Preaching to the Choir.
Ideology and Following Behavior in Social Media.

Gonzalo Rivero*
Scientific Research Group
YouGov

July 10, 2016

Abstract

Social media offers new opportunities for politicians to mobilize and persuade a large pool of potential supporters, but also allows voters to select whose messages they get directly exposed to. Knowing the factors that make individuals more likely to follow particular politicians may help understand the communication strategies of politicians and also the effects of social media on political polarization. However, information about the offline attributes of individuals in social media is not directly observable. Using a unique database with survey data about the sociodemographic characteristics and political attitudes of 5,580 Twitter users, I show that voters exposed to messages from the Members of Congress are more politically motivated and ideologically extreme than the rest of the Twitter users and that ideological distance between the user and the politician plays a major role in determining following behavior. My results provide a direct validation of previously hypothesized behavior in the literature and have implications for the discussion about how tools that enable both general and microtargeted communication with voters contribute to a fragmented political debate on the Internet.

*The research was done while the author was a member of the Scientific Research Group at YouGov. *Current contact information:* Gonzalo Rivero, Westat, 1600 Research Blvd., Suite 455. Rockville, MD 20850. E-mail: gonzalorivero@westat.com.

I. Introduction

Over the last decade, social media platforms have been quickly embraced by a large majority of the population. According to the Pew Center, 65% of the American adults in 2015 report that they they have used these services at least once (Perrin, 2015), which means that, in spite of a persisting generational gap, social media sites are progressively reflecting the sociodemographic composition of the general population (Perrin, 2015). More relevant for social scientists is the fact that social media users do not compartmentalize their political interests and partisan attachments away from the rest of their online lives. On the contrary, a significant portion of them seem willing to talk and listen to messages about public affairs and politics through the same channels they use for more personal communications, as illustrated by the fact that at least 66% say that they have engaged in political activities through social media (Rainie et al., 2012) and that 22% indicate that they resort to these tools primarily to obtain political information (Smith, 2014).

Those numbers alone are sufficient to explain why politicians join platforms like Twitter, Facebook or Instagram: social media offers to the political elites a very large audience that can be reached at a very small cost and in real time. In addition, its horizontal structure allows the elite to broadcast messages without the intermediation of traditional media brokers such as newspapers or radio stations (Bode et al., 2011; Gainous and Wagner, 2014; Lassen and Toff, 2015). As a consequence, it enables politicians to establish a direct, controlled, bidirectional communication strategy with voters that softens the need to compete with other candidates for the limited attention and space of more traditional media (Gainous and Wagner, 2014). It is far from surprising that already in 2012, nearly 90 percent of all major party candidates were active Twitter users and 98% percent of the U.S. Representatives were using at least one social media platform as a communication and outreach tool (Greenberg, 2012).

However, although the elite has obvious incentives to follow the electorate to wherever they go, the pending question is why the latter listens to the former in social media, and more specifically, what makes voters follow particular politicians on Twitter or Facebook. It would be hardly surprising for any political scientist to find that ideological proximity increases the likelihood of subscribing to a particular politician's feed (Barberá, 2015a). But the *the strength* of such relationship relative to other covariates and, more importantly from an applied perspective, our ability to exploit it in order to recover the ideology of users from behavior alone requires a thorough effort on the measurement of the effect of political attitudes on what users do on the platform.

Unfortunately, the study of the determinants of following behavior is severely limited by the lack of readily available data about the sociodemographic and political attributes of users of social media. Only in rare occasions, researchers have detailed information about the offline characteristics of the users, but in most cases they have to be inferred from their behavior on the platform (Kosinski et al., 2013). This is particularly true in the specialized literature studying the effect of political attitudes on social media behavior, which oftentimes has to

rely on indirect validation (Boutet et al., 2012; Conover et al., 2011; Bond and Messing, 2015; Barberá, 2015a).

In this article, I present a systematic study of the decision to follow politicians on Twitter using a unique, large survey data with information about the offline characteristics of Twitter users and the Members of the 114th U.S. Congress. I test the expectations the spatial model of following behavior on Twitter (Barberá, 2015a) using the reported ideal points of voters. I use a database of 5,580 Twitter users for whom I have survey data about their political attitudes and their sociodemographic profile in addition to their behavior on Twitter. This dataset, which connects offline attitudes and online behavior, is unique in the literature and offers insights into how political preferences and disposition towards political news have an effect in the decision to listen political elites on social media.

Unsurprisingly, I find support for the hypothesis that ideological distance between users and elites has a significant effect in the selection of politicians that voters listen to. However, the spatial model performs suboptimally for politically moderate users and overestimates the degree of polarization in the electorate, which in turns limits its ability to recover the state of public opinion. Part of the shortcomings of the model derive from the differential attention of Twitter users to politics. Specifically, my results show that individuals following members of the U.S. Congress on Twitter are more likely to be located in the extremes of the ideological scale, which induces a systematic bias in the sample that is available to the analyst. I am also able to identify demographic biases similar to those already reported in previous research (Smith, 2014): women, minorities, and users with lower education attainment are less likely to follow elite accounts, which may also impact the results.

The rest of the article is structured as follows. In the next section, I review the relation between the literature on political communication in social media and selective exposure, and how this research is related to the academic debates on political polarization and the retrieval of offline attributes from social media behavior. I then describe the dataset used in the empirical analysis in Section III. Results are presented in Section IV. The final section summarizes the results.

II. Theoretical framework

I. Social media, political communication, and ideological sorting

Social media has become a standard component of the communication toolkit of the political elite. Platforms like Twitter or Facebook, adopted early on by a few candidates to cater to young voters (Goldschmidt and Ochreiter, 2008), gained widespread adoption by the 2012 election campaign (Gainous and Wagner, 2014). At that point, nearly all major party candidates (including incumbents) used at least one social media platform as part of their communication strategy (Greenberg, 2012) with a small imbalance in favor of Republicans (Bode et al., 2011; Peterson, 2012).

The quick adoption of social media by politicians is not surprising. Aside from the sheer potential for outreach of social media platforms, tools like Twitter or Facebook offer a space for a low-cost, direct communication with voters, with the possibility of integrating it into microtargeting of potential supporters (see Issenberg, 2012; Nickerson and Rogers, 2014, for the role of social media in political campaigns). Furthermore, social networking sites allow politicians to escape the intermediary role of traditional media (Bode et al., 2011). As Lassen and Toff (2015) put it, now politicians "need not wait for relevant legislation, hope for a journalist's favorable pen, or pay for access to a given media market to express their position on a subject. Similarly, producing one or even a dozen tweets may be far less resource intensive for a campaign than crafting and distributing longer form newsletters or press releases."

