other factors, in which case the resulting estimates would be noisier, and the interpretation of the latent dimensions would be other than ideology. One solution to this potential problem is to add columns to the matrix indicating following decisions of the same set of users with respect to other "target accounts" with visible ideological leanings. In our application, we add all members of the current U.S. Congress, as well as Pres. Barack Obama. Evidence that our first dimension captures ideology is that the correlation with legislators' ideal point estimates based on roll-call votes is $r = .943$.

To demonstrate that this approach address any potential concerns related to this issue, we fitted a more complex model that included covariates such as journalists' location, affiliation, popularity, and section, for a sample of 500 journalists and 10,000 of their followers. In this case, the objective function we maximize is:

$$\underset{y_{1,\ldots,J}}{\arg\max}\Big[\sum_{j=1}^{J}\alpha_j(y_j) - \beta_i(y_i) + \delta_l(\text{location}_j)\ + \eta_o(\text{outlet}_j) + \tag{2}$$

$$\lambda_s(\text{section}_j)\ + \tau \times \text{followers}_j - y_j(\gamma||\theta_i - \phi_j||^2)\Big]$$

where τ measures the effect of the number of followers and δ_l, η_o, and λ_s are covariate-specific random effects for each journalist or media source j such that:

$$\delta_l \sim N(\mu_\delta, \sigma_\delta) \quad \text{for } l \in 1, \ldots, L \text{ locations}$$

$$\eta_o \sim N(\mu_\eta, \sigma_\eta) \quad \text{for } o \in 1, \ldots, O \text{ outlets}$$

$$\lambda_s \sim N(\mu_\lambda, \sigma_\lambda) \quad \text{for } s \in 1, \ldots, S \text{ sections}$$

To simplify our estimation, we group our covariates into the following categories:

Location 1 = New York City area, 2 = Los Angeles area, 3 = San Francisco Bay area, 4 = other locations.

Outlet Values 1 to 15 for outlets with 10 or more journalists in our dataset, 16 for rest of journalists.

Section 1 = political news, 2 = economic news, 3 = opinion, 4 = international news, 5 = culture and arts, 6 = other sections.

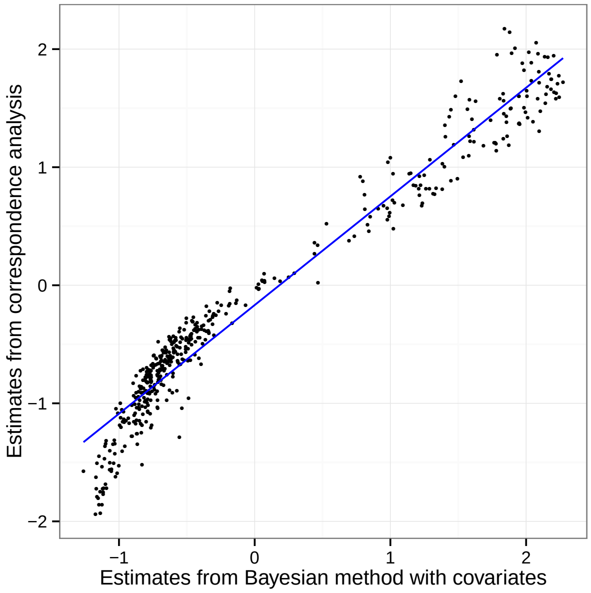
To identify the model, we use the following informative priors on the hyperparameters:

$$\mu_\delta \sim N(0, 0.2) \qquad \mu_\eta \sim N(0, 0.2) \qquad \mu_\lambda \sim N(0, 0.2)$$

$$\sigma_\delta \sim \text{InvGamma}(2, 1) \qquad \sigma_\eta \sim \text{InvGamma}(2, 1) \qquad \sigma_\lambda \sim \text{InvGamma}(2, 1)$$

As we show in Figure SI 1.2, estimates obtained using correspondence analysis and this more complex model with covariates are also highly correlated ($r = 0.976$).

Finally, it is important to note that this method yields ideology estimates for both Twitter users and media outlets, but not necessarily on the same scale. As Bonica (2013*b*) discusses, Figure SI 1.2: Comparing MCMC with covariates and CA as Estimation Methods



row and column coordinates in correspondence analysis share a common dimensionality but not a common scale —this is usually referred to as "between-sets identification problem" (see e.g. Greenacre 2010). In other words, the interpretation of the latent dimensions is the same across sets, but one is usually more "stretched out" than the other. Bonica (2013*b*) solves this issue by using contributors who are also recipients of campaign donations as "bridges" between the two sets, and then rescaling donors' estimates. If researchers were interested in comparing journalists' and media outlets' positions on the latent ideological dimension, it would be possible to apply the same method, since many journalists also follow other journalists and media outlets. In our dataset, 26% of journalists on the columns of the adjacency matrix are also present on the rows. The correlation between their ideology estimates based on who they follow (row coordinates) and who their followers are (column coordinates) is $r = 0.91$. We estimated the same regression as in Bonica (2013*b*), finding that the intercept is 0.04 and the slope is 1.06, which shows that only a minor adjustment would be necessary to rescale users' ideology estimates.

