Routledge
Taylor & Francis Group

Contemporary Social Science

Journal of the Academy of Social Sciences

ISSN: 2158-2041 (Print) 2158-205X (Online) Journal homepage: http://www.tandfonline.com/loi/rsoc21

Preaching to the choir: ideology and following behaviour in social media

Gonzalo Rivero

To cite this article: Gonzalo Rivero (2017): Preaching to the choir: ideology and following behaviour in social media, Contemporary Social Science, DOI: 10.1080/21582041.2017.1325924

To link to this article: https://doi.org/10.1080/21582041.2017.1325924

Published online: 05 Jun 2017.

CONTEMPORARY SOCIAL SCIENCE, 2017
https://doi.org/10.1080/21582041.2017.1325924

ACADEMY of SOCIAL SCIENCES
Routledge
Taylor & Francis Group

Preaching to the choir: ideology and following behaviour in social media

Gonzalo Rivero[†]

Scientific Research Group, YouGov, Redwood City, CA, USA

ABSTRACT

Social media offers a new channel to connect political elites with potential supporters. However, the flow of information between the two groups is constrained by a deliberate decision to actively listen only to some politicians based on the messages that users potentially favour. Previous research has posited that, if ideological proximity is an explanatory factor of the decision to follow certain political elites in social media, it should be possible to recover the political preferences of users from following behaviour. Using a unique database with survey data about the demographic characteristics and political attitudes of 5580 Twitter users, I show that, although ideology indeed explains following behaviour on Twitter, interest in politics induces a pattern of selection that limits what we can know about the ideological distribution in the population. My results provide direct validation of a previously hypothesised behaviour in the literature and have implications for the use of a proximity model to infer the state of public opinion using Twitter data.

ARTICLE HISTORY
Received 25 February 2017
Accepted 27 April 2017

KEYWORDS
Social media; ideology; following behaviour; polarisation

1. 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 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 socio-demographic composition of the general population (Perrin, 2015). More relevant for social scientists is the fact that social media users do not compartmentalise their political preferences and partisan attachments away from the rest of their online lives. On the contrary, a significant portion of them seem willing to be exposed to political information or engage in conversations about public affairs through the same channels they use for more personal communications, as illustrated by the fact that at least 66% of those users say that they have participated in political activities through social media (Rainie, Smith, Schlozman, & Brady, 2012).

Those numbers alone are sufficient to explain why politicians join platforms such as Twitter, Facebook, or Instagram: social media gives them access to a very large audience that can be reached at a very small cost and in real time. In addition, its horizontal structure

CONTACT Gonzalo Rivero gonzalorivero@westat.com
[†]*Current contact information:* Gonzalo Rivero, Westat, 1600 Research Blvd., Suite 455. Rockville, MD 20850, USA.

allows the elite to broadcast messages without the intermediation of traditional media brokers such as newspapers or radio stations (Bode, Lassen, Kim, Shah, & Fowler, 2011; Gainous & Wagner, 2014; Lassen & Toff, 2015).

Although the elite has obvious incentives to follow the electorate wherever they go, the question of what makes citizens follow politicians on Twitter or Facebook is still open for discussion. It would be hardly surprising for any political scientist to find that ideological proximity between the two actors increases the likelihood for the user to subscribe to the feed of a given politician (Barberá, 2015). However, *the strength* of such relationship and how relevant it is in relation to other factors require a thorough effort of measurement of the effect of political attitudes on the behaviour of users on the platform. But aside from sheer intellectual interest, the question has a practical side to it. If it is true that ideological proximity is the most relevant factor explaining following behaviour, it is conceivable to think about methods to recover the ideology of users from their behaviour by looking at whom they follow in social media. If that were to be the case and we could infer individual ideology from behaviour alone, we would then have a handle on a method to identify and correct the overrepresentation of extreme partisans in the political conversation (f.i., Conover, Ratkiewicz, et al. 2011) and therefore to make progress towards using social media to measure public opinion.

Unfortunately, the study of the determinants of following behaviour is severely limited by the lack of readily available data about the socio-demographic and political attributes of users of social media. This is particularly true in the specialised literature studying the effect of political attitudes on social media behaviour, which oftentimes has to rely on less reliable indirect validation (Barberá, 2015; Bond & Messing, 2015; Boutet, Kim, & Yoneki, 2012; Conover, Gonçalves, et al. 2011).