However, regardless of the peculiarities and innovations that social media introduces in the flow of political communication in relation to other more traditional tools, it can still be understood within a theoretical framework in which the primary aim of communication with voters is to win elections (Fenno, 1978) using *advertising*, *credit claiming*, and *position taking* (Mayhew, 1974). In particular, social media allows incumbents to advertise their name and actions in office to create a favorable image and improve their name recognition with messages that may have little legislative content (Bode et al., 2011; Peterson, 2012).

But politicians also use social media —and Twitter more specifically— to take political positions and locate themselves in the dimensions of competition of the political arena. As a matter of fact, Greenberg (2012) finds that roughly two-fifths of all tweets and Facebook posts could be classified as position-taking, making it the most popular purpose of messages in either platform, although there is considerable variability on whether these posts are from incumbents or challengers (Gainous and Wagner, 2014, chapter 5). Therefore, politicians not only "address issues of the day, respond to news and media accounts, and update supporters and followers on campaign activities" (Bode et al., 2011), they also use the platform to reinforce and gain support from their base through ideological messages (Lassen and Brown, 2010; Bode et al., 2011; Roback and Hemphill, 2013; Otterbacher et al., 2012). Precisely this ideological content of the communications on Twitter with both constituents and the general public is what allows Toff and Kim (2013) to recover the partisanship of individual legislators by using only the content of their tweets arriving to similar estimated ideal points than other studies using widely different political output like roll call data (Poole and Rosenthal, 2000), political donations (Bonica, 2014), floor speeches (Diermeier et al., 2012), or newsletters (Cormack, 2013).

As a consequence of the previous regularities, Twitter behavior reflects and does not dilute the political preferences of the elite, which opens the gate for an ideology-based interaction with followers in social media. In other words, politicians use Twitter to communicate political positions on issues, and voters are not blind to those positions when they decide who to follow. Therefore, other motivations aside, users are expected to behave in a manner that is consistent with a *proximity* or *spatial* model (Barberá, 2015a).[1] In this model, users are more likely to follow political elites whom they perceive closer to their own ideal points

[1] A similar idea is pursued by (Bond and Messing, 2015) in their study of endorsements on Facebook.

in an ideological space. Even if it is possible that the following behavior may simply reflect an expression of identity or attachment to a given political label, the platform still induces users to be directly exposed to ideological messages and it is reasonable to assume that individuals will have a preference for avoiding cognitive dissonance in their feeds (Festinger, 1957).

This notion of a proximity model of behavior is consistent with previous research on ideological sorting in online networks (Adamic and Glance, 2005; Conover et al., 2011): individuals very strongly tend to follow and interact only with users that share their same political views (Conover et al., 2011; Yardi and Boyd, 2010). This effect remains true even in we discount the *filter bubble* effects produced by recommendation engines (Pariser, 2011) and the fact that the horizontal nature of social media is also likely to expose individuals to information and sources that they would not seek out themselves and that contradict the core beliefs of the user (see Messing and Westwood, 2012; Barberá, 2015b; Bond and Messing, 2015).[2]

However, while ideological proximity may be a explanatory factor in online behavior, it is possible that the decision to follow politicians in social media is mediated by the user's interest in politics or the strength of her ideological convictions. Therefore, while users may use their ideology to decide which politicians to follow, only non-moderate users with high interest for government affairs will follow political elites on Twitter to begin with. This profile of behavior is consistent with aggregate studies on social media (Smith, 2014) but also with individual-level research on the decision to subscribe to politician's e-mails and newsletters (Cormack, 2013). Even more, it is the kind of behavior we would expect from the fact that those who prefer one-sided information are also more likely to gather political information through social media (Gainous and Wagner, 2014, chapter 2). Thus, the ideology of moderate users may prove more difficult to recover from the graph of following behavior.

The previous arguments suggest that ideology may be central to the behavior of users in regard to being exposed to the opinions of the political elite on social media. If that is the case, it should be possible to make the way in the other direction and infer political preferences from online behavior (Barberá, 2015a). But the users who engage in political activities online is not necessarily a representative sample of all users and is likely to over-represent highly motivated, political engaged individuals. This differential behavior will pose obstacles for using a proximity-based model to recover ideological preferences from users.

In the following section, I empirically tackle the question about the validity of the spatial model to recover individual ideology from social media by exploiting a unique dataset of 5,580 Twitter users for whom I have survey data about their demographic and socio-political attitudes. This survey data allows me to estimate individual models for their decision funnel to follow accounts from Members of the U.S. Congress. In particular, it allows me to test the three steps leading to the decision of subscribing to the feeds from particular political elites.

[2]Not surprisingly, 38% of social media users report to have discovered through that channel that their friends' political beliefs were different than they thought (Rainie et al., 2012).

I first study why individuals follow accounts from particular politicians. My model replicates the structure of the spatial model of following behavior (Barberá, 2015a), which assumes that individuals will follow politicians who are closer to them in an ideological space. My dataset, which is described in the following section, provides a unique opportunity for a direct test of the theory, as it contains the reported ideological location of individual users in addition to their online behavior.

I then analyze the cause of some discrepancies between observed and reported ideology by taking a look at the extent to which political ideology and interest for public affairs affects the probability of following at least one account from a Member of Congress. Therefore, the analysis gives us an insight on the decision to be exposed to partisan messages on Twitter in the first place. Building on the results about watching, reading, and listening to news in offline and online traditional media (Kohut et al., 2012), the expectation is that more educated individuals with higher interest for public affairs are also more likely to subscribe to a politician's feed. Also, it has been found that white, older individuals have a higher propensity of following news (Kohut et al., 2012) and subscribing to political newsletters (Cormack, 2013) and that prediction carries forward to this case of study.

Finally, I study the variables that affect the number of accounts from Members of Congress followed by a given user. The interest is now in the *level* of interest for receiving partisan messages and how it affects the amount of exposure for a given social media user. It is natural to expect that the same variables that affect in the model above influence how many feeds the user will subscribe to.

II. Related literature

This research is related to the literature on the link between selective exposure and political polarization (Prior, 2013; Bennett and Iyengar, 2008; Stroud, 2011). One of the dominant topics in recent American Politics has been how representatives have become more extreme ideological than ever (Poole and Rosenthal, 2011; McCarty et al., 2006), regardless of the relative stability and moderation of mass public opinion (Fiorina et al., 2006). It has been argued that media reinforces political polarization through the limited variation of partisan news to which readers are exposed (Iyengar and Hahn, 2009). To the extent that social media reflects the same pattern and that conversations about political issues seem to be dominated by people with extreme ideological positions (Barberá and Rivero, 2014; Conover et al., 2012), it is possible that it has similar effects. As a result, social media reinforces polarization through the creation of entire networks of reinforced beliefs (Sunstein, 2008): "The end result is very different groups of people, living in entirely different informational networks, creating increasingly isolated cultural and ideological bubbles" (Gainous and Wagner, 2014).

