Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data.*

Pablo Barberá†
Richard Bonneau
Patrick Egan
John T. Jost
Jonathan Nagler
Joshua Tucker

Prepared for delivery at the 2014 Annual Meeting of the American Political Science Association, August 28-31, 2014.

Abstract

Are legislators responsive to their constituents in their public communication? To what extent are they able to shape the agenda that the mass public cares about, as expressed by the issues they discuss? We address this twofold question with an analysis of all tweets sent by Members of the U.S. Congress and a random sample of their followers from January 2013 to March 2014. Using a Latent Dirichlet Allocation model, we extract topics that represent the diversity of issues that legislators and ordinary citizens discuss on this social networking site. Then, we exploit variation in the distribution of topics over time to test whether Members of Congress lead or follow their constituents in their selection of issues to discuss, employing a Granger-causality framework. We find that legislators are responsive in their public statements to their constituents, but also that they have limited influence on their followers' public agenda. To further understand the mechanisms that explain political responsiveness, we also examine whether Members of Congress are more responsive to specific constituents groups, showing that they are more influenced by co-partisans, politically interested citizens, and social media users located within their constituency.

*We thank Andrew Eggers, Dustin Tingley, Drew Dimmery, Ken Benoit, Nick Beauchamp, Gaurav Sood, Michael Lewis-Beck, and the participants at the Text as Data Workshop at LSE and the Social Media and Political Participation Conference at NYU for their helpful comments and suggestions. We also gratefully acknowledge financial support from the INSPIRE program of the National Science Foundation (Award #1248077).

†Corresponding author. Email: pablo.barbera@nyu.edu

1 Introduction

An enduring question in the study of democratic polities is the level of responsiveness of government to the preferences of the public. In order for members of a legislature to be responsive to public preferences, they need to be paying attention to the policy views and preferences of the public. In this paper we analyze whether or not they are doing that by examining whether the issues discussed by Members of Congress via Twitter are influenced by the issues discussed by their constituents on the same social networking platform.

We argue that legislators' tweets constitute a standardized representation of their political activity. Virtually all members of the U.S. Congress are active Twitter users, posting more than once per day, either in person or through their communication office. Twitter is used as an additional platform to stay in touch with their constituents, to express their policy positions, and to broadcast political messages.

The main advantage of using Twitter as a source of information about legislators' communication is that *both* Members of Congress and their constituents are present on this platform, sending tweets that follow the same format and symbolic references, and often re-posting each other's content through "retweets." This allows to examine whether legislators adapt the issues they emphasize in response to changes in their constituents' public agenda; and also whether Members of Congress influence the agenda of their constituents, by looking at whether or not the content of Tweets by members of the mass public responds to changes in the topics that Members of Congress emphasize on Twitter.

We analyze all tweets by Members of Congress over a 15 month period. Using a Latent Dirichlet Allocation model, we extract topics that represent the diversity of issues legislators discuss on this social networking site. We find that this method is able to classify legislators' tweets on a limited number of topics, with meaningful variation over time and across parties.

Next, we exploit the interactive nature of Twitter to understand to what extent legislators' expressed political agenda is affected by what their constituents discuss publicly on this same platform; and vice versa. Using tweets sent by users from different partisanship groups that follow members of Congress, we test whether longitudinal changes in the importance that legislators attribute to different issues affect voters' discussions about this same set of topics. We are able to show that Members of Congress are surprisingly responsive in their public statements to their constituents. On many issues when co-partisans among the mass public devote more twitter posts to an issue, Members of Congress follow suit. With few exceptions, the opposite is not true: Members of Congress are able to exert only a minimal influence in the political issues their followers discuss.

We then explore whether they are more responsive to specific subgroups of followers, depending on their political ideology, interest in congressional politics, and geographic location. Our results show that Members of Congress are more responsive to their co-partisans, to followers who are highly interested in congressional politics (which we call "hardcore followers"), and to their own constituents rather than the general population of followers.

The rest of the paper proceeds as follows. In section 2 we discuss the existing literature on political responsiveness and agenda control, and present our theory and hypotheses. Section 3 describes our dataset of tweets sent by members of Congress and their followers. Section 4 introduces our topic modeling method and how we apply it when using Twitter data. Results of our analysis are shown in section 6. The article concludes in 7 with a summary of findings and a list of possible paths for future research.

2 Theory

There are two distinct questions for democratic polities that our paper relates to. The first important question is that of responsiveness. Does the government respond to what members of the public want? But the second important question is that of agenda-setting. How do issues get on the agenda? If the mass public talks about something, do members of the legislature observe that and react? And do members of the public shift their concerns based on the behavior, or speech, of their legislators? There is also a more narrow question related to United States politics that we consider: who do members of the legislature listen to – co-partisans, or all voters; constituents, or the entire population of voters?

Many scholars have looked at the US legislature to see who individual legislators are responsive to: examining them primarily to see whether legislators served co-partisans, or their district more broadly (see Clinton, 2006 for a recent example). The broader question of overall government responsiveness, rather than that of individual legislator responsiveness, has most recently shifted to discussion of *whom* it is responsive to. Gilens (2012) shows that government is more responsive to the policy preferences of the wealthy than to the policy preferences of the poor. Ezrow et al. (2011), find that European parties are more responsive to shifts in the mean voter position than to changes in the preferences of their supporters. There are similarly broad and long-standing literatures on how different issues reach the political agenda (see for instance Baumgartner and Jones, 2009).

We approach the current analysis from an agnostic perspective, but with hypotheses to test. First, do members of the legislature observe, and respond, to the views of the mass public? If so, do they respond more to the views of co-partisans than of the general electorate; to their core supporters or also to the mean voter in the electorate; to their own constituents or to voters in the entire country? And, do members of the legislature have the ability to influence the agenda? Does the public follow issues that members of the legislature view as important?

We note several limitations before proceeding. We do not consider the role of the traditional mass media here. Thus it may well be that both the mass public and members of the legislature are following the mass media.[1]

[1] See Gerber, Karlan and Bergan (2009); Ladd and Lenz (2009) or Habel (2012) for examples of studies documenting media effects on voters and political actors.

3 Data

3.1 Members of Congress on Twitter

To test our hypotheses, we use tweets sent by members of the 113th House and Senate of the U.S. (2013–2014). Twitter use in Congress has increased steadily over the past few years (Golbeck, Grimes and Rogers, 2010; Chi and Yang, 2010; Shapiro, Hemphill and Otterbacher, 2012; Evans, Cordova and Sipole, 2013). Of all legislators that have served in the current Congress, around 95% of all Representatives (417 of 440)[2] and Senators (100 of 105)[3] have active Twitter accounts.[4] This proportion is similar across parties: 94% of republicans (268 of 284 serving members) and 94% of democrats (248 of 264 serving members).

The interest that Members of Congress show in using Twitter to communicate with their constituents is illustrated by the high number of tweets they send, a total of 354,860 during our period of study (from January 1st, 2013 to March 15th, 2014), which results in approximately 800 tweets per day. If we extend this period to the moment each of them created their accounts, this number increases to 894,659 tweets, with an average of 1,740 tweets per member and a median of 1,187 tweets.[5] By party, Republicans have sent a higher total of tweets (501,986 tweets vs 382,916), in part because they represent a majority of the current House, but also because they are more active, with a median of 1,296 tweets vs 1,088 tweets for Democrats.

There appears to be only mild public interest on what Members of Congress are writing on Twitter. The median Representative or Senator has 7,329 followers, although the average of followers is 24,875 due to the existence of a few outliers, such as Sen. McCain with 1,827,023 followers, Sen. Cory Booker with 1,457,320 followers, Rep. Boehner with 630,999 followers, Sen. Marco Rubio with 542,145 followers, Rep. Pelosi with 464,118 followers, Sen. Rand Paul with 411,031 followers, and Rep. Ryan with 363,765 followers. There is ample variation in their number of followers (see Figure 2 in the Supplementary Materials), which Shapiro, Hemphill and Otterbacher (2012) found to be related to ideological positions of the members, with more extreme Members of Congress having larger audiences.

Golbeck, Grimes and Rogers (2010) argue that Members of Congress use Twitter primarily to advertise their policy positions and to provide information about their activities. However, more recent studies have shown that this online tool can also be a tool for Members of Congress to be

[2] We include in our analysis Jason T. Smith (MO-8), who won a special election in June 2013 after the previous incumbent resigned; David Jolly (FL-13), who substituted Bill Young; Catherine Clark (MA-5), who substituted Edward Markey after he was elected senator; Bradley Byrne (AL-1), who substituted Jo Bonner after he resigned; and Vance McAllister (LA-5), who substituted Rodney Alexander after his resignation.

[3] We include in our analysis William Cowan, who substituted John Kerry as junior Senator from Massachusetts; Edward Markey, who substituted William Cowan after he declined to run in a special election; Jeffrey Chiesa, who substituted Frank Lautenberg as junior senator from New Jersey; and was in turn substituted by Cory Booker; and John Walsh, who substituted Max Baucus after his appointment as U.S. Ambassador in China.

[4] The list of Twitter handles of Members of Congress was collected through the New York Times Congress API and then revised for errors. All the figures in this section are reported as of March 15th, 2014.

[5] Figure A.1 in the Appendix displays the distribution of this variable, where Cory Booker represents a clear outlier with over 35,000 tweets sent (2,000 tweets since he swore in as Senator).

responsive to their constituents (Hemphill, Otterbacher and Shapiro, 2013), to exercise control of the offline and online political agenda, to interact publicly with other Representatives and Senators, and to report on their constituency service (Evans, Cordova and Sipole, 2013). Figure 1 shows examples of each of these five types of tweets.

These descriptive statistics provide preliminary evidence that tweets sent by Members of Congress can be considered a meaningful representation of how legislators communicate with their constituents: both Members of Congress themselves and their constituents show interest in this platform, and a preliminary analysis of the content of the messages they exchange shows a sophisticated use of this tool.

The dataset of tweets we use in this paper consists of all tweets sent by Members of Congress since January 1st, 2013 until March 15th, 2014. These tweets were collected by the Social Media and Political Participation Lab (SMaPP) at New York University using Twitter's Streaming API. As we show in Table 1, our dataset contains a total of 354,860 tweets.

Table 1: Number of tweets in dataset (Members of Congress)

Members of Congress	N	Average	Std.Dev.	Min	Max	Tweets
House Republicans	224	686.1	681.0	0	5432	153684
Senate Republicans	44	874.2	787.0	53	4113	38465
House Democrats	194	620.8	550.1	0	3472	120434
Senate Democrats	54	782.9	562.9	96	2761	42277
Total	517	686.4	635.7	0	5432	354860

Period of analysis: January 1st, 2013 to March 15th, 2014.