SI 2. Validation

As we note in the main text of the article, we compared our results to two other measures. Here we briefly describe the method behind each of the measures and results of comparison with our measures.

SI 2.1. Comparison with Gentzkow and Shapiro, 2011

Figure SI 2.1 plots our estimates of ideological location of media sources against those in Gentzkow and Shapiro (2011), which were computed using a statistical model of website visits.[11] We find that both are highly correlated: Pearson's r is $.785$ and Spearman's ρ is $.750$ ($N = 76$).

SI 2.2. Comparison with Groseclose and Milyo, 2005

Figure SI 2.2 compares our estimates of media ideology with those in Groseclose and Milyo (2005), computed based on mentions of political think tanks in news articles and newscasts. In general, the estimates overlap. However, there are a few differences. While Groseclose and Milyo (2005) find that the *Wall Street Journal* is the most liberal newspaper, our method locates it slightly to the right of the ideological center. Similarly, rather than finding *Drudge Report* to be ideologically similar to ABC or NBC Nightly News, our method estimates it to be the most conservative among prominent media outlets. On the ideological left, rather than finding *New York Times* and CBS News to be the most liberal, we find NPR Morning Edition to be the most liberal. Our analysis locates the main Twitter account for the *New York Times* close to the *Washington Post* or the *Los Angeles Times*. Though, note that when we consider all Twitter accounts within each media outlet, the *New York Times* does appear to be more liberal the Washington Post. Overall, our estimates are essentially the same as those in Groseclose and Milyo (2005) and, when they differ, they do consistently with other studies of media ideology (Gentzkow and Shapiro 2010; Habel 2012).

[11] Note that we exclude foreign news outlets and websites that do not cover political news.

Figure SI 2.1: Comparing Twitter-Based Ideology Estimates and Estimated Media Slant in Gentzkow and Shapiro (2011)