In this article, I present a systematic study of the decision to follow politicians on Twitter using a data set with information about the demographic characteristics and political attitudes of 5580 Twitter users. I also provide the first direct test of the expectations of the spatial model of following behaviour on Twitter (see below) using the reported ideal points of voters and the DW NOMINATE scores of Members of the 114th US Congress. This data set, which connects offline attitudes and online behaviour, is unique in the literature and offers insights into how individual political preferences and disposition towards political news have an effect on the decision to listen to political elites on social media.

Perhaps unsurprisingly, I find support for the hypothesis that ideological proximity between users and elites has a significant effect in the pool of politicians that voters directly listen to. However, my results offer a more nuanced picture that poses a challenge to our ability to reverse-engineer the political preferences of users. In particular, I show that an ideology-only model tends to misclassify users self-identified as *moderates* and largely overestimates the degree of political polarisation in the electorate. I then investigate a *two-step model* of following behaviour as an extension of the ideology framework and find evidence consistent with the notion that only the most politically motivated users on Twitter engage in political following to begin with.

The rest of the article is structured as follows. In the next section, I review the literature on how ideology affects following behaviour of accounts from the political elite and discuss the pitfalls of the spatial model of following behaviour. I also discuss some related literature that is connected to this research. I then describe the data set used in the empirical analysis in Section 3. Results are presented in Section 4. I first estimate the expected behaviour under the spatial model and then show how a two-step approach explains the empirical problems that appear in the validation of the model. The final section summarises the results and offers some conclusions.

2. Theoretical framework

2.1. Ideology and following behaviour in social media

Social media platforms such as Twitter and Facebook offer a vehicle for low-cost, direct communication with voters that can also be integrated within modern encompassing microtargeting strategies (see Issenberg, 2012; Nickerson & Rogers, 2014). Furthermore, social networking sites allow politicians to escape the gatekeeping, intermediary role of traditional media (Bode et al., 2011).

From the perspective of communication strategy, social media can still be understood within a theoretical framework in which the primary aim of politicians' communication with voters is winning elections (Fenno, 1978) using *advertising*, *credit claiming*, and *position-taking* (Mayhew, 1974). Social media allows incumbents to advertise their name and actions in office to create a favourable image and improve their name recognition with messages that may have little legislative content (Bode et al., 2011; Peterson, 2012). But they also build their social media image by taking political positions. For instance, Greenberg (2012) finds that roughly two-fifths of all tweets and Facebook posts could be classified as position-taking, making it the most common type of message in either platform, although there seems to be variations depending on whether these posts are from incumbents or challengers (Gainous & Wagner, 2014, chap. 5). 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), but also use the platform to reinforce and gain support from their base through ideological messages (Bode et al., 2011; Lassen & Brown, 2011; Otterbacher, Shapiro, & Hemphill, 2012; Roback & Hemphill, 2013).

To put it differently, the behaviour of the political elites in social media is expected to reflect their political priorities and ideology. In consequence, it is almost a truism to indicate that the interaction between users and politicians is mediated by their respective ideology, and it should be uncontroversial to say that we should expect users to behave in a manner that is consistent with a *proximity* model when it comes to the decision of whose feeds to follow in social media (Barberá, 2015).[1] It may be the case that some users will want to receive political messages for as wide a range of different politicians as possible. But, generally, users are more likely to follow political elites whom they perceive closer to them in some metric that orders individuals according to their ideology. In other words, left-wing users are more likely to follow left-wing politicians and, not only that, the more politically extreme a user is, the more likely he or she is to follow more extreme politicians instead of moderate ones. Even if it is possible that the following behaviour may simply reflect an expression of identity or attachment to a given political label, the platform still induces users to choose whose messages they are exposed to, avoiding cognitive dissonance in their feeds (Festinger, 1957).

The proximity model of following behaviour is attractive because it relates to similar theories in other areas in political science, namely, the spatial model of voting behaviour

– a landmark for the study of individual decision-making, party competition, and political representation – which posits that voters are more likely to pick parties that better reflect their own political ideas (Downs, 1957). In addition, the model is consistent with previous research on ideological sorting in online networks (Adamic & Glance, 2005; Conover, Ratkiewicz, et al. 2011), that is, individuals tend to follow and interact only with users who share their same political views (Conover, Ratkiewicz, et al. 2011; Yardi & Boyd, 2010), even if we account for the fact that the horizontal nature of social media is also likely to expose individuals to information that can potentially contradict the beliefs of the user (see Bond & Messing, 2015; Messing & Westwood, 2014).[2]

2.2. Extending the proximity model

The elegance of the proximity model comes at the price of not being able to explain the decision to follow politicians or politics-related accounts in the first place. The model can only tell us whom the user is likely to follow if they indeed choose to include politicians in their feeds, but it does not answer how the first part of the problem for the user (i.e. whether to follow any politicians or not) is resolved. Of course, it does mean that the proximity model is an incomplete theory – after all it produces an answer for the question it is designed to tackle – but one must acknowledge that it produces a partial view of the following behaviour that limits its applicability, especially if one thinks about the possible practical implications of the model.