This article also speaks to the vast literature estimating offline traits for social media users using supervised and unsupervised learning with the aim of using Twitter as a source for better understanding the opinion of the general population in real time. The literature has tried to recover age (Nguyen et al., 2013), gender (Ciot et al., 2013; Liu and Ruths, 2013),

ethnicity (Pennacchiotti and Popescu, 2011) and political orientation (Boutet et al., 2012; Conover et al., 2011; Barberá, 2015a; Ecker, 2015) with different degrees of success (most notably Kosinski et al., 2013). In our research, this information is known for each user and used to estimate the observed behavior and therefore permits a direct validation of a behavioral model.

Finally, this article has direct consequences for the literature predicting aggregate political behavior using social media —namely, elections (Gayo-Avello et al., 2013). Starting with Tumasjan et al. (2010), the literature has tried to find ways to use sentiment, following behavior, and retweet behavior as ways to estimate the expected behavior of users in the offline political world. However, it has been quickly recognized that for the enterprise to be successful, the research needs information about the sociodemographic attributes of the user and her online presence in order to correct the differences between Internet/social media population and the general population (Barberá and Rivero, 2014), and between accounts and usage of those accounts (Gayo-Avello, 2012), especially considering that younger, urban, and minority populations (Mislove et al., 2011; Lenhart, 2009) and also ideologically extreme users (Barberá and Rivero, 2014) are overrepresented on Twitter.

III. Data

Data was collected from the database of panelists of YouGov. The YouGov U.S. panel, a proprietary opt-in survey panel, is comprised of U.S. residents who have agreed to participate in YouGov's web surveys. Panel members are recruited by a number of methods to help ensure diversity in the panel population, including web advertising campaigns and permission-based email campaigns. Panelists are subjected to a double opt-in procedure where they are informed of the privacy policy and agree to receive survey invitations. All panelists are profiled on basic socioeconomic demographics, political attitudes and behavior, health status and consumer behavior. Participants are not paid to join the YouGov panel, but they receive incentives through a loyalty program to take individual surveys. Verifications against public voter rolls (through studies that require registered voters and vote turnout data) and controls against fraudulent are in place to flag and remove panelists who do not provide verifiable information.

The construction of the data set used in this article proceeded in two stages. First, a convenience sample of US YouGov panelists was asked about their Internet usage. The questionnaire included a section about their use of social networking sites. Respondents who said they used Twitter were then followed up to provide their Twitter handle in a open-text box. Those handles were then validated using the Twitter API. For set the validated users I then collected information about the people they followed as of June, 6, 2015. A total of 5,580 YouGov panelists provided valid Twitter handles.

The second stage consisted on retrieving *offline* data for the YouGov panelists with validated Twitter profiles. YouGov periodically asks all panelists a number of questions with identical wording and response options that are used in the sampling strategy. These questions

include basic sociodemographic data and sociopolitical attitudes. For each of the users in the Twitter sample, I collected the latest available information in the database regarding their gender, age, education, racial self-identification, ideology, partisanship, and general interest in government and public affairs.

A descriptive summary of the data is presented in Table I.[3]

Table I: Summary statistics of the dataset

	N	Mean	Min	Max
Birth year	5525	1960	1935	2000
Gender	5580	1.54	1	2
Education	5580	3.99	1	6
Race	5580	1.5	1	5
Ideology	5578	3.03	1	6
Party ID	5576	3.47	1	8
News interest	5478	1.6	1	4

I also built a dataset on political elite accounts which I restricted to members of the U.S. Congress (US House of Representatives and Senate) as of August, 1, 2015. The decision to limit the set of politicians and political actors to Members of Congress was driven by availability of ideal point estimates from the DW NOMINATE database of roll call voting behavior (Poole and Rosenthal, 2000), that is standard in the political literature. Therefore, each account was matched against their ideological ideal point in the DW NOMINATE scores for the 114th Congress.

By restricting to members of the U.S. Congress, my sample of political actors is only slightly more limited than the one used by Barberá (2015a), which includes political representatives, political parties, and media outlets and journalists who tweet about politics.

[3] The variables are coded as follows:

- Gender: 1: Male, 2: Female.
- Education: 1: Did not graduate from high school, 2: High school graduate, 3: Some college, but no degree (yet), 4: 2-year college degree, 5: 4-year college degree, 6: Postgraduate degree (MA, MBA, MD, JD, PhD, etc.).
- Race: 1: White, 2: Black or African-American, 3: Hispanic or Latino, 4: Asian or Asian-American, 5: Other.
- Ideology: 1: Very liberal, 2: Liberal, 3: Moderate, 4: Conservative, 5: Very conservative, 6: Not sure.
- Party ID: 1: Strong Democrat, 2: Weak Democrat, 3: Lean Democrat, 4 : Independent, 5: Lean Republican, 6: Weak Republican, 7: Strong Republican, 8: Not sure.
- Frequency with which the respondent follows political information: 1: Most of the time, 2: Some of the time, 3: Only now and then, 4: Hardly at all.

IV. Results

I. Who to follow?

Table II shows the proportion of individuals of a given ideological group who follow a Member of Congress. For instance, we see that 27.4% of the respondents who consider themselves very liberals follow at least one Representative or Senator from the Democratic party, while 30.5% of very conservatives follow one from the GOP.[4] Similarly, only 9.2% of the moderates follow a Democrat and 8.3% follow a Republican. Therefore, what the table shows is that individuals from the extremes of the ideological scale are more likely to follow their own party and only a small percentage is exposed to messages from the other side of the aisle. In addition, moderates are less likely to follow anyone, and in fact they are equally likely to follow accounts from either party. Note how this number relates to the 19% of respondents reported signing up for official messages from their Representative and 14% from their Senators at some point in their lives (Cormack, 2013), or the 20% of social media users that report following elected officials and candidates for office (Rainie et al., 2012).

Table II: Probability of following a Member of Congress of a given party by ideology

	Very liberal	Liberal	Moderate	Conservative	Very conservative
Democrats	27.4	19.7	9.2	3.4	5.0
Independent	13.7	7.7	2.7	0.2	0.3
Republicans	5.6	5.4	8.3	22.1	30.5

This notion of selection into ideological exposure to the respondent's favorite party is even clearer in Table III which shows the diversity of parties to which a respondent from each individual group listens. In particular, it counts the number of different Members of Congress from different parties a individual follows.[5] It shows that only around 4% of respondents are exposed to voices from both sides, although interestingly enough those in the extremes are just slightly more likely to follow both parties rather than one.