Figure 2 shows how the total number of tweets sent by Members of Congress, grouped by chamber and party, vary over time. We find that their activity increases during the legislative period and, in particular, during relevant political events, such as the State of the Union address by the President in late January of 2013 and 2014, the government shutdown in early October 2013, and the IRS scandal in May 2013.[6]

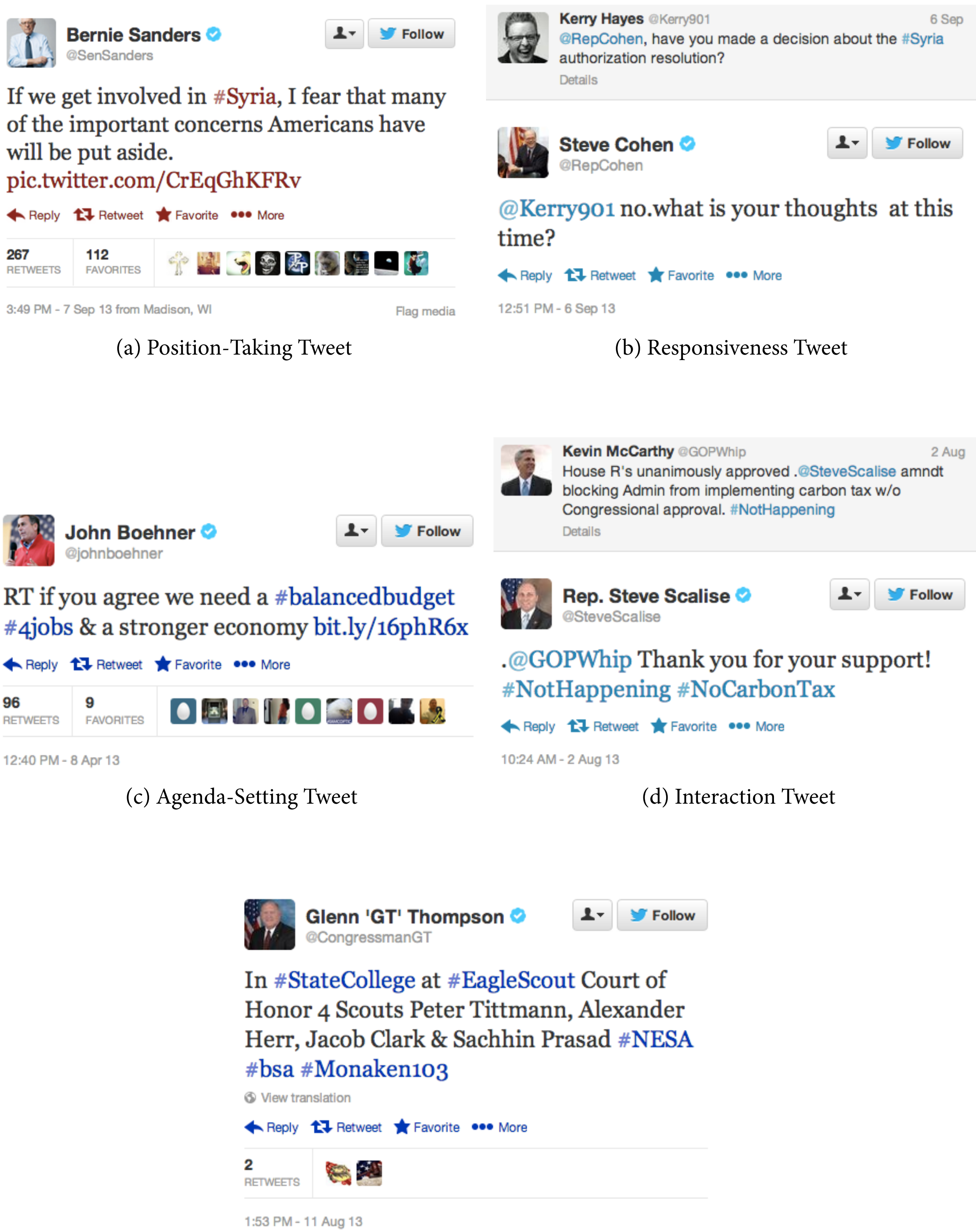
3.2 Followers of Members of Congress

In addition to tweets sent by Members of Congress, we also collected tweets sent by a random sample of their followers. This will allow us to examine to what extent changes in what ordinary citizens are discussing affect the expressed political agenda of their representatives, and vice versa.

As of March 15th, 2014, a total of 5,583,795 unique Twitter users follow at least one member of

[6] Figure A.3 in the Appendix plots the distribution of tweets sent by day, showing strong seasonality, with only 8.6% of all tweets sent during the weekends.

Figure 1: Examples of Tweets Sent by or to Members of Congress



(a) Position-Taking Tweet

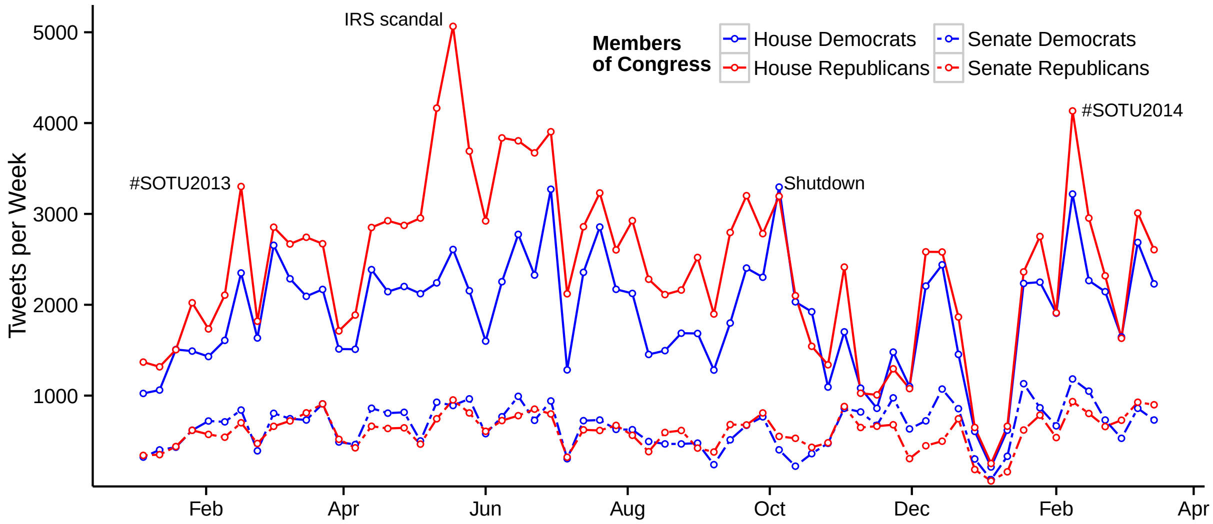
(b) Responsiveness Tweet

(c) Agenda-Setting Tweet

(d) Interaction Tweet

(e) Constituency Service Tweet

Figure 2: Number of tweets in dataset, by week



congress.[7] Of these, 3,893,168 follow only one account, 804,189 follow two accounts, 288,902 follow three accounts, 167,111 follow four accounts, 101,284 follow five accounts, and 329,141 follow more than five accounts. There are relevant differences across parties: Republicans have more unique followers than Democrats (3,443,028 vs 2,584,720 users). However, this difference is smaller if we exclude the two outliers in each party, Senators McCain and Booker: 1,861,083 unique followers of Republicans vs 1,353,451 unique followers of Democrats.

Using Twitter's REST API, we collected the user information for all 2,948,817 Twitter users who follow at least one Member of Congress.[8] Then, we applied a simple spam and location filter to ensure that all users in our random sample are active accounts located in the United States.[9] Our final sampling frame includes 866,185 users.[10]

As we show in Table 2, we subdivided the filtered population of followers in six groups, according to the number of Members of Congress they follow (from one to three, "interested" users; more than three, "hardcore" users), and their party (democrats, republicans, and independents).[11] Ad-

[7] We collected the lists of followers for all Representatives and Senators with a Twitter account using Twitter's REST API

[8] We didn't consider users who *only* follow Senators McCain or Booker, since many of them are likely to be located outside of the U.S.

[9] More in detail, we only considered users with at least 25 followers and whose most recent tweet was less than one month old. Additionally, we used a random sample of 150 million geolocated tweets captured in November, 2013 to identify users outside of the US, which we excluded from our sample.

[10] Of these, 365,543 users follow only Republicans, 297,293 users follow only Democrats, and 204,349 follow both.

[11] We assume that Twitter users who follow Members of Congress of only one party identify with that party. See Barberá (2013) for justification of why following behavior is a strong predictor of partisanship.

ditionally, we chose three groups of politically interested Twitter users who follow the New York Times, Fox News, MSNBC and CNN on Twitter, but do not follow any Member of Congress. This group will be our benchmark, which we conceptualize as active Twitter users with some interest in politics in general, but not in congressional politics. We then took a random sample of 5,000 users from each group. Finally, in order to divide our control groups according to partisanship, we applied the estimation method outlined in Barberá (2013) and classified users in the bottom 40% of the ideology distribution as democrats; those in the top 40% as republicans; and the other 20% as independents.

Table 2: Distribution of followers

Followers of Members of Congress	N	%	n	tweets
Hardcore democrats (follow > 3 Dem. MCs and 0 Rep. MCs)	32,781	3.7	5,000	6,366,549
Hardcore independents (follow > 3 MCs of either party)	119,408	13.8	5,000	6,594,749
Hardcore republicans (follow > 3 Rep. MCs and 0 Rep MCs)	74,115	8.5	5,000	4,428,818
Interested democrats (follow $1 - 3$ Dem. MCs and 0 Rep. MCs)	264,512	30.5	5,000	4,829,713
Interested independents (follow $1 - 3$ MCs of either party)	84,941	9.8	5,000	4,560,266
Interested republicans (follow $1 - 3$ Rep. MCs and 0 Dem. MCs)	290,428	33.53	5,000	4,448,887
Not following Members of Congress				
Uninterested democrats (follow no MCs; ideology on bottom 40%)	–	–	5,000	5,533,821
Uninterested independents (follow no MCs; ideology on middle 20%)	–	–	5,000	5,954,227
Uninterested republicans (follow no MCs; ideology on top 40%)	–	–	5,000	5,081,810

n = sample size of Twitter users included in analysis.

Sampling frame: users who follow at least one Member of Congress who is not Sen. McCain or Sen. Booker.

"Uninterested" users are those who do not follow a Member of Congress but do follow a major media outlet.

Our final data collection step consisted on using the Twitter REST API to capture the nearly 44 million tweets sent since January 1st, 2013 by the 45,000 users in our sample of followers. Given the rate limits of the Twitter API (only up to 3,200 recent tweets can be downloaded for each user), we conducted our data collection in two waves (September, 2013 and March, 2014), to ensure we were able to recover all tweets for users in our sample. Table 2 also includes the total of tweets for each subgroup. We find that the average user in our sample sent around 1,000 tweets in our period of study (around 72 tweets per month). Democrats appear to be more active than Republicans (between 10 and 35% more tweets), and this difference is consistent across all groups based on interest in congressional politics.

Finally, in order to test our hypothesis about legislators' responsiveness to their constituents, we tried to determine the state in which each of the users in our sample is located. We implemented a two-step process. First, we checked whether any of tweets sent by users were geolocated; if so, we assigned the user to that state. Second, for those without any geolocated tweet, we parsed the location they reported in their profiles into a set of coordinates using the Data Science Toolkit geocoder, and then assign a state based on those coordinates. This method allowed us to locate 24,321 users in our sample (54%). The number of users by state varies between 2,460 (California) and 16 (Montana).

4 Modeling Legislators' and Followers' Topic Distributions

Our purpose in this paper is to characterize the different issues that Members of Congress and their followers discuss on Twitter, and how their importance varies over time and across groups defined by their partisanship, political interest, and geographic location. To extract these categories, we estimate a probabilistic model of word occurrences in documents called Latent Dirichlet Allocation (Blei, Ng and Jordan, 2003), which belongs to a general category of latent variable models that infer topics from documents using a "bag-of-words" approach.

This approach is adequate for our application for at least two reasons. First, since it is an unsupervised machine learning technique, it is less subject to the biases inherent to imposing a set of topics pre-defined by the researcher. Second, it allows us to include the entire text of all documents in the analysis, therefore improving the validity and accuracy of our estimates for the topic distributions.

This model considers each document as a sequence of N words, denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, extracted from a vector of length V containing all possible words in the corpus. (Note that the order of words is irrelevant). LDA treats each document as a random mixture over latent topics, and each topic as a distribution over words. Each document w in the corpus is the result of the following generative model (Blei, Ng and Jordan, 2003, p.96):

1. The topic distribution for document w is determined by: $\theta \sim \text{Dirichlet}(\alpha)$
2. The word distribution for topic k is determined by: $\beta \sim \text{Dirichlet}(\delta)$
3. For each of the words in document w
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on z_n.