If the assumption that following behaviour is driven by ideology is correct, an analyst should then have the ability to recover the ideology of *both* users and politicians on a comparable scale by observing *only* who follows who (Barberá, 2015). In practical terms, it means that we would now have the crucial piece that was needed in order to use social media data for the study of public opinion. If the fact that ideologically extreme users are overrepresented in political conversations (Conover, Gonçalves, et al. 2011) is one of the main source of bias[3] that hinders our ability to use Twitter as a source to measure public opinion (Gayo-Avello, 2012), the proximity model offers us the tools to first estimate the ideology of each user on Twitter and then adjust the user weight to reflect the prevalence in the population of each ideological group – using the same methods that are common in non-probability surveys (see Dorfman & Valliant, 2005; Rivers, 2006 or Wang, Rothschild, Goel, & Gelman, 2015, for alternative approaches).

The easiest, most obvious way to extend the proximity model is to hypothesise that the following decision is mediated by either the users' interest in politics or the strength of their ideological convictions. Even if users resort to their ideology to decide which politicians to follow, it is reasonable to expect that only motivated users with high interest for government affairs will follow political elites in social media. Therefore, the following behaviour of politicians and political elites on social media would conform to a *two-step model,* in which the user's interest triggers the decision to listen to politicians on social media, and then proximity in the ideological space informs the selection of politicians who are finally included in the feed. In consequence, a potentially large group of users – those who are not interested in politics, regardless of their ideology – will prefer to not follow any politician whatsoever.

In the following section, I empirically study the relation between ideology and following behaviour in social media from the perspective of the *two-step model* by exploiting a unique data set of 5580 Twitter users for whom I have survey data about their demographic and sociopolitical attitudes. The richness of information contained in this data set allows me to estimate individual-level models for their decision funnel that leads users to follow accounts from Members of the US Congress.

3. Data

The data for this study were collected from the database of panelists of YouGov. The YouGov US panel, a proprietary opt-in survey panel, comprises US 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 go through 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 socio-economic demographics, political attitudes and behaviour, health status, and consumer behaviour. Participants are not paid to join the YouGov panel, but they receive incentives through a loyalty programme to take individual surveys.

The construction of the data set used in this article proceeded in two stages. First, a 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 with a question in which they were asked to provide their Twitter handle in an open-text box. Those handles were validated using the Twitter API. For set the validated users, I collected the list of people they followed. A total of 5580 YouGov panelists provided valid Twitter handles.

The second stage consisted on retrieving user data for the YouGov panelists who provided valid 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 socio-demographic data and a small battery of 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 user data is presented in Table 1.[4]

Finally, I built a data set on political elite accounts, which I restricted to members of the US Congress (US House of Representatives and Senate) as of 1 August 2015. The decision to limit the set of politicians and political actors to Members of Congress was driven by the availability of ideal point estimates from the DW NOMINATE database of roll-call voting behaviour (Poole & Rosenthal, 2000). The data set is standard in the political science

Table 1. Summary statistics of the data set.

	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

literature. Each elite Twitter account was matched against its corresponding ideological ideal point from the DW NOMINATE scores for the 114th Congress. Note that by using only accounts from Members of Congress, my sample of political actors resembles that of Barberá (2015), although I have not included accounts from political parties, or media outlets and journalists who tweet about politics.