Table III: Diversity of parties followed by ideology

	Very liberal	Liberal	Moderate	Conservative	Very conservative
No parties	69.6	77.4	84.9	76.2	68.9
1 party	25.8	18.7	11.5	21.8	26.4
2 parties	4.5	3.8	3.4	1.9	4.6

Both tables point to the same idea: individuals tend to select only accounts from their party and the proportion of individuals listening to messages from both sides of the aisle

[4] Note that Independents in the 114th Congress caucus with the Democrats.

[5] I have collapsed Independents with Democrats.

on Twitter is no greater than 5%. Therefore, these results suggest that it is possible to "go the other way," i.e., to recover the political affinity of a given individual by looking at the politicians a given users follows on Twitter. In order to test whether reverse engineering ideology from behavior on Twitter is possible we need to verify two different things. First, we have to verify that indeed ideology is a predictor of following behavior. And second, we need to evaluate the predictions of the model against some ground truth model. Those are the two tasks that this section tries to accomplish.

I first start by testing the spatial model of behavior. In particular, I follow here the specification by Barberá (2015a) and assume that my decision to subscribe to a given politician's account is mostly the result of the distance between the politician's ideal point and my own. In a Item Response Theory framework the model takes the following structure:

$$\Pr(y_{ij} = 1) = \text{logit}(\alpha_i + \eta_j + \gamma|\theta_j - \theta_i|); \tag{1}$$

where y_{ij} is an boolean variable for whether respondent i follows account from Member of Congress j, α_i is a panelist-specific random-effect that captures the propensity of i to follow, η_j is a elite-specific random effect that measures the differential popularity of Member of Congress j, γ is a parameter that scales the effect of the ideological distance on the decision to follow an account, θ_j is the ideal point of the Member of Congress j, and θ_i is the ideal point of the panelist.

Using Equation 1, I have run two separate models depending on whether θ_i was assumed to be observed or not. In both cases, the value of θ_j was taken to be the DW NOMINATE score for Member of Congress j. In the first model, I used the answer to the ideology question in a 5-point scale for each of the YouGov panelists (see Table I) as the ideal point θ_i, after rescaling it to match the $[-10, 10]$ range of values in DW NOMINATE. Therefore, in the first model, the distance between the politician and the respondent is assumed to be known. Results are shown in the top block of Table IV. The second model, shown in the lower block of Table IV replicates[6] the original specification in Barberá (2015a) and assumes that θ_i is a parameter to be estimated from the data.[7] It is very relevant to remark that the first model is applied to all respondents in the sample, including those who do not follow any politician. However, the second model uses only respondents who follow at least one Member of Congress, as otherwise the ideology is not defined according to the model.

Table IV shows the γ parameter, which captures the effect of ideological distance in the probability of following, and also the hyper-parameters for the individual- (indiced by α) and politician-specific (indiced by η) random effects. The main message from the models is clear: γ, the scale parameter that measures the effect of distance, is negative and statistically significant in both Model 1 and 2. We therefore have direct, primary evidence in favor of

[6] In his paper, Barberá (2015a) sets himself a more difficult task and attempts to estimate θ_j from data, while here I assume the value is known.

[7] Both models were run though the `R` interface to `Stan` using standard priors. The specification included the following priors: α_i was assumed to be $N(\mu_\alpha, \sigma_\alpha)$, where μ_α follows a $N(0, 100)$ and σ_α follows a Cauchy$(0, 5)$. A similar structure of priors and hyperpriors was used for η_j. γ was specified as a $N(0, 100)$. For the second model, the priors for the ideology of each individual was assumed to be $U(-10, 10)$.

Table IV: Estimation of the decision to model a given Member of Congress account.

		Coefficient	2.5%	97.5%
Model 1	μ_α	-3.126	-4.074	-1.744
	σ_α	2.957	2.816	3.097
	μ_η	-6.821	-8.212	-5.914
	σ_η	1.696	1.569	1.839
	γ	-0.257	-0.263	-0.246
Model 2	μ_α	-2.078	-2.909	-1.628
	σ_α	1.061	1.003	1.121
	μ_η	-3.500	-3.984	-2.681
	σ_η	1.733	1.608	1.884
	γ	-0.426	-0.446	-0.405

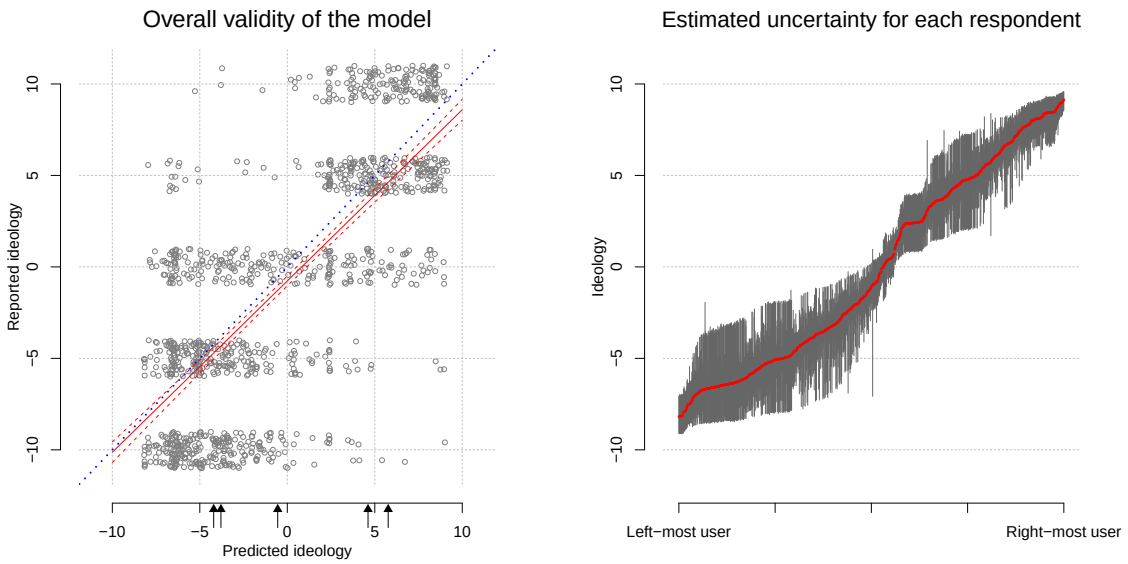
the first requirement, namely that politicians further apart from the respondent are less likely to be followed.