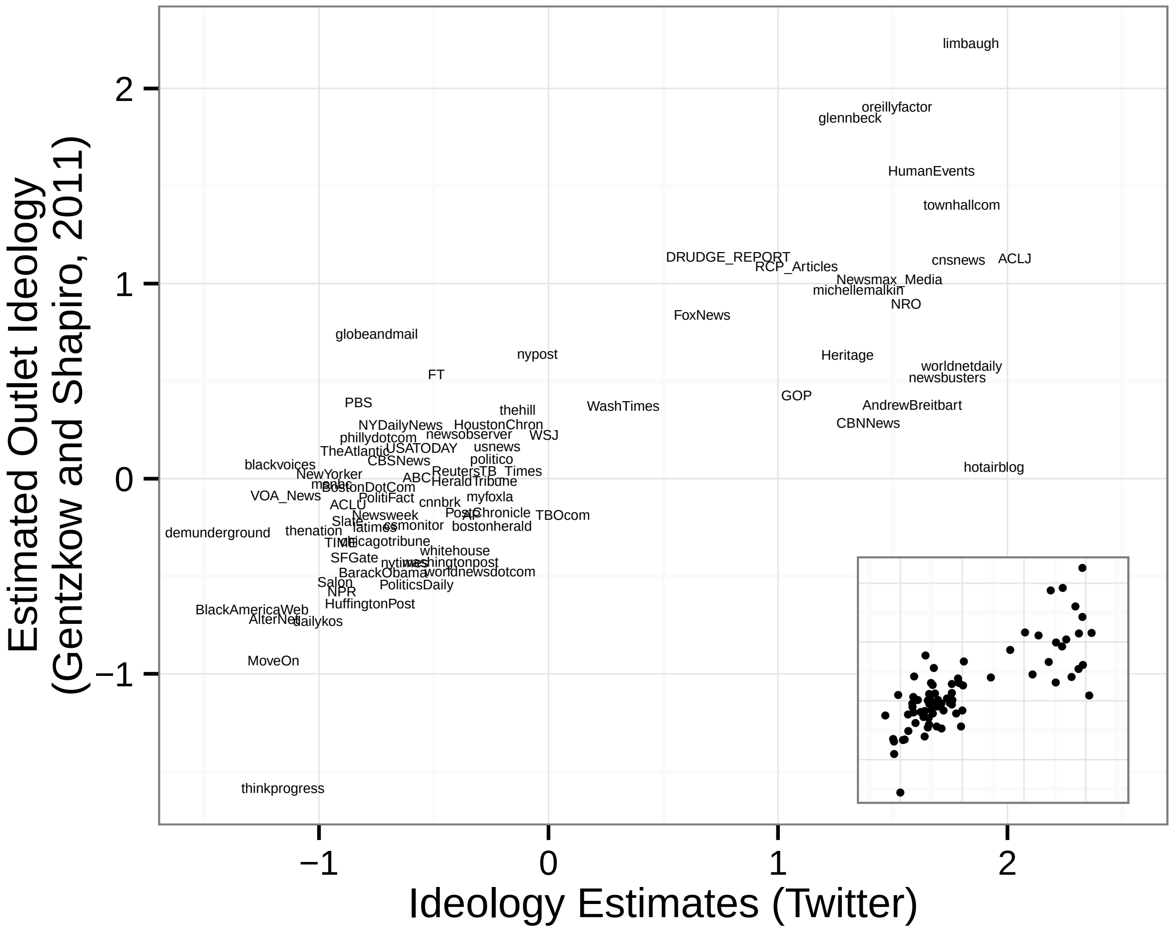
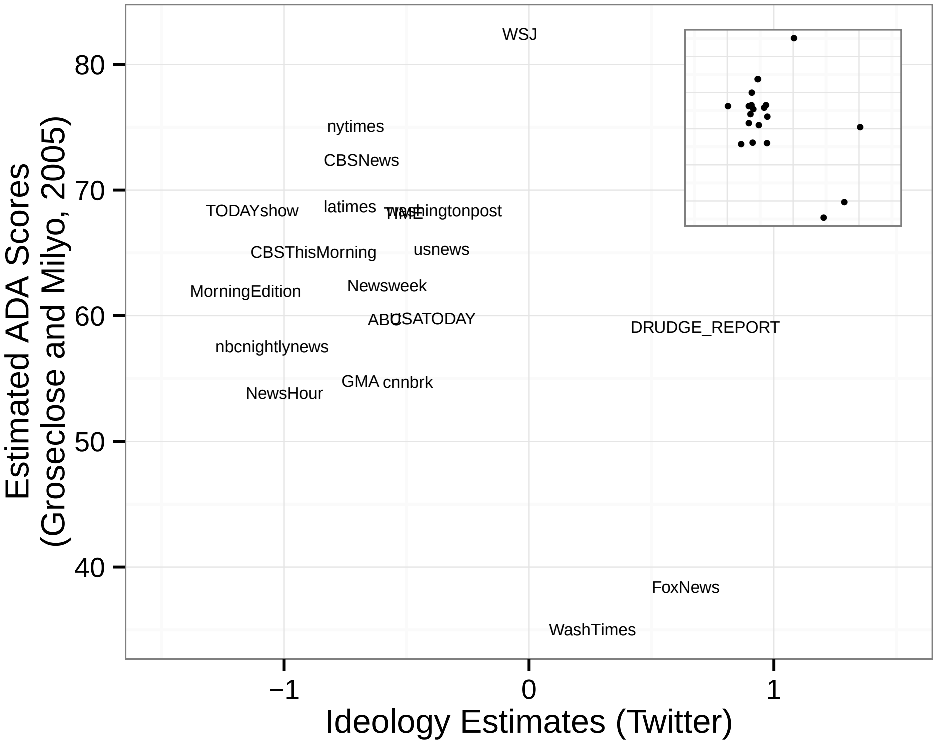
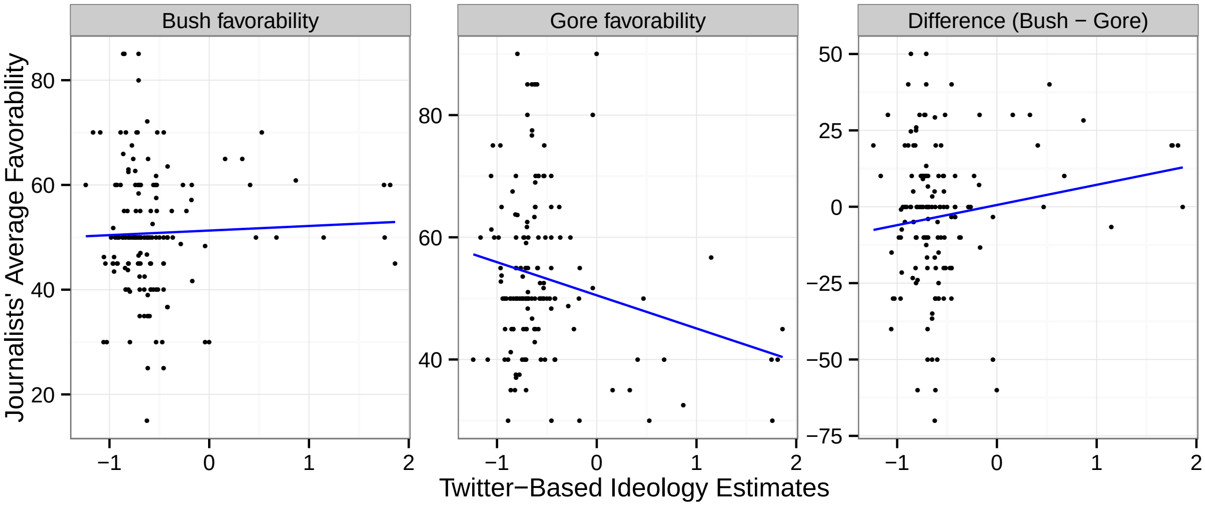


Figure SI 2.2: Comparing Twitter-Based Ideology Estimates and Estimated ADA Scores in Groseclose and Milyo (2005)



SI 3. Comparison with Shaw and Sparrow Estimates

Figure SI 3.1: Distribution of Ideology Estimates, by Media Outlet


Bush favorability
Gore favorability
Difference (Bush − Gore)
Journalists' Average Favorability
Twitter−Based Ideology Estimates