4. Results

4.1. Who to follow?

Let me start with a description of the data set that gives us the stylised facts that the more sophisticated models below will try to explain more carefully. Table 2 shows the proportion of individuals by self-reported ideology who follow any Member of Congress in my data set. We see that 27.4% of the respondents who consider themselves to be very liberal follow at least one Representative or Senator from the Democratic Party, while 30.5% of the very conservatives follow one from the GOP.[5] Similarly, only 9.2% of the moderates follow a Democrat and 8.3% follow a Republican. There are two things to notice in this table: First, that even within the very ideological group, only a minority follows Members of Congress; second, that, as one would expect, individuals from the extremes of the ideological scale are more likely to follow their own party. In fact, only a small percentage of them are exposed to messages from the other side of the aisle. In addition, moderates are less likely to follow anyone and they are equally likely to follow accounts from either party. It is also interesting to note how these numbers are similar to the 20% of social media users who reported following elected officials and candidates for office in other studies (Rainie et al., 2012).

This notion of following behaviour through ideological proximity is even clearer in Table 3, 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 that an individual follows. The main message is that only around 4% of respondents are exposed to voices from ideological groups, and the large majority prefers to isolate themselves from other political positions. Interestingly enough, and contrary to the overall expectation of the proximity model, it is those in the *extremes* who are just slightly more likely to follow both parties. This fact is likely explained by a correlation between ideological extremism and motivation to read about public affairs, which may

Table 2. 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

Table 3. 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

push those in the extremes rather than the moderates to seek information about what the other group is saying. However, the proximity model would have expected, independently of other covariates, moderate users to be listening to a mix of opinions.

In any case, the results in the previous tables are not new and they should not come as a surprise: individuals avoid dissonant messages in their Twitter feeds and only a small proportion of individuals decide to actively listen to politicians from their least preferred party. The tables also suggest that it is possible to go in the opposite direction to recover the political affinity of a given individual by looking at the politicians a given user follows on Twitter. Not only that, given that we can order politicians by their relative ideological position (we can make statements such as 'Congressperson A is more liberal than Congressperson B who is more liberal than Congressperson C'), we should also be able to produce a fine-grained estimate of the user's ideology that goes beyond labelling users as 'liberal' or 'conservative'. However, before taking that step, we first need to verify that ideology is a predictor of following behaviour and then test how the predictions of a proximity-based model compare against some of the ground truth. These are the two tasks that this section tries to accomplish.

In order to do a first test of the proximity model of behaviour, I follow the specification from Barberá (2015) and assume that the individual's decision to subscribe to a given politician's account is a function of the distance between the politician's ideal point and her own. In consequence, my empirical 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} the dependent variable is a Boolean variable for whether respondent i follows the account from Member of Congress j, α_i is a panelist-specific random effect that captures the propensity of i to follow on Twitter, η_j is an elite-specific random effect that measures the differential popularity of the 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. Therefore, model (1) makes explicit the relation between proximity and following behaviour and allows us to recover the position of the individual.

I ran two different versions of the model in Equation (1), 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 the Member of Congress j. However, the first model used the answer to the ideology question on a 5-point scale for each of the YouGov panelists (see Table 1) as the ideal point for the respondent. That is, in the first model, the distance between the politician and the respondent is assumed to be known, and the concern is to recover the parameter that links distance with following behaviour. The results for this model are shown in the top block of Table 4. The second model, shown in the lower block of Table 4, replicates the original specification in Barberá (2015) and assumes that θ_i is in fact a parameter to be estimated from the data. The second model will, therefore, output the prediction of ideology at the individual level, which we can then use to compare with the raw data from YouGov. Before moving to the results, it is 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.

Table 4. Estimation of the decision to follow a given Member of Congress.

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

Table 4 shows the parameter that captures the effect of ideological distance on the probability of following, and also the hyper-parameters for the individual- and politician-specific random effects. The main message from the models is clear: the effect of distance on the probability of following is negative and statistically significant in both Models 1 and 2, that is, regardless of whether we take the individual responses as input or whether we simply impose the behaviour model on the data set. In itself, this result constitutes direct, primary evidence in favour of the proximity model, at least as a description of following behaviour on Twitter.

However, a significant coefficient is weak evidence. We need to check the performance of the model by comparing what the model predicts as individual ideology against what the respondents told us. Figure 1 shows a scatterplot of the survey responses to the ideology question for the respondents in Model 2 against the predictions of Model 2. The blue