Table V: Standard deviation of the predictions by reported ideology

Reported ideology	Mean	Standard deviation
Very liberal	-4.204	2.844
Liberal	-3.788	3.140
Moderate	-0.550	4.737
Conservative	4.611	3.543
Very conservative	5.767	2.574

However, a significant coefficient is not enough support for the model. The next step is to verify the predictive performance of the model. Panel (a) in Figure 2 shows the predicted values (plus jitter) of ideology for each respondent according to Model 2. The blue dashed line represents an identity function and the solid red line a the estimated regression of observed and predicted values. The intercept of the regression line is $\beta_0 = -0.76, s.e.(\beta_0) = 0.135$ and the slope $\beta_1 = 0.937, s.e.(\beta_1) = 0.025$. The close correspondence between the two lines is some initial evidence that the predictions are consistent at least in the sense that the model recovers some basic structure. In fact, the model does an excellent job at sorting people with identifications other than moderate: liberals and very liberals are consistently assigned predicted values below 0 and conservatives and very conservatives are above 0.

Unfortunately, there are some obvious issues with the predictions. As it can be readily seen from the figure, the moderates are placed all through the political spectrum. In particular, as Table V shows, the standard deviation of the predicted ideology for people who consider themselves moderates is about 1.7 times larger than the standard deviation for very liberals or very conservatives. The result is consistent with Table II and the idea that moderates are equally likely to follow politicians from either party. Not only that, the model attributes very low uncertainty to the predictions of moderates (Panel (b) of Figure 1).



(a) Predicted and observed ideology (b) Estimated uncertainty

Figure 1: Predictions of the model

A second problem is related to intra-group classification. The arrows in the lower margin indicate the median of the predicted values according to Table IV for each of the categories of ideology in the raw data. The model seems unable to reproduce the very liberal and very conservative groups or, more correctly, it tends to pull extremists and moderates closer than they should.

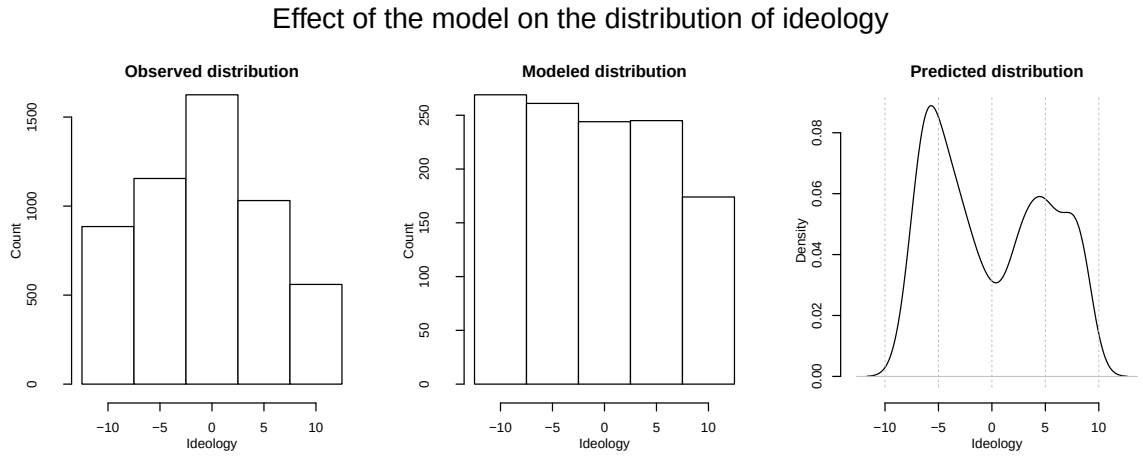


Figure 2: Predicted and observed ideology

The third and final problem is related to the overall, population-wide estimate of polarization. As Figure 2 shows, the ideological distribution of the full sample (left panel) shows a clear bell shape, that disappears once we select only individuals that follow at least one

Member of Congress (center panel), mostly because the moderate group is trimmed the ideological distribution. Finally, the predicted distribution from Model 2 (right panel) is strongly bimodal, even when the original data that enters into the model is approximately uniform. In other words, recovering ideology based only on following behavior exaggerates the degree of polarization in the population.

There are two results to extract from the analysis above. On the one hand, a positive theoretical message: we have some evidence in favor of the spatial model of behavior on Twitter and the notion that users expose themselves only to like-minded politicians. On the other, the fact that the model induces some systematic biases when trying to recover individual- and group-level estimates. In the next section I explore the most obvious candidate as explanation for the problems of the model, namely the selection induced by political interest and ideology and following behavior.

II. Follow any accounts?

I start the analysis by looking at the probability with which Twitter users follow accounts from Members of Congress. Let be an y_{ij} indicator variable for whether user i follows account j in the set of Twitter accounts of members of the 114th Congress. I then model the probability of following one of these political accounts as:

$$\Pr\left(\sum_j y_{ij} > 0\right) = \text{logit}(x_i^T \beta), \tag{2}$$

where x_i^T is the vector of covariates observed for individual i indicated in the section above, and β is a vector of parameters associated with each covariate.

The main results are shown in Table VI. The model uses the demographic and political information available for each respondent and removes the few individuals for which at least one covariate has not been observed. Note that the age of the respondent has been shifted so that 18 is represented by a value of 0, and transformed with a cubic B-spline to capture non-linearities in the following behavior.[8]

Results depict an image that is similar to the intuitions that are commonly found in the literature about social media and politics (Mislove et al., 2011). The model finds a gender bias and indicates that men are 10% more likely to follow political accounts than women. Also, the probability of following Members of Congress monotonically increases with education, as can be seen in the coefficients. In the dataset, we can observe a 16% difference between users with less than a high school diploma users with post-graduate education, although the only significant variable in Table VI is the indicator for high school graduates. Interestingly enough I do not find a effect of age. Even if older users are less likely to use the tool in the first place, once they have an active Twitter account they seem equally likely

[8] In order to avoid the B-spline to be affected by extreme observations, cases above the 99% of the age distribution were also removed from the analysis.