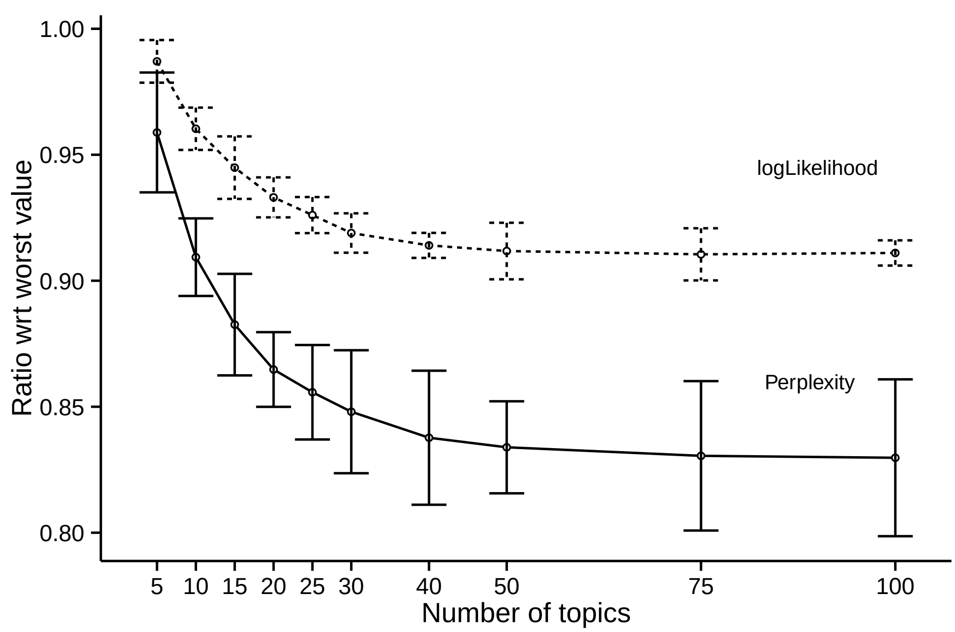
This model requires us to fix K, the number of possible topics. There are two main parameters of interest: β, a matrix of dimensions $K \times V$ indicating the distribution of words over topics; and θ, a matrix of dimensions $K \times N$ indicating the distribution of topics over documents.

Our definition of "documents" is the aggregated total of tweets sent by members of Congress each day, by party and chamber. There are two reasons for this. First, LDA assumes that each document is a mixture of topics, which is appropriate for our conceptualization of each day's tweets as the political agenda that each party within each legislative chamber is trying to push for that specific day. Second, conducting an analysis at the tweet level is complex, given its very limited length. The existing literature on topic modeling of tweets has found that applications that aggregate tweets by author or day outperform those that rely on individual tweets (Hong and Davison, 2010).

Note two additional features of our analysis. First, we fit the model at first only for Members of Congress and then, as we describe in Section 5.2, use the estimated parameters to compute the posterior topic distributions of their followers' tweets, also aggregated by day, based on their observed words. We do so to make sure the topics we estimate are political in nature, and because our main focus is language use by Members of Congress.[12] Second, in our estimation we assume that topic distributions are independent over time, and that the number and content of each topic is constant. As we discuss in Section 7, an alternative approach would be to fit a Dynamic Topic Model (Blei and Lafferty, 2006; Quinn et al., 2010).

To fix the number of topics, we ran our model multiple times with different values of K, using 10-fold cross-validation and computing the log likelihood and estimated perplexity on the holdout sample. Figure 3 reports these two measures of model fit when estimating the model with 5, 10, 15, 20, 25, 30, 40, 50, 76, and 100 topics. We find that K=50 fits the data best. A higher value of K would minimize perplexity, but we choose a conservative K in order to avoid overfitting (Hastie, Tibshirani and Friedman, 2009). We fit the model with a collapsed Gibbs sampler (Griffiths and Steyvers, 2004; Phan, Nguyen and Horiguchi, 2008), implemented in R (Grün and Hornik, 2011). We ran a single chain for 1000 iterations. We apply the usual pre-processing text techniques (converting all words to lowercase and removing stopwords, all words shorter than 3 characters, and all words that appear less than 3 times in all documents), which gives us a vocabulary of N=73,059 words.

Figure 3: LDA model fit with different number of topics



[12] We note that there is a limitation to this method. If mass followers discuss some political topics that Members of Congress never mention, we will not observe this. However, the focus on the paper is only on that Members of Congress mention on Twitter.

5 Validation of Discovered Topics

This section demonstrates that the topics that are discovered by the Latent Dirichlet Allocation model are valid representations of the political issues that legislators and voters discuss. Following Quinn et al. (2010), we discuss how our results meet different notions of validity. First, we analyze the top scoring words for each topic to demonstrate that the topics that emerge from the model have a coherent meaning (semantic validity). Then, we examine whether topic usage corresponds correctly to external events (predictive validity). As expected, we find that topic usage is coherent with party identification for both legislators and followers, and that spikes in their probability distribution can be matched to relevant political events. In Appendix D we compare our results with those from a topic model applied to legislators' speeches, finding strong support for our claim that tweets by legislators are a valid substitute for alternative sources of text used in other studies (convergent construct validity).

5.1 The Political Agenda of Members of Congress

The results of fitting our topic model are summarized in Table 3. Each row displays the top 10 words for each topic.[13] The last column displays the percentage of time each topic is used by Democrats (as opposed to Republicans), which allows us to examine what topics are more "bipartisan" and what topics are used exclusively (or "owned") by members of one party. Figure B.2 on page 31 complements our visualization of the results. Here, to facilitate the interpretation, the size of the font for each word is proportional to its normalized score, this is, an index that measures how good a word w is at predicting whether document n belongs to topic k. We also computed the specificity for each word, which indicates to what extent a word w_n is specific of a single topic k. The formulas for each of these indicators are:

$$\text{score}_{w,k} = \beta_{w,k}\left(\log \beta_{w,k} - \frac{\sum_{k'} \log \beta_{w,k'}}{K}\right)$$

$$\text{specificity}_{w,k} = \frac{\beta_{w,k}\sum_n \theta_{k,n}}{\sum_k \beta_{w,k}}$$

Table 3: Top scoring words in each topic

Topic		%D	%F
47	**Senate Republicans:** sen, senate, obamacare, mcconnell, idpol, obama, senator, kentucky, johncornyn, via	3.9	5.5
28	**House Republicans:** house, obamacare, obama, tcot, rep, president, 4jobs, hearing, will, week	4.4	4.8
17	**Obamacare Website:** obamacare, health, insurance, keep, care, website, plan, sebelius, plans, obama	14.6	1.6

[13] Due to space constraints, our figures and tables here display only the 25 most frequently used topics, after removing those that we identified as "stopword topics", whose top scoring words are not informative about politics (topics 34, 41, 45). These top 25 topics account for approximately 80% of all tweets legislators sent. Full figures and tables can be found in the Supplementary Materials. Note that our analysis in the following section does include all 50 topics.

Table 3: *(continued)*

	Topic	%D	%F
44	**Budget and Sequester (Republicans):** budget, spending, sequester, debt, president, senate, nobudgetnopay, obama, cuts, taxes	14.6	1.3
8	**Obamacare Implementation:** obamacare, delay, mandate, fairnessforall, employer, businesses, individual, another, law, senatemustact	17.3	0.8
4	**IRS scandal:** irs, targeting, groups, scandal, obamacare, benghazi, hearing, conservative, keystonexl, political	17.7	1.0
49	**Foreign Policy:** ukraine, bills, stopgovtabuse, venezuela, transparency, russia, seniors, senatorreid, sosvenezuela, report	23.6	0.7
35	**NSA Surveillance Scandal:** nsa, stopgovtabuse, abetterbargain, amendment, august, amash, privacy, surveillance, speechesdonthire, repjustinamash	32.0	0.6
14	**Republicans on TV:** game, facethenation, thisweekabc, watch, meetthepress, sunday, weekly, address, cnnsotu, tune	32.6	2.1
18	**Tax Reform:** tax, code, taxreform, taxes, taxday, cispa, budget, west, fairer, simpler	34.4	0.4
30	**Sequester and Airline Industry:** faa, life, obamaflightdelays, prolife, marchforlife, roe, furloughs, wade, delays, flight	36.7	0.5
16	**Budget Deal I:** budget, deal, christmas, holiday, agreement, fast4families, year, research, kidsfirst, tree	37.9	0.7
32	**Ronald Reagan:** reagan, prayer, ronald, presidents, presidentsday, nationaldayofprayer, breakfast, fl19, pray, internettax	38.8	0.4
36	**Benghazi Scandal:** benghazi, hearing, teacher, yourtime, flexibility, working, gopoversight, families, teachers, appreciation	39.0	0.6
48	**KeystoneXL Pipeline and weather:** cbo, tours, snow, report, keystonexl, standwithrand, pipeline, keystone, weather, due	39.7	0.5
7	**Military Intervention in Syria:** syria, military, war, action, congress, vote, weapons, use, chemical, intervention	41.0	1.1
31	**State of the Union Address:** sotu, president, tonight, obama, address, union, state, potus, hear, jobs	41.8	0.7
13	**Budget Deal II:** budget, balancedbudget, balance, tax, ryan, medicare, balanced, plan, skillsact, reppaulryan	43.3	0.7
43	**Congress Inauguration:** congress, sandy, 113th, new, debt, fiscal, sworn, spending, relief, cliff	45.4	0.9
19	**September 11:** neverforget, constitution, shall, benghazi, lost, remember, ago, constitutionday, victims, states	45.7	0.4
34	**Congress Stopwords I:** nsa, ndaa, amendment, dday, dingell, longest, john, defense, game, pjnet	45.7	0.5
37	**Birthdays and celebrations:** birthday, happy, flag, usarmy, 238th, flagday, army, ndaa, usairforce, colleague	46.6	0.5
3	**Iran Nuclear Deal:** iran, deal, nuclear, cathymcmorris, freeamir, registration, catherine, unga, voter, brynn	47.4	0.4
45	**Congress Stopwords II:** today, will, bill, house, can, now, watch, act, support, time	47.8	20.8
33	**National Holidays:** day, today, honor, veterans, happy, thank, remember, memorial, men, women	49.2	2.0
10	**Boston Marathon Attack:** prayers, thoughts, boston, victims, families, affected, today, oklahoma, safe, responders	51.5	1.1

Table 3: *(continued)*

Topic		%D	%F
15	**Death of Nelson Mandela:** mandela, nelson, world, passing, family, leader, lost, prayers, saddened, legacy	54.1	0.6
9	**Government Shutdown:** shutdown, house, government, gopshutdown, senate, end, govt, open, gop, debt	54.2	2.3
41	**Congress Stopwords III:** great, today, will, thanks, new, can, day, good, see, morning	54.5	22.5
2	**Immigration Reform:** immigration, reform, border, cir, bill, immigrationreform, amendment, timeisnow, comprehensive, broken	55.0	0.8
23	**United for Marriage Campaign:** april, marriage, monument, autism, passover, art, march, unitedformarriage, celebrating, doma	55.0	0.5
38	**Super Bowl and Death of Margaret Thatcher:** thatcher, entrepreneursday, margaret, superbowl, bowl, game, seahawks, sharon, ravens, holocaust	56.9	0.6
27	**Small Business and Farm Bill:** farmbill, farm, snap, bill, small, smallbiz, food, business, student, nsa	57.9	0.5
22	**Olympic Games:** valentine, teamusa, sochi2014, luck, whatwomenneed, olympics, debt, sochi, usa, day	58.0	0.7
26	**Holidays:** happy, year, family, day, thanksgiving, wishing, christmas, mother, holiday, everyone	59.0	1.6
1	**Mel Watt Confirmation for FHFA:** investinkids, vra, fhfa, nwlc, strongstartact, melwattnc12, 48th, fpaction, confirmed, cirmeansjobs	60.9	0.3
24	**Martin Luther King Day:** king, today, mlk, anniversary, martin, dream, president, inauguration, luther, legacy	66.5	1.0
29	**Education and Marriage Equality:** student, loan, rates, dontdoublemyrate, scotus, students, interest, loans, doma, equality	67.4	1.0
46	**Family and Medical Leave Act:** leave, fmla, family, familyact, medical, fmla20, 20th, paid, times, echovt	70.2	0.4
25	**Social Security:** socialsecurity, nhcw2013, 78th, makecollegeaffordable, social, egypt, randchat, aug, voter, mandatory	71.2	0.4
42	**Climate Change:** earthday, chicagoclimate, p2c, citizenship, art, earth, bridge, heritage, rethinktheborder, asian	74.1	0.5
40	**Gun Control:** gun, violence, background, checks, newtown, ban, guns, weapons, laws, sandy	74.5	0.9
20	**Women's Issues:** women, equal, pay, act, equalpay, work, womensucceed, every, paycheckfairness, men	76.0	0.6
50	**Poverty:** snap, enda, food, passenda, cuts, discrimination, cut, endhungernow, actonclimate, workplace	76.1	0.5
5	**Violence Against Women:** vawa, sequester, women, sequestration, violence, stopthesequester, cuts, victims, domestic, real	78.0	0.9
6	**Unemployment Benefits:** renewui, unemployment, insurance, americans, benefits, lost, gop, poverty, unemployed, must	78.2	0.9
21	**Minimum Wage:** raisethewage, wage, minimum, workers, raise, raising, petition, timefor1010, time, minimumwage	79.0	0.6
11	**Affordable Care Act:** aca, health, care, getcovered, insurance, obamacare, affordable, repeal, coverage, medicare	79.7	0.8
39	**Senate Democrats:** senate, sen, military, murray, sexual, senator, assault, help, women, proud	93.8	5.3