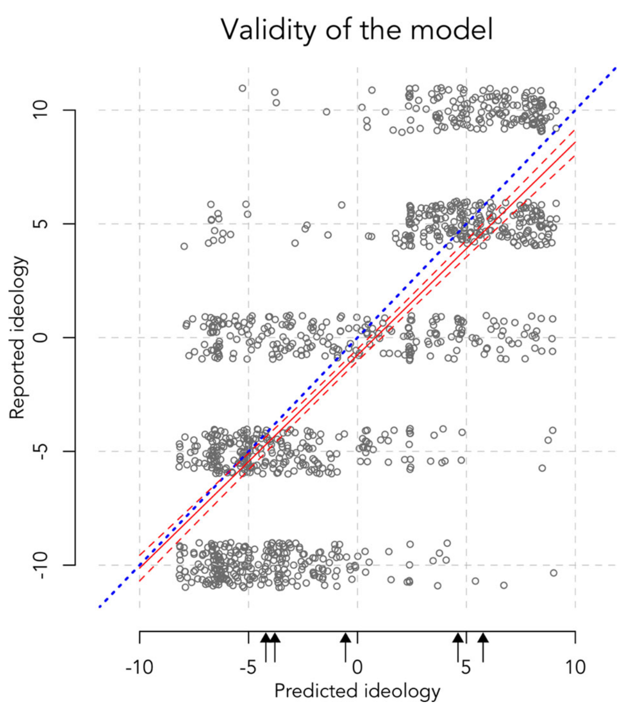

Figure 1. Predictions of the model (overall validity).

dashed line represents a 45° line and the solid red line the estimated regression of observed and predicted values. 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 the same general structure of the ideological distribution in the survey data. As a matter of fact, the model does an excellent job at sorting individuals who identify with labels other than 'moderate.'

But there are some obvious issues with the predictions. As it can be readily seen, Model 2 predicts moderates in a wide range of positions. In particular, as Table 5 shows, Model 2 does a far better job at recovering the ideology of people on the extremes than that of people on the centre. If we look at the standard deviation of the predicted ideology for people who consider themselves moderates, we see that it is about 1.7 times larger than the standard deviation for very liberals or very conservatives. However, the model makes this error with very high conviction.

A second problem is related to the population-wide estimate of polarisation. As Figure 2 shows, the ideological distribution of the full sample (left panel) shows a clear bell shape. The distribution is similar to what is commonly found in other studies, with more people displaying moderate opinions surrounded by a smaller proportion of individuals in either extreme. However, if we look at the sample of individuals who follow at least one Member of Congress (centre panel), that is, the sample on which we can make predictions using the proximity model, we clearly see that the distribution of ideology becomes almost flat. This is a natural consequence of, as we saw in Table 1, moderates being far less likely to follow politicians that those in the ideological extremes. Not only that, if we move to the next stage and look at the predicted distribution from Model 2 (right panel), we observe that

Table 5. 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

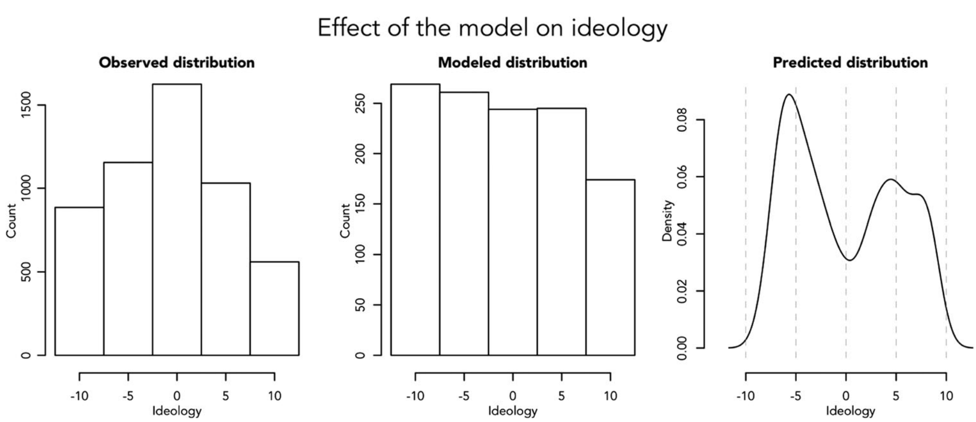

Figure 2. Estimated distribution of ideology.

it is strongly bimodal, even when the original data that enter into the model are approximately uniform. What we have seen in the process is that in trying to recover ideology from the population, we have (1) reduced the size of the group of moderate individuals because it is less likely to listen to politicians on Twitter and (2) exaggerated the degree of polarisation in the population because the model tends to push moderates to one of the ideological sides.