Table VI: Probability of following any elite accounts

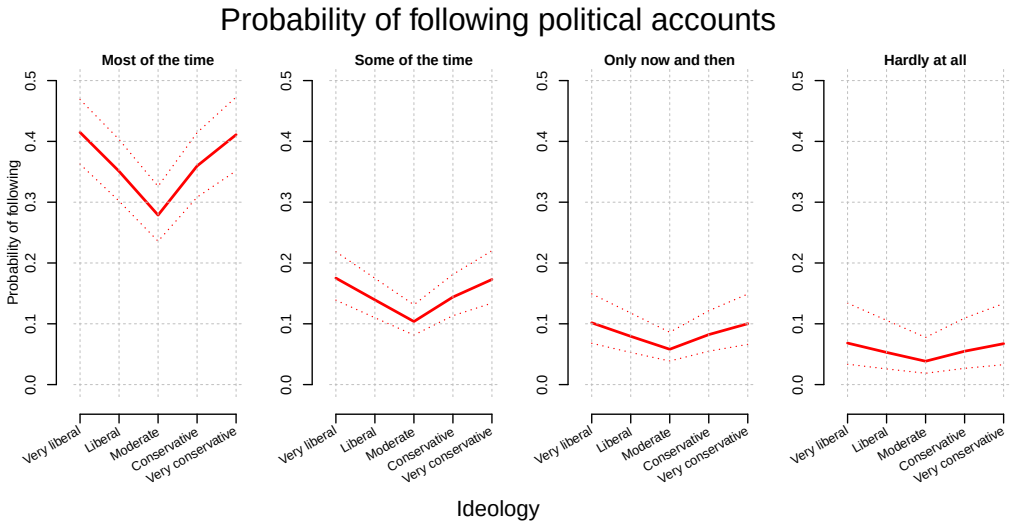
Variable	Coefficient	2.5%	97.5%
(Intercept)	-0.152	-0.834	0.510
Age: 1st piece	-0.572	-1.971	0.858
Age: 2nd piece	0.315	-0.260	0.898
Age: 3rd piece	-0.476	-1.335	0.395
Ideology: Liberal	-0.271	-0.479	-0.063
Ideology: Moderate	-0.606	-0.815	-0.397
Ideology: Conservative	-0.233	-0.450	-0.015
Ideology: Very conservative	-0.015	-0.260	0.227
Ideology: Not sure	-1.047	-1.673	-0.495
Gender: Female	-0.190	-0.332	-0.048
Education: Less than High school	-0.195	-0.824	0.381
Education: High school graduate	-0.470	-0.768	-0.175
Education: Some college	-0.111	-0.348	0.128
Education: 4-year college	0.013	-0.218	0.249
Education: Postgraduate degree	0.025	-0.222	0.277
News interest: Some of the time	-1.205	-1.402	-1.013
News interest: Only now and then	-1.835	-2.248	-1.460
News interest: Hardly at all	-2.268	-3.075	-1.611
Race: Black or African-American	-0.338	-0.606	-0.081
Race: Hispanic or Latino	-0.041	-0.355	0.259
Race: Asian or Asian-American	-0.439	-1.074	0.126
Race: Other	-0.099	-0.415	0.203
N	5423		
AIC	5183		
Balance	21%		

Notes: Balance indicates the proportion of observations with value 1 in the outcome variable.

as younger users follow Members of Congress in Twitter. Finally, the data clearly shows a difference between whites and minorities. Even if minorities may be overrepresented on Twitter (Mislove et al., 2011), in my dataset whites are 7% more likely to follow Members of Congress than any minority.

But more importantly, I find that the users on the extremes of the ideological scale are also the most likely to follow elite accounts (Figure 3). Those who consider themselves very liberal or very conservative are *on average* 15% more likely to follow Members of Congress than moderates, from a 31% to about 15%. Therefore, part of the reason the model in Section I overestimates the level of polarization has to do with the fact that it eliminates moderates from the sample as they are less likely to follow politicians to begin with and therefore cannot be scaled. It is also remarkable how the difference between not reading

Figure 3: Predicted probability of following a Member of Congress on Twitter



political news at all and following them most of the time nearly doubles the size of the full effect of ideology (29% versus 15%).

Therefore, the results support the usual intuition found in the literature about the fact that the people with the most active participation in the political debate on Twitter actually comes from the political extremes and not the center (Barberá and Rivero, 2014; Conover et al., 2011, 2012), which in turn causes obstacles for the spatial model to be able to recover the distribution of ideology.

III. How many accounts to follow?

Now the question is about whether this bias is accentuated or attenuated by the behavior *once engaged* in political following in the platform. Therefore, we are interested in estimating the determinants of the number of accounts followed by each respondent.

Let n_i be the number of accounts from Members of Congress that respondent i follows. The model was specified as a zero-inflated binomial regression model:

$$\mathrm{E}\left(n_j = k\right) = \Pr(n_i = 0) + \Pr(n_i > 0)\mathrm{E}\left(n_i = k | n_i > 0\right) \tag{3}$$

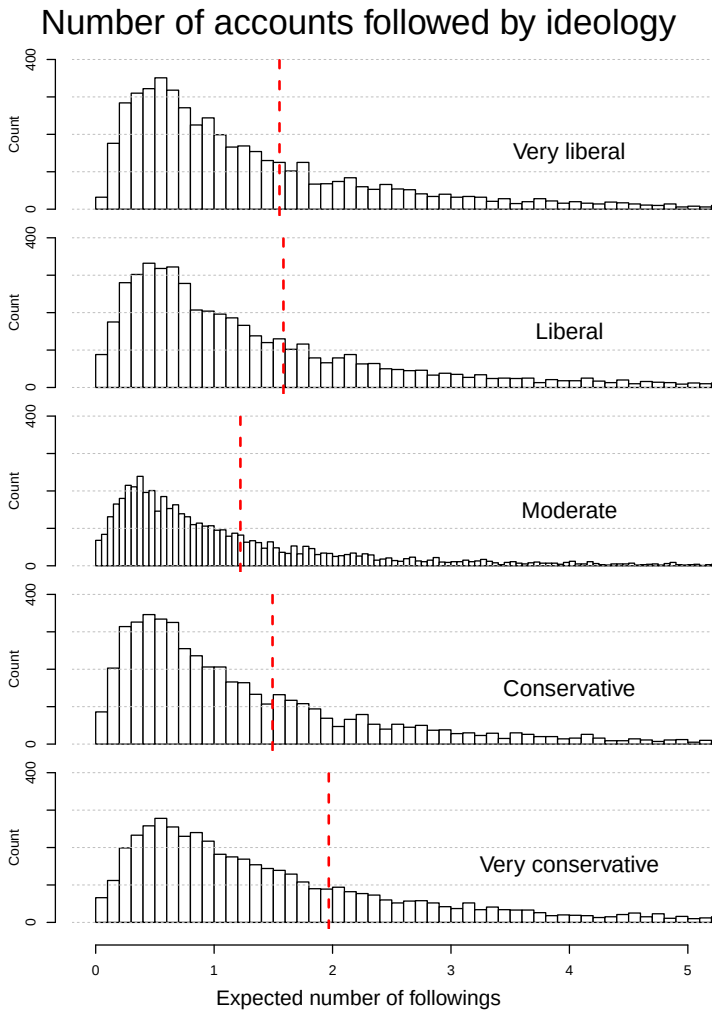
where $\Pr(n_i = 0)$ is modeled as a logistic regression, and $\mathrm{E}\left(n_i = k | n_i > 0\right)$ follows a negative binomial distribution, where I have omitted the condition on the covariates to simplify the notation.

The main results are shown in Table VII. The first column shows the determinants of the number of followed accounts and the second equation shows the model generating the zero inflation. The model includes the same covariates used in the previous section. Unsurprisingly, the inflation equation shows a similar structure to what was already reported.