Table 3: *(continued)*

Topic		%D	%F
12	**House Democrats:** rep, congress, house, thank, thanks, gop, women, proud, congressman, immigration	95.7	4.9

Each line shows the top 10 words associated with each topic, as well as the label we assigned to each of them. The third column indicates the proportion of each topic that is associated to Democrats in Congress. For example, 95.7% of the time that legislators talk about topic 12, it is due to Democrats. The last column indicates frequency of use of each topic. For example, topic 7, "Military Intervention in Syria", corresponds to approximately 1.1% of all the tweets legislators sent during this period.

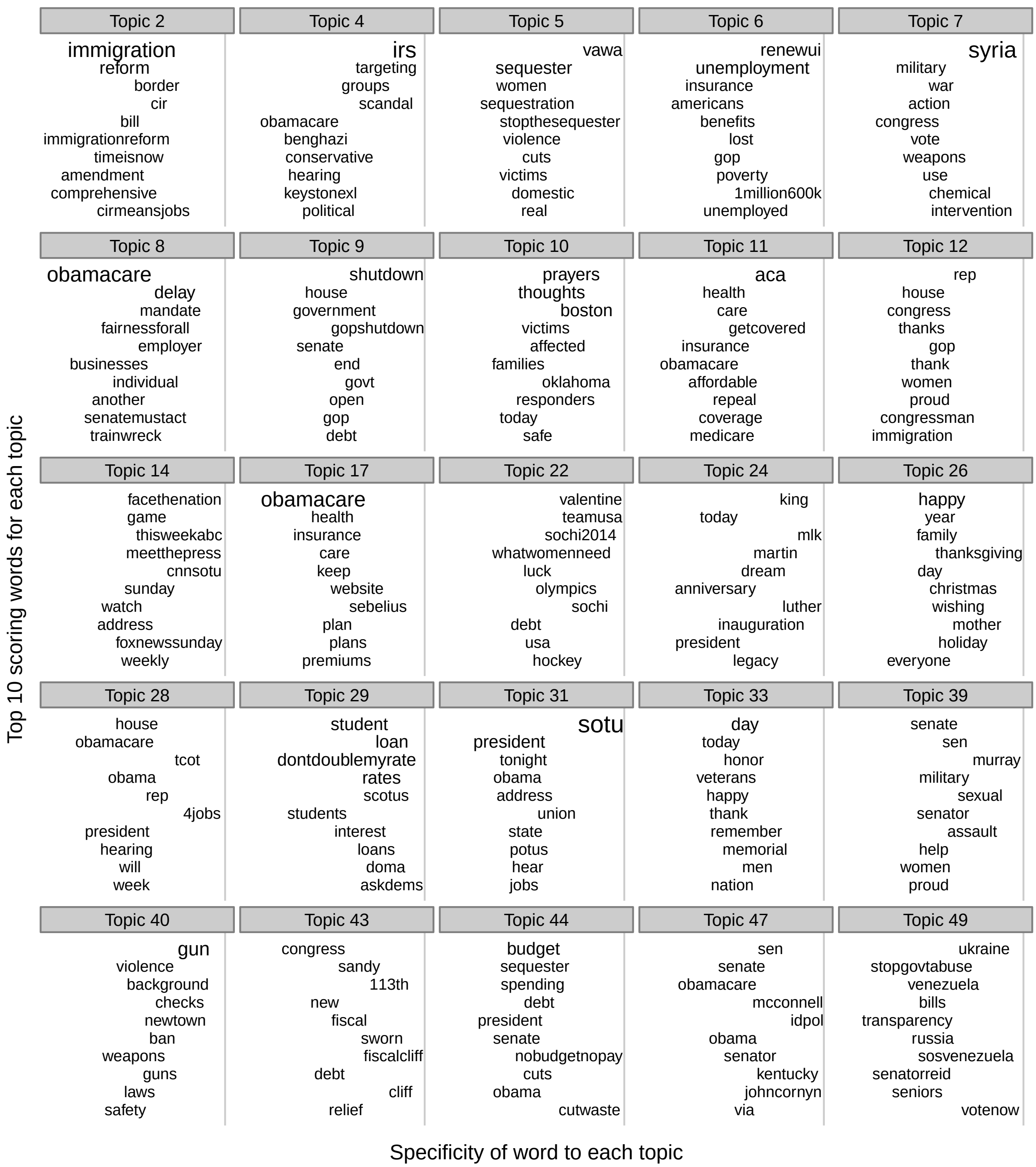
Although there is some variation in the correspondences between each estimated topic and sets of political issues, we can identify the following topics:

1. Mel Watt Confirmation for FHFA
2. Immigration Reform
3. Iran Nuclear Deal
4. IRS Scandal
5. Violence Against Women Act
6. Unemployment Benefits
7. Military Intervention in Syria
8. Obamacare Implementation
9. Government Shutdown
10. Boston Marathon Attack
11. Affordable Care Act
12. House Democrats
13. Budget Deal II
14. Republicans on TV
15. Death of Nelson Mandela
16. Budget Deal I
17. Obamacare Website
18. Tax Reform
19. September 11
20. Women's Issues
21. Minimum Wage
22. Olympic Games
23. United for Marriage Campaign
24. Martin Luther King Day
25. Social Security
26. Holidays
27. Small Business and Farm Bill
28. House Republicans
29. Education and Marriage Equality
30. Sequester and Airline Industry
31. State of the Union Address
32. Ronald Reagan
33. National Holidays
34. Congress Stopwords I
35. NSA Surveillance Scandal
36. Benghazi Scandal
37. Birthdays and celebrations
38. Super Bowl and Death of Margaret Thatcher
39. Senate Democrats
40. Gun Control
41. Congress Stopwords III
42. Climate Change
43. Congress Inauguration
44. Budget and Sequester (Republicans)
45. Congress Stopwords II
46. Family and Medical Leave Act
47. Senate Republicans
48. KeystoneXL Pipeline and weather
49. Foreign Policy
50. Poverty

Based on the top scoring words for each topic, and the empirical estimates of what party owns each of them, we classify them in four different categories: topics that are more related to the agenda of the Democratic (1, 5, 6, 11, 12, 20, 21, 24, 25, 29, 39, 40, 42, 46, 50) or Republican (4, 8, 14, 16, 17, 18, 28, 30, 32, 35, 36, 44, 47, 48, 49) party, bipartisan topics (2, 3, 7, 9, 13, 19, 23, 27, 31, 34, 41, 43, 45), and topics related to non-political events (10, 15, 22, 26, 33, 37, 38).[14]

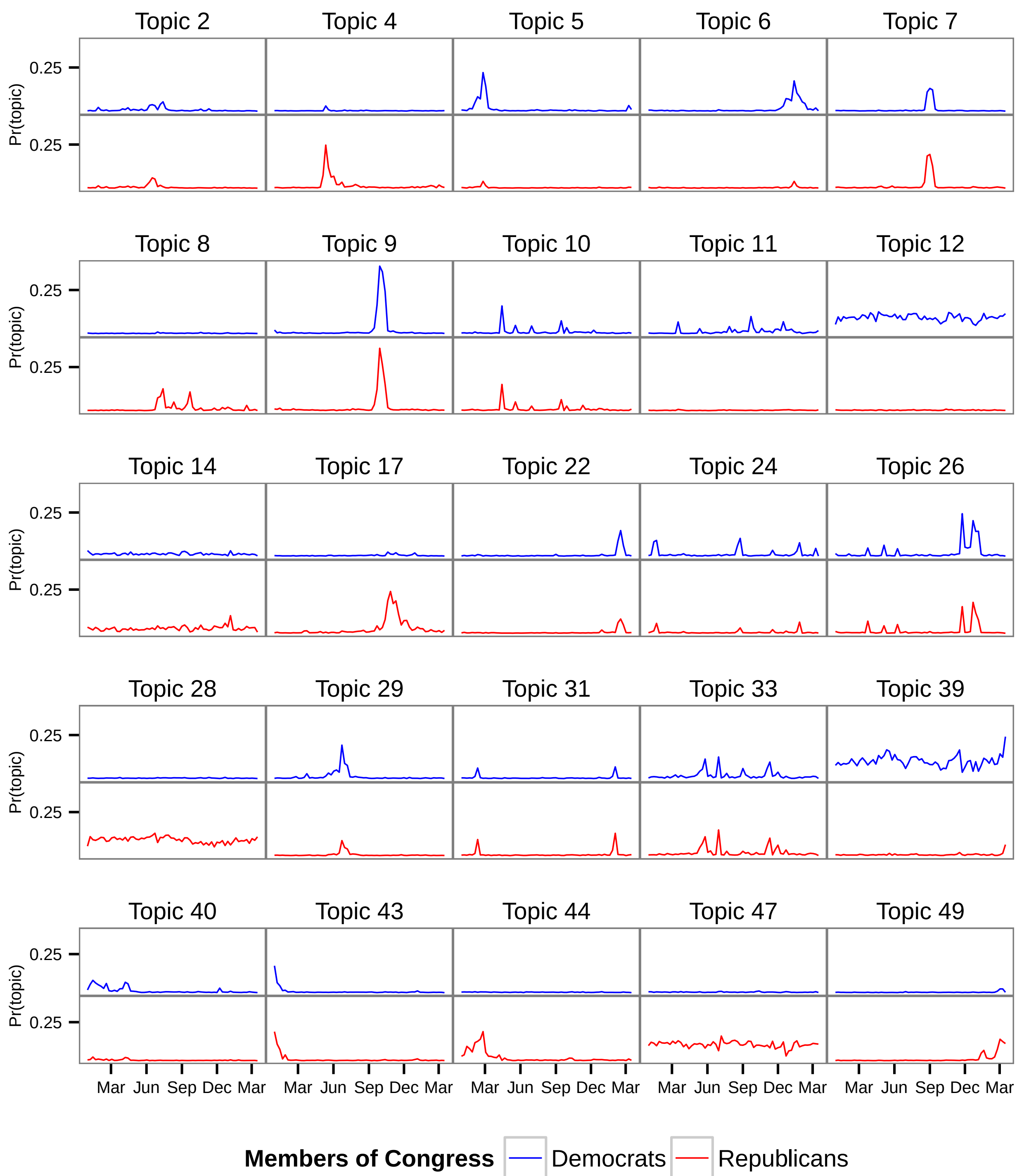
[14] We define as topics "owned" by a party those that are used by one of the parties more than 60% of the time, according

Figure 4: Top Scoring Words in each Topic



to the results on Table 3.