There are two ideas to highlight from the discussion above. On the one hand, a positive message: we have direct evidence in favour of the spatial model of behaviour on Twitter and the notion that users expose themselves only to like-minded politicians. On the other, an observation that the vanilla spatial model induces systematic biases in our attempt to recover individual- and group-level estimates. In the next section, I explore the most obvious candidate as an explanation for the problems of the model, following the discussion in Section 2. In particular, I study how political interest mediates the decision to follow any account from the political elite and, relatedly, the number of accounts to follow.

4.2. Follow any accounts?

I start the analysis by looking at the probability with which Twitter users follow accounts from Members of Congress. Let y_{ij} be an 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:

$$Pr(y_{ij} = 1) = \text{logit}(x_i^T \beta), \tag{2}$$

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

The main results are shown in Table 6. The model uses the demographic and political information available for each respondent and removes the few individuals for whom 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 also transformed with a cubic B-spline to capture non-linearities in the following behaviour.[6]

Results depict an image that is similar to the intuitions that are commonly found in the literature about social media and politics (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 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 data set, we can observe a 16% difference between users with less than a high school diploma and users with postgraduate education, although the only significant variable in Table 6 is the indicator for high school graduates. Interestingly enough I do not find an 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 as younger users to follow Members of Congress on Twitter. Finally, the data clearly show a difference between whites and minorities. Even if minorities may be overrepresented on Twitter (Mislove et al., 2011), in my data set whites are 7% more likely to follow Members of Congress than any minority.

But more importantly, I find a remarkable effect of the variable measuring political interest. The difference between not reading political news at all and following them most of

Table 6. Probability of following any elite accounts.

	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%	

the time changes the probability of following political accounts on average by around 14% (29% vs. 15%). When compared to the effect of ideology on the same variable, we see that, although individuals at either extreme of the ideological scale are indeed the most likely to follow elite accounts (Figure 3), the difference in effect between a moderate and an extremist is smaller than the effect of being interested versus not being interested in politics.

There are two additional consequences to this result. First, it seems that part of the reason the model in Section 4.1 overestimates the level of polarisation has to do with

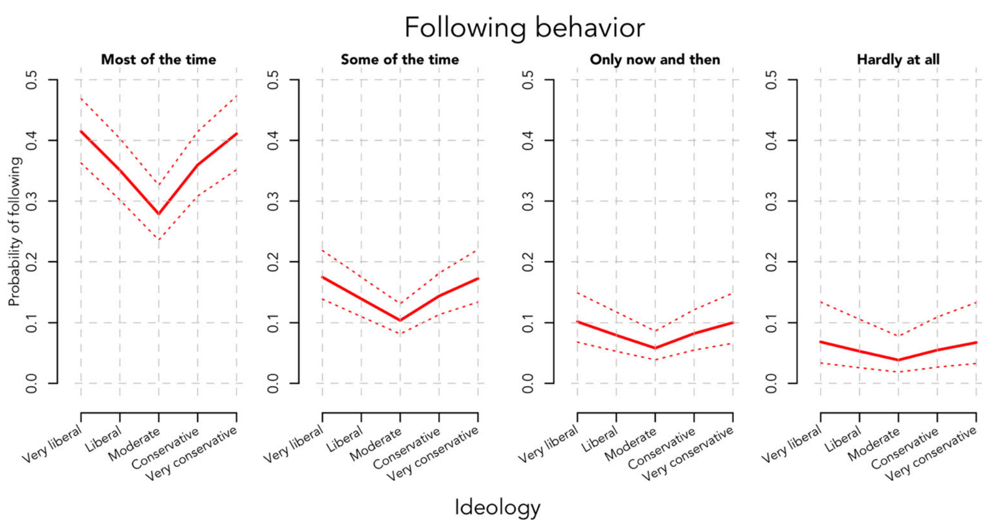

Figure 3. Predicted probability of following a Member of Congress by ideology and news attention.

the fact that it eliminates moderates from the sample as they are less likely to follow politicians to begin with. Finally, the results provide some different support to the result that the most active people in the political debate on Twitter actually come from the political extremes and not the centre (Barberá & Rivero, 2015; Conover, Ratkiewicz, et al. 2011; Conover, Gonçalves, Flammini, & Menczer, 2012). In this particular case, this result causes obstacles for the spatial model to be able to recover the distribution of ideology.