Table VII: Expected number of elite accounts followed

	Count model			Zeros model		
Variable	Coef.	2.5%	97.5%	Coef.	2.5%	97.5%
(Intercept)	0.324	-0.689	1.339	-4.392	-7.380	-1.403
Age: 1st piece	0.495	-1.584	2.575	-0.028	-5.900	5.843
Age: 2nd piece	-0.216	-1.037	0.604	0.057	-3.866	3.980
Age: 3rd piece	-0.506	-1.733	0.720	-4.978	-11.500	1.543
Ideology: Liberal	0.050	-0.217	0.317	2.087	0.432	3.742
Ideology: Moderate	-0.191	-0.481	0.097	2.679	1.046	4.313
Ideology: Conservative	-0.012	-0.285	0.261	2.080	0.445	3.715
Ideology: Very conservative	0.261	-0.050	0.572	2.214	0.486	3.942
Ideology: Not sure	-1.127	-1.894	-0.360	1.766	-0.745	4.278
Gender: Female	-0.154	-0.350	0.040	1.210	0.503	1.917
Education: Less than High school	-0.072	-0.898	0.753	-0.928	-3.954	2.096
Education: High school graduate	-0.055	-0.463	0.352	1.411	0.266	2.556
Education: Some college	0.544	0.225	0.862	1.030	0.025	2.036
Education: 4-year college	0.276	-0.028	0.581	0.513	-0.513	1.540
Education: Postgraduate degree	0.220	-0.094	0.536	-0.185	-1.467	1.097
News interest: Some of the time	-1.225	-1.535	-0.915	1.375	0.653	2.097
News interest: Only now and then	-1.047	-1.701	-0.392	2.827	1.809	3.846
News interest: Hardly at all	-3.620	-4.824	-2.417	-1.536	-10.400	7.327
Race: Black or African-American	-0.427	-0.777	-0.077	-0.127	-1.247	0.992
Race: Hispanic or Latino	0.674	0.220	1.129	0.567	-0.373	1.507
Race: Asian or Asian-American	0.149	-0.855	1.155	1.529	-0.299	3.357
Race: Other	0.042	-0.353	0.439	0.442	-0.620	1.506
$\log(\theta)$	-1.577	-1.703	-1.450			
N			5423			
AIC			10332			
Vuong statistic			4.666			

Figure 4: Effect of news interest in the number of accounts followed.



The process generating the zeroes in the our data is driven by news interest and the political ideology of the respondent, with a significant effect of gender and some categories of education. Note that a Vuong test comparing a negative binomial with and without an inflation equation returns a test statistic of 4.6 (p-value < 0.001) therefore supporting the specification in Table VII.

The count model (first column in Table VII) tells a story in which *the number* of political accounts the user follows is mostly an effect of an interest for political news. As the first column of Table VII shows, while all the categories of news interest are statistically significant, for the ideology variables only "Not sure" category is. Therefore, once the user has decided to follow political accounts, the number of accounts actually followed depends mostly on her interest in government and political affairs and not on her political preferences.

Figure 4 shows the substantive effect of news interest in the number of accounts followed. In the figure, I plot the predictive distribution of the model for an individual with different values of news interest. The vertical dashed line marks the median of the distribution. It can be seen the big jump that happens between the "most of the time" and the "some of the time" categories, with the former group following three times as many accounts (1.02 and 0.29 accounts, respectively). As expected, people with no interest in political news are not expected to follow any Member of Congress. Also, while there is a difference between very liberals/conservatives and moderates, the size of the difference is remarkably smaller. In particular, using Table VII and an average baseline of the sample, respondents identifying as very liberal follow 1.01 accounts, moderates follow 0.79, and very conservatives 1.24.

In summary, the number of political accounts followed is mostly an effect of news interest and ideological attachment, with the former having a larger effect than the latter. As a consequences, this differential contributes to the bias that is already produced by the process of following political accounts in the first place.

V. Conclusions

Social media enables a direct communication between politicians and voters without the brokerage of traditional media and in real time. It allows individuals to fine tune the sources they want to be exposed to, while politicians benefit from an unexpensive platform that reachers a large proportion of the electorate. As a consequences, social media has the ability to increase the enagement in horizontal and vertical political discussions, which may positively affect the individual propensity for political participation (Gil de Zúñiga et al., 2012; Gainous et al., 2013; Valenzuela et al., 2014). However, political communication in social media is inserted into the same biases that affect all other media-related activities: users select sources based on ideological proximity, shielding them from different opinions and reducing their exposure to opposing views.

This article advances in the understanding of the following behavior in social media by exploiting a unique database of Twitter users for whom we know their political attitudes and

demographic profile from surveys. In particular, the dataset allows a direct validation test of the spatial model of following behavior (Barberá, 2015a). Consistent with the theoretical expectations, I find a strong negative effect of ideology as a predictor of subscribing to a politician's feed, which supports the logic of a selection of content based on proximity.

However, the model faces a number of very relevant challenges for its application to recovering the ideology of Twitter users. In particular, it understates differences within groups in either side of the ideological scale and doe not separate clearly between hard- and soft-liners. In addition, it misrepresents the moderate group, partly because moderates are less likely to follow political accounts in the first place, and also because moderates are more likely to be incorrectly classified. Therefore, while the essential logic of the spatial model works, the limitation of using only people following political accounts induces it to exaggerate the polarization that we actually observe in society.

In spite of its limitations, the model gets us closer to the goal of being able to take advantage of real-time streams of data to gauge public opinion. By recovering attitudes from user behavior, the model is another step to programatically limit the effect of those with most extreme attitudes who also tend to be the most active participants in the political conversation on social media (Conover et al., 2012; Barberá and Rivero, 2014).

References

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 US election: Divided they blog. *Proceedings of the 3rd international workshop on Link discovery*, 36–43.

Arceneaux, K., M. Johnson, and C. Murphy (2012). Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics 74*(1), 174–186.

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

Barberá, P. (2015b). How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. Unpublished manuscript.

Barberá, P. and G. Rivero (2014). Understanding the political representativeness of Twitter users. *Social Science Computer Review XX*(XX), xxx–xxx.

Bennett, W. L. and S. Iyengar (2008). A new era of minimal effects? The changing foundations of political communication. *Journal of Communication 58*(4), 707–731.

Bode, L., D. Lassen, Y. M. Kim, D. Shah, and E. F. Fowler (2011). Social and broadcast media in 2010 midterms: The expanding repertoire of Senate candidates' campaign strategies. Annual Conference of the American Political Science Association.

Bond, R. and S. Messing (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review 109*(01), 62–78.

Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Boutet, A., H. Kim, and E. Yoneki (2012). What's in your tweets? I know who you supported in the UK 2010 General Election. *ICWSM*.