Figure 5: Distribution of Topics over Time, by Party (Members of Congress)



This categorization is helpful because it generates specific hypotheses about what the distribution of the attention devoted to each of these topics over time should be for each party, which will allow us to assess their predictive validity. That is precisely the purpose of Figure B.3 on page 32, where we plot the estimated $\hat{\theta}$, this is, the distribution of topics over time separately for each party, which roughly corresponds with the proportion of tweets that are classified in each of the 50 topics for a given day and party.[15]

These results are consistent with our theoretical expectations. Topics 5 (Violence Against Women Act), 6 (Unemployment Benefits), 29 (Education and Marriage Equality) and 40 (Gun Control) are discussed by Democratic Members of Congress. Topics 4 (IRS Scandal), 17 (Obamacare Website), and 44 (Budget and Sequester), on the contrary, seem to be used only by Republicans. At the same time, the distribution for non-partisan topics, such as 10 (Boston marathon attack), 33 (national holidays), and 43 (Congress Inauguration), is similar across both parties.

5.2 The Political Agenda of Their Followers

We now turn to the validation of our topic model results after applying the estimated parameters for Members of Congress to the observed distributions of the words used by their followers and our control groups.[16] Using simulation, we compute $p(\hat{\theta}_F|\hat{\alpha}, \hat{\beta}, \mathbf{w}_F)$, the posterior distribution of topics over documents, $\hat{\theta}_F$ (i.e. tweets, aggregated by day and group), given the Dirichlet prior $\hat{\alpha}$, the estimated distribution of words over topics $\hat{\beta}$, and the observed word counts $\mathbf{w}_F$. This process gives us θ_F, this is, an estimate of how much each topic was discussed by our groups of ordinary Twitter users over the period of analysis.

Figures 6 and 7 display a selection of our results. The first plot compares the relative importance of topics that we identified as being "owned" by Democratic Members of Congress in the political agenda of each of our three groups of Democratic users. This plot confirms our intuition that users who follow more Members of Congress write more tweets with a political content. The differences are particularly evident for Topic 9, "government shutdown", as we would expect given that it's very related to Congress' activities.

[15] We combine the topic distributions from both chambers by computing a simple average that weighs each chamber equally.

[16] We apply identical pre-processing techniques to those described in Section 4, and include only words that appear in the vocabulary of words used by Members of Congress.

Figure 6: Distribution of Topics (Selection), for Democratic Followers

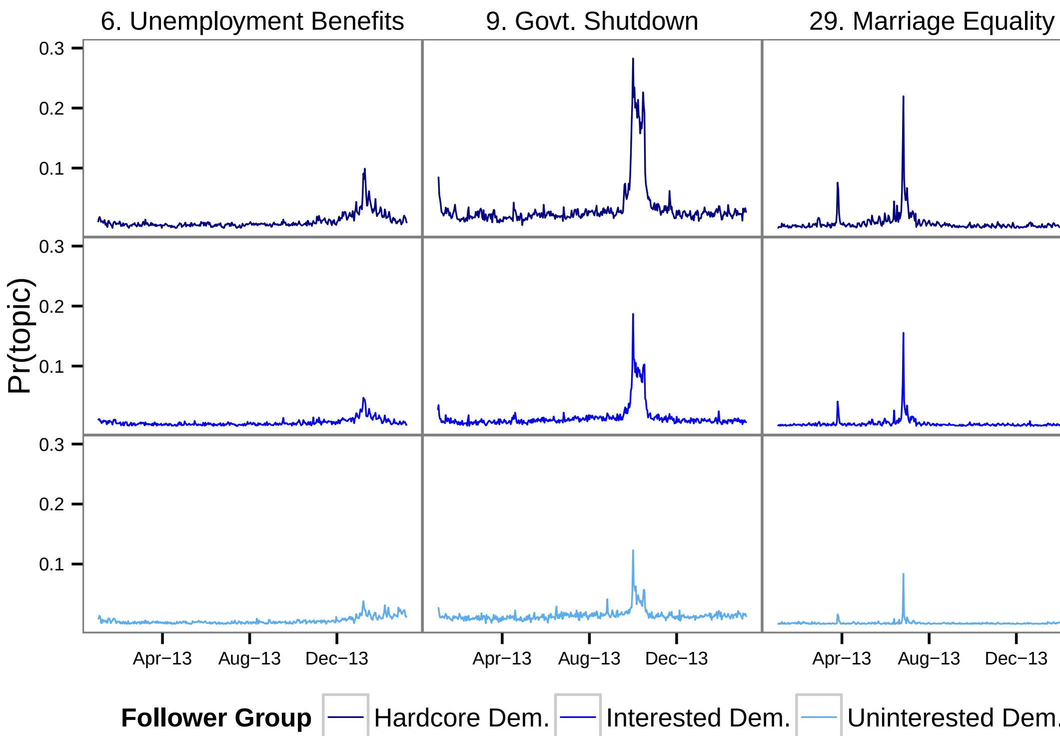
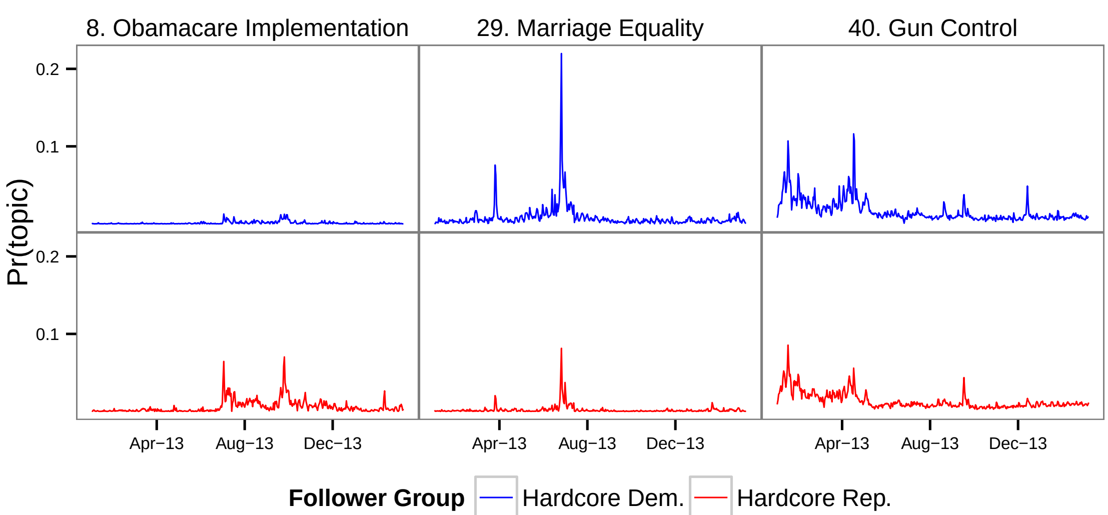


Figure 7: Distribution of Topics (Selection), for Hardcore Followers



The second plot makes the complementary comparison, by showing the distribution of three political topics for hardcore democrats and hardcore republicans. As we expected, Topics 29 and 40 (marriage equality; gun control) are more frequently discussed by hardcore Democrats, while the result is the opposite for Topic 8 (Obamacare implementation). However, note that these differences are not as wide as for Members of Congress, which is consistent with some of the insights of the literature on mass and elite political polarization (McCarty, Poole and Rosenthal, 2006; Fiorina and Abrams, 2008).

Tables 4 and 5 complement our analysis by examining the most frequent topics across follower groups and states; and the legislative party to which each group resembles more based on their topic distribution. "Hardcore" followers use topics about congressional politics more often than "uninterested." Similarly, those who we classified as Republicans use topics owned by Republican legislators more frequently, and have topic distributions that are closer to theirs; and vice versa for Democrats. As we show in Figure 8, when we replicate this computation at the state level, we find that "red" states use Republican topics more frequently than "blue" states.[17]

Table 4: Topic distribution by group

Group	T	C
Hardcore democrats	39	D
Hardcore independents	47	D
Hardcore republicans	28	R
Interested democrats	12	D
Interested independents	48	D
Interested republicans	28	R
Uninterested democrats	37	D
Uninterested independents	37	D
Uninterested republicans	37	D

T = most common topic by group. C = party to which topic distribution in each group is closest.

Overall, these results suggest that our measure of the political agenda of ordinary Twitter users has face validity, and that our classification is properly capturing political ideology and interest in congressional politics, which allows us to draw meaningful comparisons across groups and also with Members of Congress.

[17]There is also some interesting variation across states. For example, gun violence is the most frequent topic in Connecticut, and the Boston marathon attack is the most common topic in Massachusetts. However, note that the sample size in some of these states is small (50–100 users in most cases), so these results should be taken with caution.

Table 5: Topic distribution by state

State	T	C	State	T	C	State	T	C
Alabama	28	R	Kentucky	44	R	North Dakota	38	R
Alaska	44	R	Louisiana	37	R	Ohio	23	R
Arizona	44	R	Maine	44	D	Oklahoma	10	R
Arkansas	28	R	Maryland	37	D	Oregon	40	D
California	12	D	Massachusetts	10	D	Pennsylvania	34	D
Colorado	42	D	Michigan	12	D	Rhode Island	38	D
Connecticut	40	D	Minnesota	21	D	South Carolina	44	R
Delaware	38	D	Mississippi	19	R	South Dakota	45	R
DC	12	D	Missouri	17	R	Tennessee	32	R
Florida	28	R	Montana	5	D	Texas	28	R
Georgia	28	R	Nebraska	8	R	Utah	46	R
Hawaii	38	D	Nevada	41	D	Vermont	38	D
Idaho	44	R	New Hampshire	1	R	Virginia	35	R
Illinois	12	D	New Jersey	41	D	Washington	37	D
Indiana	32	R	New Mexico	40	R	West Virginia	16	D
Iowa	25	D	New York	12	D	Wisconsin	20	D
Kansas	28	R	North Carolina	14	R			

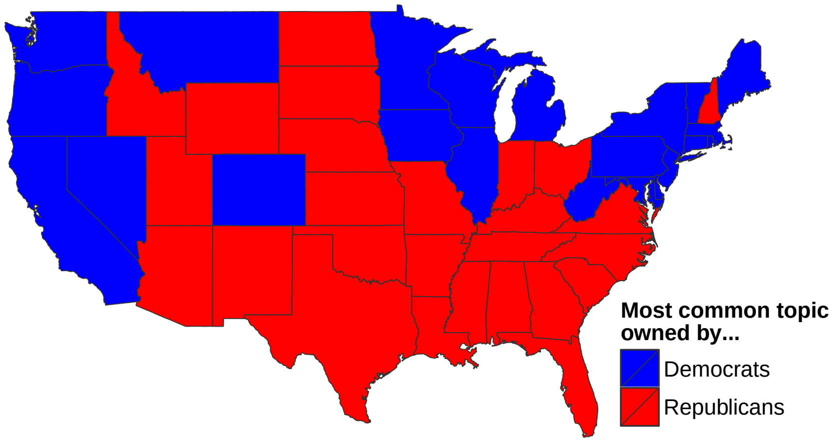
T = most common topic by state. For example, topic 28 (topic associated to House Republicans) is the most frequently used in Alabama. C = party to which topic distribution in each state is closest. For example, the topics used in Alabama are more similar to those used by Republicans than by Democrats.

6 Members of Congress and their Followers: Who's Listening to Whom?

The key substantive question we want to answer is whether the distribution of topics discussed by Members of Congress leads or follows that of their constituents. Are members reacting to their constituents, or vice versa?