4.3. How many accounts to follow?

Given the results in the previous section, now the question is about whether this bias is accentuated or attenuated by the behaviour *once engaged* in political following on 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 a respondent follows. The model was specified as a zero-inflated binomial regression model:

$$E(n_j = k) = Pr(n_i = 0) + Pr(n_i > 0)E(n_i = k|n_i > 0), \qquad (3)$$

where $\Pr(n_i = 0)$ is modelled as a logistic regression, and $E\ (n_i = k|n_i > 0)$ 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 7. 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.

Table 7. Expected number of accounts followed.

	Count model			Zeros model		
	Coef.	2.5%	97.5%	Coef.	2.5%	97.5%
	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
	−1.577	−1.703	−1.450			
N			5423			
AIC			10332			
Vuong statistic			4.666			

Unsurprisingly, the inflation equation shows a similar structure to what was already reported. The process generating the zeroes in 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 < .001), therefore supporting the specification in Table 7.

The count model (first column in Table 7) 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 7 shows, while all the categories of news interest are statistically significant, for the ideology variables only the 'Not sure' category is. Therefore, once the user has decided to follow political accounts, the number of accounts actually followed depends mostly on the user's interest in government and political affairs and not on the user's political preferences.

Figure 4 shows the substantive effect of news interest on 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 7 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 consequence, this differential contributes to the bias that is already produced by the process of following political accounts in the first place.

5. 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 inexpensive platform that reaches a large proportion of the electorate. As a consequence, social media has the ability to increase the engagement in horizontal and vertical political discussions, which may positively affect the individual propensity for political participation (Gainous, Marlowe, & Wagner, 2013; Gil de Zúñiga, Jung, & Valenzuela, 2012; Valenzuela, Arriagada, & Scherman, 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 the understanding of the following behaviour 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 data set allows a direct validation test of the spatial model of following behaviour. Consistent with the theoretical

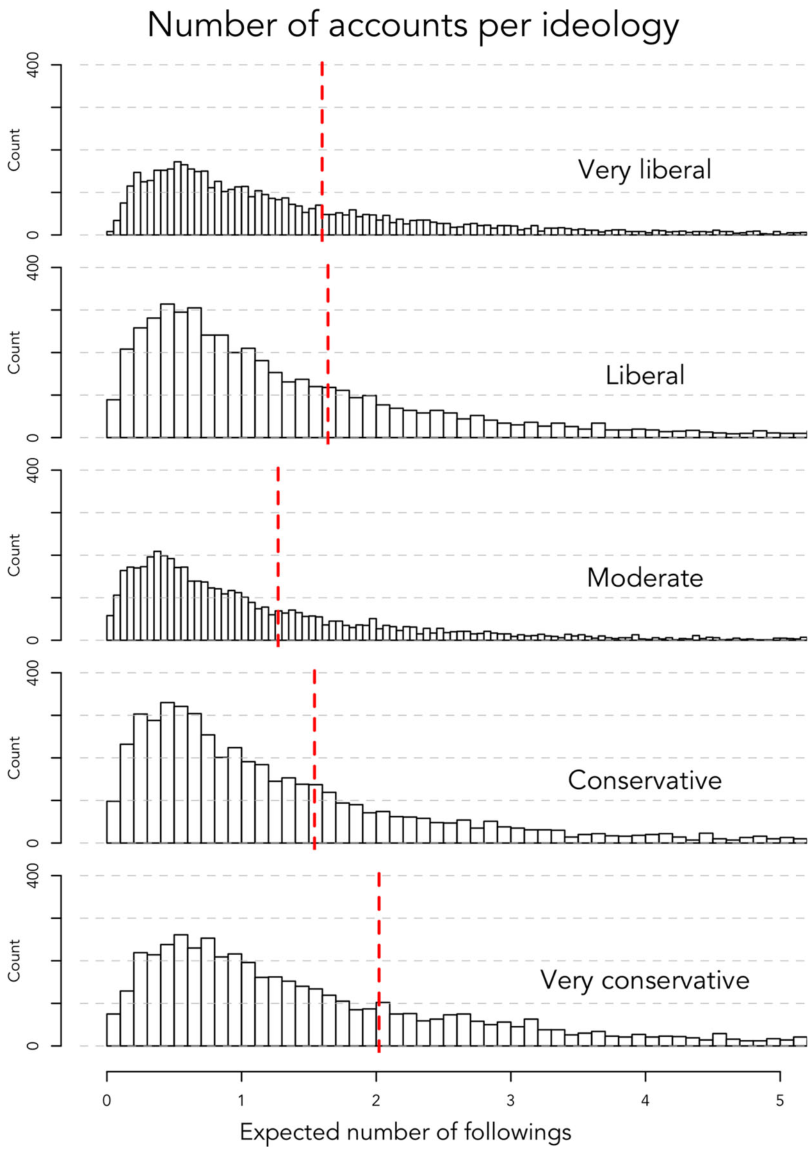

Figure 4. Effect of news interest on the number of accounts followed.

expectations, I find a strong negative effect of ideology as a predictor of subscribing to a politician's feed, which supports the logic of 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 on either side of the ideological scale and does 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 polarisation 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 behaviour, the model is another step to programmatically 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.