Ciot, M., M. Sonderegger, and D. Ruths (2013). Gender inference of Twitter users in non-English contexts. *Conference on Empirical Methods on Natural Language Processing*, 1136–1145.

Conover, M., J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini (2011). Political polarization on Twitter. *ICWSM*.

Conover, M. D., B. Gonçalves, A. Flammini, and F. Menczer (2012). Partisan asymmetries in online political activity. *EPJ Data Science 1*(1), 1–19.

Conover, M. D., B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer (2011). Predicting the political alignment of Twitter users. *IEEE Third International Conference on Social Computing*, 192–199.

Cormack, L. (2013). Congressional communication: The ideological nature of official e-mail. Unpublished manuscript.

Diermeier, D., J.-F. Godbout, B. Yu, and S. Kaufmann (2012). Language and ideology in Congress. *British Journal of Political Science 42*(1), 31–55.

Druckman, James N., M. J. K. and M. Parkin (2010). Timeless strategy meets new medium: Going negative on Congressional campaign Web sites, 2002–2006. *Political Communication 1*(27), 88–103.

Ecker, A. (2015). Estimating Policy Positions Using Social Network Data: Cross-Validating Position Estimates of Political Parties and Individual Legislators in the Polish Parliament. *Social Science Computer Review XX*(XX), xxx–xxx.

Fenno, R. (1978). *Home Style: House Members in Their Districts*. Addison-Wesley Educational Publishers Inc.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford university press.

Fiorina, M. P., J. Abrams, S. Pope, and J. C. Pope (2006). *Culture War? The Myth of a Polarized America*. Pearson Education Inc.

Gainous, J., A. D. Marlowe, and K. M. Wagner (2013). Traditional cleavages or a new world: Does online social networking bridge the political participation divide? *International Journal of Politics, Culture, and Society 26*(2), 145–158.

Gainous, J. and K. Wagner (2014). *Tweeting to Power. The Social Media Revolution in American Politics*. Oxford University Press.

Gayo-Avello, D. (2012). I wanted to predict elections with Twitter and all I got was this lousy paper–A balanced survey on election prediction using Twitter data. arXiv preprint `arXiv`:1204.6441.

Gayo-Avello, D., P. T. Metaxas, E. Mustafaraj, H. S. Markus Strohmaier and, and P. Gloor (2013). The power of prediction with social media. *Internet Research 23*(5), 528–543.

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

Gil de Zúñiga, H., N. Jung, and S. Valenzuela (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication 17*(3), 319–336.

Goldschmidt, K. and L. Ochreiter (2008). *Communicating with Congress: How the Internet has changed citizen engagement*. Congressional Management Foundation.

Graham, T., M. Broersma, K. Hazelhoff, and G. van't Haar (2013). Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign. *Information, Communication & Society 16*(5), 692–716.

Greenberg, S. (2012). *Congress and Social Media*. University of Texas, Lyndon B. Johnson School of Public Affairs.

Issenberg, S. (2012). *The Victory Lab: The Secret Science of Winning Campaigns*. Broadway Books.

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

Kohut, A., C. Doherty, and M. Dimock (2012). *In Changing News Landscape, Even Television Is Vulnerable*. Pew Center for the People and the Press.

Kosinski, M., D. Stillwell, and T. Graepel (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences 110*(15), 5802–5805.

Kreiss, D. (2014). Seizing the Moment: The Presidential Campaigns' Use of Twitter During the 2012 Electoral Cycle. *New Media & Society XX*(XX).

Lassen, D. S. and A. R. Brown (2010). Twitter: The electoral connection? *Social Science Computer Review 29*(4), 419–436.

Lassen, D. S. and B. J. Toff (2015). Elite ideology across media: Constructing a measure of Congressional candidates' ideological self-presentation on social media. Unpublished manuscript.

Lenhart, A. (2009). *Teens and Mobile Phones over the Past Five Years: Pew Internet Looks Back*. Pew Internet & American Life Project Washington, DC.

Liu, W. and D. Ruths (2013). What's in a name? Using first names as features for gender inference in Twitter. *AAAI Spring Symposium: Analyzing Microtext*.

Mayhew, D. R. (1974). *Congress: The Electoral Connection*. Yale University Press.

McCarty, N., K. T. Poole, and H. Rosenthal (2006). *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

Messing, S. and S. J. Westwood (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research 41*(8), 1042–1063.

Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011). Understanding the demographics of Twitter users. *International AAAI Conference on Web and Social Media 11*, 5th.

Nguyen, D., R. Gravel, D. Trieschnigg, and T. Meder (2013). "How old do you think I am?:" A study of language and age in Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Nickerson, D. W. and T. Rogers (2014). Political campaigns and big data. *The Journal of Economic Perspectives 28*(2), 51–73.

Otterbacher, J., M. A. Shapiro, and L. Hemphill (2012). Tweeting vertically? *International Conference on e-Democracy and Open Government-Asia*.

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin.

Pennacchiotti, M. and A.-M. Popescu (2011). Democrats, Republicans and Starbucks aficionados: User classification in Twitter. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 430–438.

Perrin, A. (2015). *Social Media Usage: 2005-2015*. Pew Internet & American Life Project Washington, DC.

Peterson, R. D. (2012). To tweet or not to tweet: Exploring the determinants of early adoption of Twitter by House members in the 111th Congress. *The Social Science Journal 49*(4), 430–438.

Poole, K. T. and H. Rosenthal (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press.

Poole, K. T. and H. L. Rosenthal (2011). *Ideology and Congress*. Transaction Publishers.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science 16*, 101–127.

Rainie, L., A. Smith, K. Schlozman, and H. Brady (2012). *Social Media and Political Engagement*. Pew Internet & American Life Project Washington, DC.

Roback, A. and L. Hemphill (2013). I'd have to vote against you: Issue campaigning via Twitter. pp. 259–262.

Smith, A. (2014). *Cell phones, social media and campaign 2014*. Pew Internet & American Life Project Washington, DC.

Stroud, N. J. (2008). Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior 30*, 341–366.

Stroud, N. J. (2011). *Niche news: The politics of news choice*. Oxford University Press.

Sunstein, C. R. (2008). Neither Hayek nor Habermas. *Public Choice 134*(1-2), 87–95.

Toff, B. J. and Y. M. Kim (2013). Words That Matter: Twitter and Partisan Polarization. Unpublished manuscript.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review 29*(4), 402–418.

Valenzuela, S., A. Arriagada, and A. Scherman (2014). Facebook, Twitter, and youth engagement: A quasi-experimental study of social media use and protest behavior using propensity score matching. *International Journal of Communication 8*, 25.

Yardi, S. and D. Boyd (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society 30*(5), 316–327.