We start by examining the similarity in the topic distribution across groups defined by partisanship and interest in congressional politics. Figure 9 displays our estimates of how each group uses the most common topics used by Republicans and Democrats in Congress. For each party, we found the five most common political topics after excluding "stop words" topics. These five topics account for around 47% of all their tweets. Then, we examined what proportion of the tweets within each group belongs to that same set of topics. Our results demonstrate that topics discussed by legislators and topics discussed by their followers are related. We observe that fewer than 20% of the tweets sent by hardcore Republicans are among the top 5 topics of Democratic Members of Congress, but more than 40% of those tweets are on the top 5 topics of Republican Members of Congress. Fig-

Figure 8: Party Ownership of Most Common Topic in Each State



ure 9 also shows that, as we move from hardcore partisans to only interested partisans, and then to uninterested partisans, the gap in congruence of topic use with each party decreases, as we would expect.[18]

Figure 9 shows that topic distributions across groups are similar, and that the differences are consistent with our theoretical expectations, but does not address whether longitudinal changes in these distributions are also related. In order to assess whether Members of Congress are responsive to what their constituents discuss on Twitter, it is necessary to examine whether changes in topic usage by constituents affect future topic usage by legislator.

We do so by applying the standard Granger-causality framework (Granger, 1969). To limit the scale of our analysis, we focus our attention on six groups: Democratic and Republican Members of Congress, hardcore Democrat and Republican constituents, and uninterested Democrat and Republican constituents.

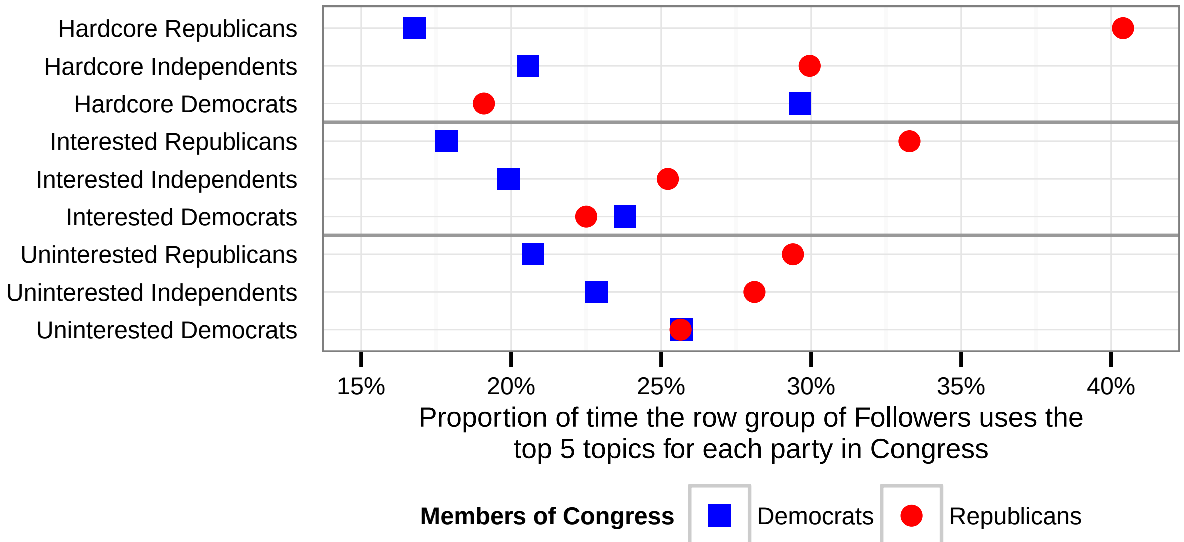
For each of our six groups, we regressed the proportion of tweets on a set of topics J at time t sent by that group on the proportions of tweets for each of the six groups (including itself) in each of the preceding five days.[19] Each of these regressions was run including a different set of topics: all political topics, topics related to non-political events, topics owned by Democrats, and topics owned by Republicans (see page 15 for a full list of topics in each category).[20]

[18] Figure A.4 in the Appendix replicates this analysis by computing Pearson correlation coefficients across topic distributions, finding almost identical results.

[19] We chose 5 days as the appropriate lag structure after examining the fit of models of Members' tweets as a function of different lag lengths of Members' tweets. The optimal model fit (measured using the AIC criterion) was obtained with 7 lags. However, we chose a more conservative set of lags to avoid overfitting the data, and because from a theoretical perspective it is unlikely that any substantively relevant effects take place with a lag of more than 5 days.

[20] In other words, we computed 4 topic groups $\times$ 6 user groups $=$ 24 regressions, with 5 lags $\times$ 6 groups $=$ 50 independent variables; where topic distributions for each group of users were "stacked."

Figure 9: Similarity in Topic Usage Between Members of Congress and their Followers

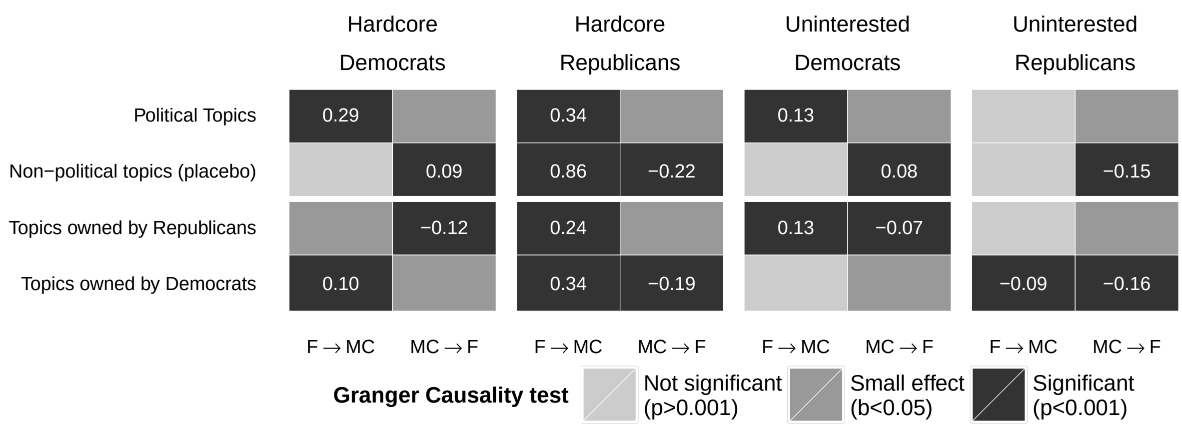


We then performed a standard F-test for the joint significance of each set of five lagged variables corresponding to our group of interest. When the F statistic is significant, we can reject the null hypothesis that the topic distribution of the group to which the lagged variables correspond has an effect on the topic distribution of the group to which the dependent variable corresponds. When the test suggests that one group's tweets help predict the other group's tweets, but not vice versa, then we conclude that one group's tweets granger-causes the other group's tweets. If we conclude that both groups' previous tweets help predict contemporaneous tweets of the other group, then we conclude that both groups are influenced by previous behavior of the other.

In Figure 10 we report the main results of our analysis. Each cell corresponds to one test; for example, the first cell on the top left-hand corner indicates whether changes in the topic distribution for hardcore Democrats have a significant effect on the topic distribution of Democratic Members of Congress; and the second cell displays the same result when the analysis is conducted in the opposite direction. The color of the cells indicates whether the F statistic leads us to reject the null hypothesis (black cells) or not (grey cells).[21] When significant, the magnitude of the sum of coefficients for the lags is reported. This coefficient can be roughly interpreted as the change in the topic distribution, in percentage points, as a result of a one-percentage increase in the topic distribution of the group of users on the right-hand side of the equation. To facilitate the interpretation, we do not report coefficients of small magnitude ($b < 0.05$), indicated by a pale gray shading.

[21] Given the large size of our dataset, we choose conservative significance levels ($p < 0.001$) to reject our null hypothesis.

Figure 10: Results of Granger Causality Tests, by Type of Topic



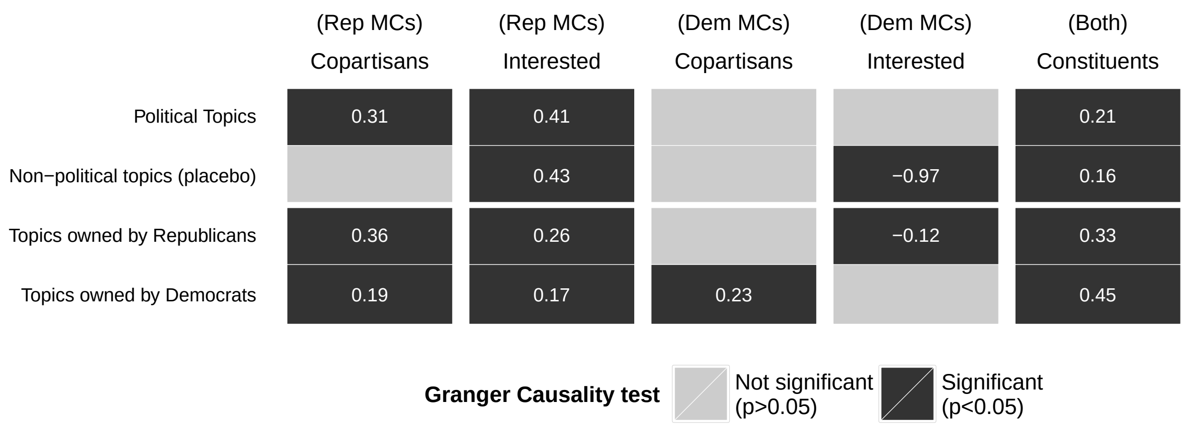
We find evidence of political responsiveness in the issues Members of Congress discussed: legislators of both parties increase the attention they pay to political topics in response to shifts in hardcore followers' attention to that same set of topics. The magnitude of this effect is larger for topic owned by each party, particularly in the case of Democrats. Despite our expectations, we do find significant coefficients in the case of non-political topics, which we consider a "placebo." The results concerning followers uninterested in congressional politics are similar but lower in magnitude, as we expected bases on their degree of attentiveness. Some coefficients are even negative, which would imply that Members of Congress talk *less* about these topics in response to shifts in attention by uninterested followers.

Our results are robust to different methodological decisions. Appendix C replicates our analysis using word counts for a selected set of issues, with similar findings. Figure C.1 displays our estimated coefficients when we split our sample according to each of the 50 topics.

Having found that Members of Congress are responsive to their followers in the issues they discuss, we now turn to examine to what groups they are more responsive. Figure 11 report the results of a series of F-tests for the difference between the estimated coefficients for different groups of followers. The first two columns show that Republican legislators are significantly more responsive to co-partisans (followers with conservative ideology) and to followers interested in congressional politics (following more than 3 Members of Congress). Columns 3 and 4 display the results after replicating this estimation with Democratic legislators. In this case, we find that they are more responsive to co-partisans only for topics owned by Democrats, and that they are at least as equally responsive to the entire electorate than to "hardcore" followers. These results suggest that, while Republicans pay close attention to what their core supporters care about, Democrats respond to the entire electorate.

The final column of Table 11 shows the results of a slightly different type of analysis. Here, we split legislators (both Senators and Representative) and their followers by state, and then compute their posterior distribution of topics and following the same procedure described in Section 5.2. We also aggregate all the topic distributions for followers in order to obtain the average topic distribution for all followers. We then "stack" all topic distributions and run the same Granger causality model where the dependent variable is the topic distribution of legislators, and the independent variables are the lagged topic distributions of legislators, of constituents from the state they represent, and of all constituents. Our results clearly indicate that legislators are significantly more responsive to voters from their state than to the entire U.S. electorate.

Figure 11: Differences in Estimated Responsiveness to Different Follower Groups



7 Discussion and Conclusions

In this paper we have characterized the political agenda of Members of Congress and their constituents using latent topic modeling applied to the text of their tweets. We have shown that legislators are responsive to their followers (and their constituents) in the issues they discuss, but have minimal influence on their followers' public agenda. This effect varies across parties: Republicans appear to be more influenced by their core supporters than by all conservative Twitter users, while Democrats are equally responsive to both groups. These findings demonstrate that meaningful insights can be extracted from an analysis of how Representatives and Senators communicate through social media.