Notes

1. A similar idea is pursued by Bond and Messing (2015) in their study of endorsements on Facebook.
2. Not surprisingly, 38% of social media users report to have discovered through that channel that their friends' political beliefs were different from what they thought (Rainie et al., 2012).
3. Other biases, like the difference in demographic composition of Twitter users with respect to the general population, can be corrected using similar methods.
4. 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.
5. Note that Independents in the 114th Congress caucus with the Democrats.
6. 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.

Notes on contributor

Gonzalo Rivero is the senior data scientist of the Statistics and Survey Methodology Unit at Westat. He is interested in the limits of political accountability when voters are faced with multiple, conflicting dimensions, and in how social media amplifies the offline biases of political participation. His research has appeared in *Electoral Studies*, *Political Science Research and Methods*, and the *European Journal of Political Economy*. The research was done while the author was a member of the Scientific Research Group at YouGov.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

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

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729.

Bode, L., Lassen, D., Kim, Y. M., Shah, D., & Fowler, E. F. (2011). *Social and broadcast media in 2010 Midterms: The expanding repertoire of senate candidates' campaign strategies*. Unpublished manuscript. Georgetown University.

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

Boutet, A., Kim, H., & Yoneki, E. (2012). *What's in your tweets? I know who you supported in the UK 2010 General Election*. Sixth International AAAI Conference on Weblogs and Social Media. Dublin: AAI.

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

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

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). *Political polarization on Twitter*. ICWSM. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847

Dorfman, A., & Valliant, R. (2005). Superpopulation models in survey sampling. *International Encyclopedia of Statistical Science*, *8*, 1575–1577.

Downs, A. (1957). *An economic theory of democracy*. New York: Harper.

Fenno, R. (1978). *Home style: House members in their districts*. Boston, MA: Little, Brown and Company.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Gainous, J., Marlowe, A. D., & Wagner, K. (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., & Wagner, K. (2014). *Tweeting to power. The social media revolution in American politics*. New York: 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 paper 1204.6441.

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

Greenberg, S. A. (2012). *Congress and social media*. Austin: University of Texas, Lyndon B. Johnson School of Public Affairs.

Issenberg, S. (2012). *The victory lab: The secret science of winning campaigns*. New York: Broadway Books.

Lassen, D. S., & Brown, A. R. (2011). Twitter: The electoral connection? *Social Science Computer Review*, *29*(4), 419–436.

Lassen, D. S., & Toff, B. J. (2015). *Elite ideology across media: Constructing a measure of congressional candidates' ideological self-presentation on social media*. Unpublished manuscript. University of Wisconsin-Madison.

Mayhew, D. R. (1974). *Congress: The electoral connection*. New Haven: Yale University Press.

Messing, S., & Westwood, S. J. (2014). 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., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). *Understanding the demographics of Twitter users*. International AAAI Conference on Web and Social Media 11: 5th.

Nickerson, D. W., & Rogers, T. (2014). Political campaigns and big data. *Journal of Economic Perspectives*, *28*(2), 51–74.

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

Perrin, A. (2015). *Social media usage: 2005–2015*. Washington, DC: Pew Internet & American Life Project.
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., & Rosenthal, H. (2000). *Congress: A political-economic history of roll call voting*. New York: Oxford University Press.
Rainie, L., Smith, A., Schlozman, K. L., & Brady, H. (2012). *Social media and political engagement*. Washington, DC: Pew Internet & American Life Project.
Rivers, D. (2006). *Sample matching: Representative sampling from Internet panels*. Polimetrix White Paper Series.
Roback, A., & Hemphill, L. (2013). *I'd have to vote against you: Issue campaigning via Twitter*. Proceedings of the 2013 conference on computer supported cooperative work companion: 259–262.
Valenzuela, S., Arriagada, A., & Scherman, A. (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*, 2046–2070.
Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.
Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327.