We plan to extend and improve our present study in two directions. First, one important advantage of using social media messages as a source of information about political discussion is that it allows us to increase the granularity of our observations. So far our unit of analysis was tweets aggregated by day, but it would be easy to reduce it half days or even hours in order to adapt to the fast pace of Twitter; and also aggregate it to weeks or months to study long-term changes in language use. Similarly, we aggregated all Democratic and Republican Members of Congress, but we could also focus on individual legislators and their followers. In doing so, we need to be aware of the trade-offs that these choices present: more granularity can also be accompanied by more measurement error in the estimation of topic distributions.

A second possible improvement would lie in our methodology. As we discussed in Section 4, our model assumes independence over time and that topics are stable in number and content. Other latent topic models such as the dynamic topic model introduced by Blei and Lafferty (2006), whose most prominent application in Political Science can be found in Quinn et al. (2010), would allow us to relax these assumption and thus could be a more appropriate choice that might improve the validity of our estimates. Similarly, we could also complement our empirical strategy with supervised approaches, such as focusing on the use of specific words that belong to a set of pre-defined issues, or manual classification of tweets in topics. Along the same lines, our estimates require further validation, which could be achieved by splitting Members of Congress and their followers into different groups (e.g. senior vs junior Members of Congress, male vs. female, by geographic location, by number of followers...) and then examining whether their posterior topic distributions are consistent with our theoretical expectations. It is also important to note that we did not exclude retweets from our analysis, which could be an important source of common variation in language use across groups.

The scope of this paper is limited to an analysis of Members of Congress and their constituents, but the method we have developed could be applied to any group of Twitter users, including media outlets and journalists, politicians and ordinary users in different countries, celebrities, etc. Our analysis could also be replicated in the context of a study of social influence: to what extent do media outlets follow or set the public agenda? what types of political issues spread across borders? what characteristics make specific groups of users influential in patterns of language use? These are some of the questions that we hope to address in our future work.

References

Barberá, P. 2013. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Unpublished manuscript, New York University.* .

Baumgartner, Frank and Bryan D. Jones. 2009. *Agendas and Instability in American Politics*. 2nd ed. Chicago, Illinois: University of Chicago Press.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3:993–1022.

Blei, David M and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM pp. 113–120.

Chi, F. and N. Yang. 2010. "Twitter adoption in congress." *Review of Network Economics, Vol. 10, No. 1* .

Clinton, Joshua D. 2006. "Representation in Congress: Constituents and Roll Calls in the 106th House." *Journal of Politics* 68:397–409.

Evans, Heather K., Victoria Cordova and Savannah Sipole. 2013. Twitter Style: An Analysis of How House Candidates used Twitter in their 2012 Campaigns. In *Paper Presented at the 2013 Midwest Political Science Association Conference*.

Ezrow, Lawrence, Catherine De Vries, Marco Steenbergen and Erica Edwards. 2011. "Mean voter representation and partisan constituency representation: Do parties respond to the mean voter position or to their supporters?" *Party Politics* 17(3):275–301.

Fiorina, Morris P and Samuel J Abrams. 2008. "Political polarization in the American public." *Annu. Rev. Polit. Sci.* 11:563–588.

Gerber, Alan S, Dean Karlan and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1(2):35–52.

Gilens, Martin. 2012. *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton, New Jersey: Princeton University Press.

Golbeck, Jennifer, Justin M Grimes and Anthony Rogers. 2010. "Twitter use by the US Congress." *Journal of the American Society for Information Science and Technology* 61(8):1612–1621.

Granger, Clive WJ. 1969. "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica: Journal of the Econometric Society* pp. 424–438.

Griffiths, Thomas L and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National academy of Sciences of the United States of America* 101(Suppl 1):5228–5235.

Grün, Bettina and Kurt Hornik. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13):1–30.

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

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. Springer.

Hemphill, Libby, Jahna Otterbacher and Matthew Shapiro. 2013. What's congress doing on twitter? In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM pp. 877–886.

Hong, Liangjie and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*. ACM pp. 80–88.

Ladd, Jonathan McDonald and Gabriel S Lenz. 2009. "Exploiting a rare communication shift to document the persuasive power of the news media." *American Journal of Political Science* 53(2):394–410.

McCarty, Nolan M, Keith T Poole and Howard Rosenthal. 2006. *Polarized America: The dance of ideology and unequal riches*. MIT Press Cambridge.

Phan, Xuan-Hieu, Le-Minh Nguyen and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*. ACM pp. 91–100.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Shapiro, M.A., L. Hemphill and J. Otterbacher. 2012. "Doing What I Say: Connecting Congressional Social Media Behavior and Congressional Voting.".

A Additional Figures

Figure A.1: Distribution of number of tweets sent by Members of Congress

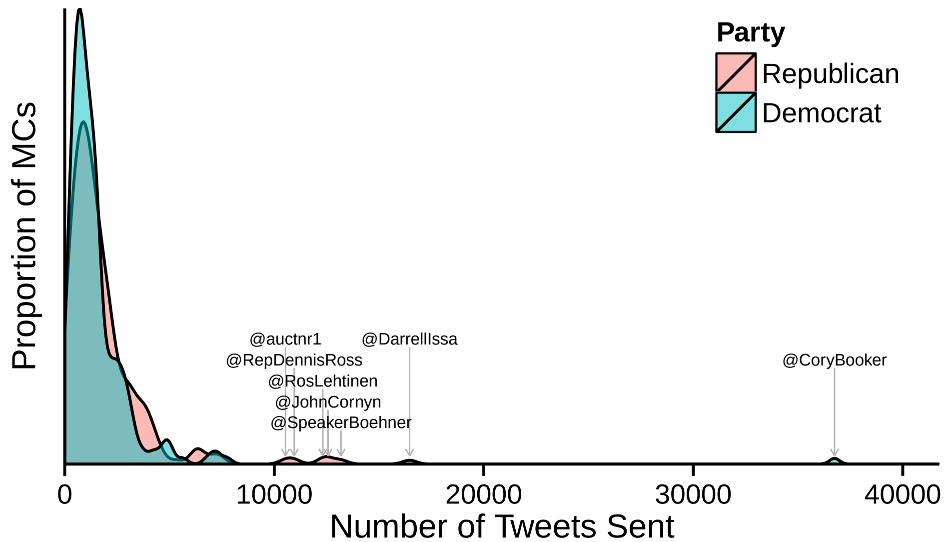


Figure A.2: Distribution of number of followers of Members of Congress

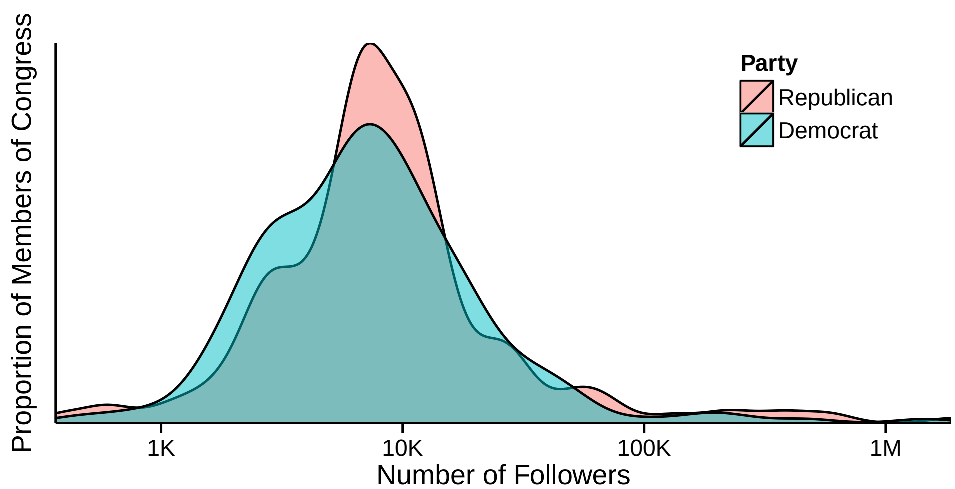


Figure A.3: Number of tweets in dataset, by day

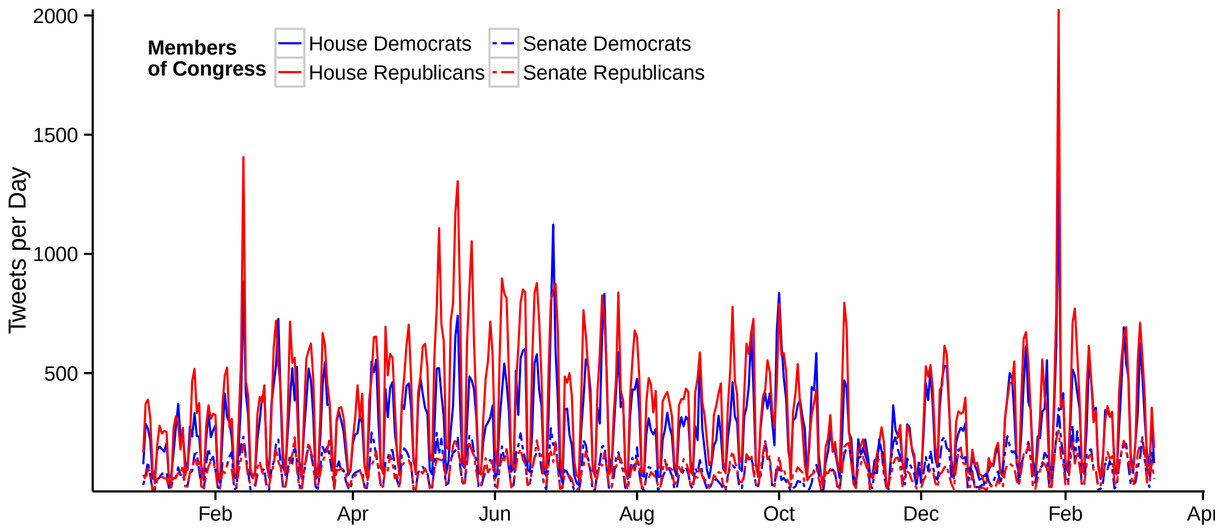


Figure A.4: Correlations in Contemporaneous Topic Distributions of Members of Congress and their Followers

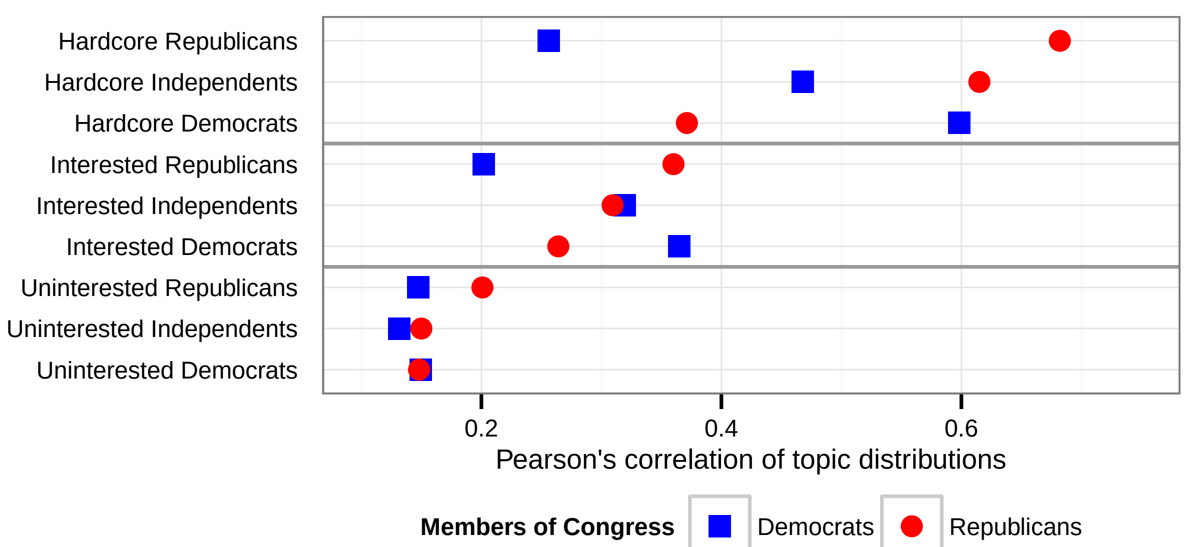
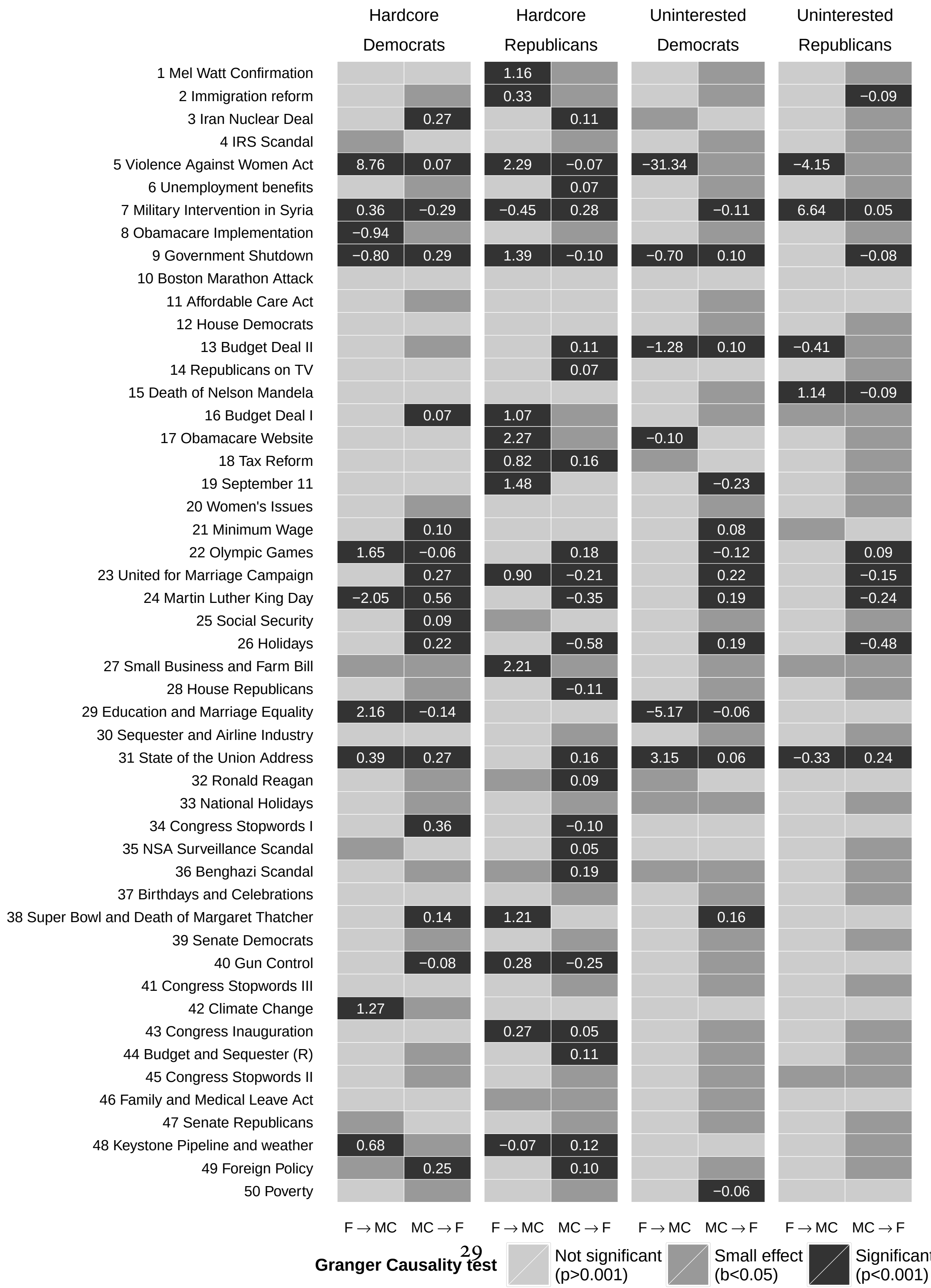


Figure A.5: Results of Granger Causality Tests, by Topic



B Full results of topic model

Figure B.1: Top Scoring Words in each Topic (Topics 1 to 25)

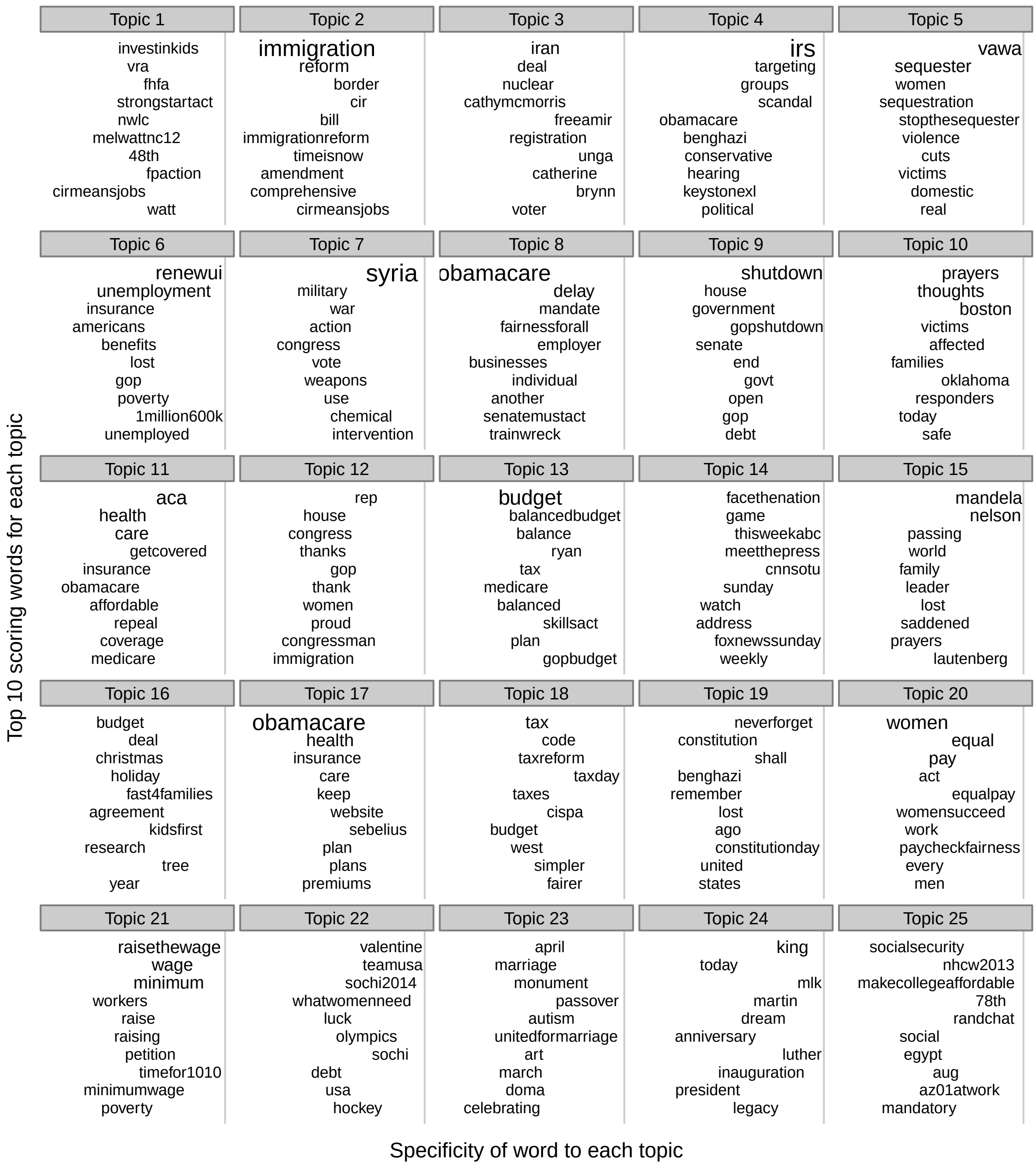


Figure B.2: Top Scoring Words in each Topic (Topics 26 to 50)

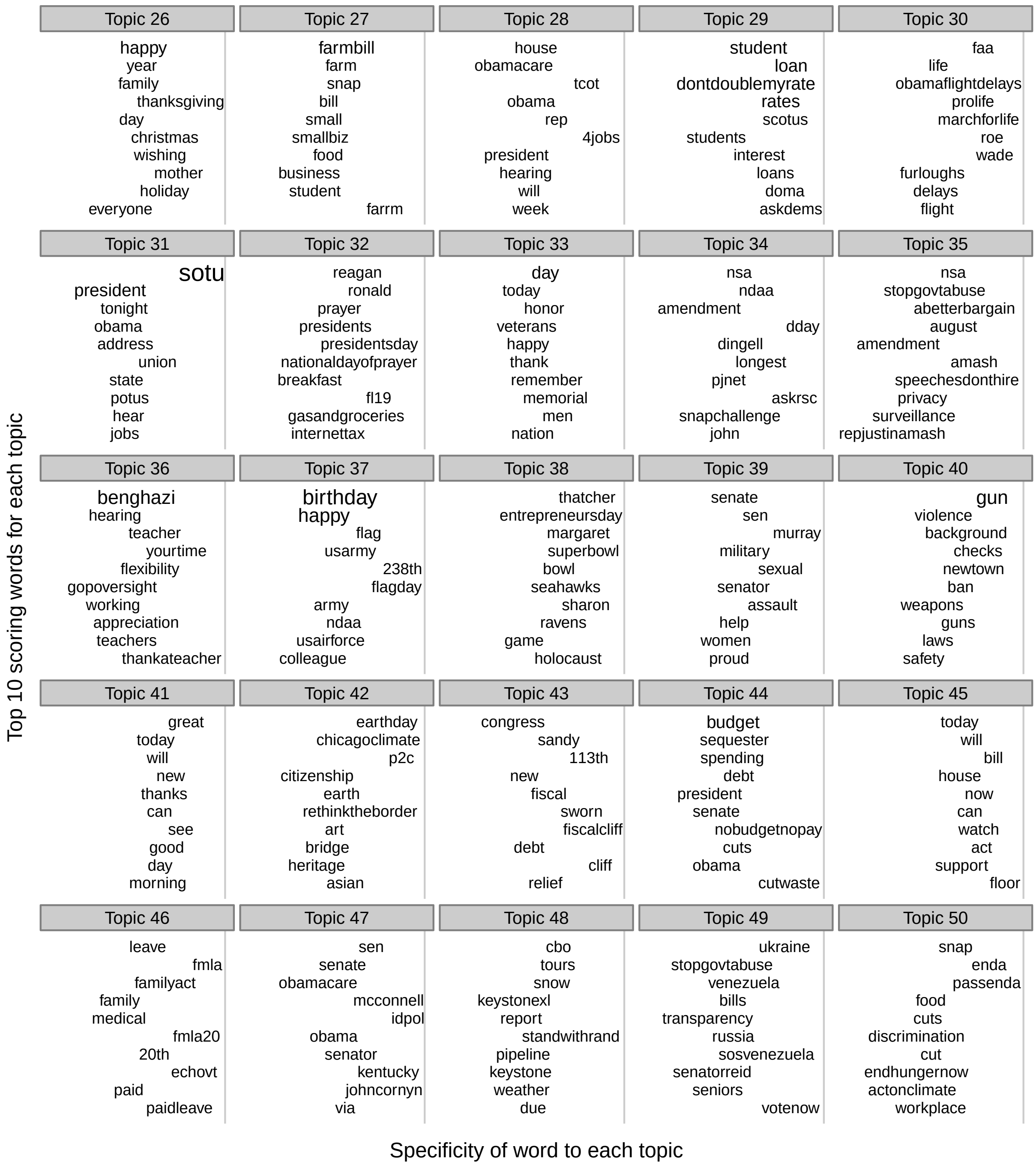
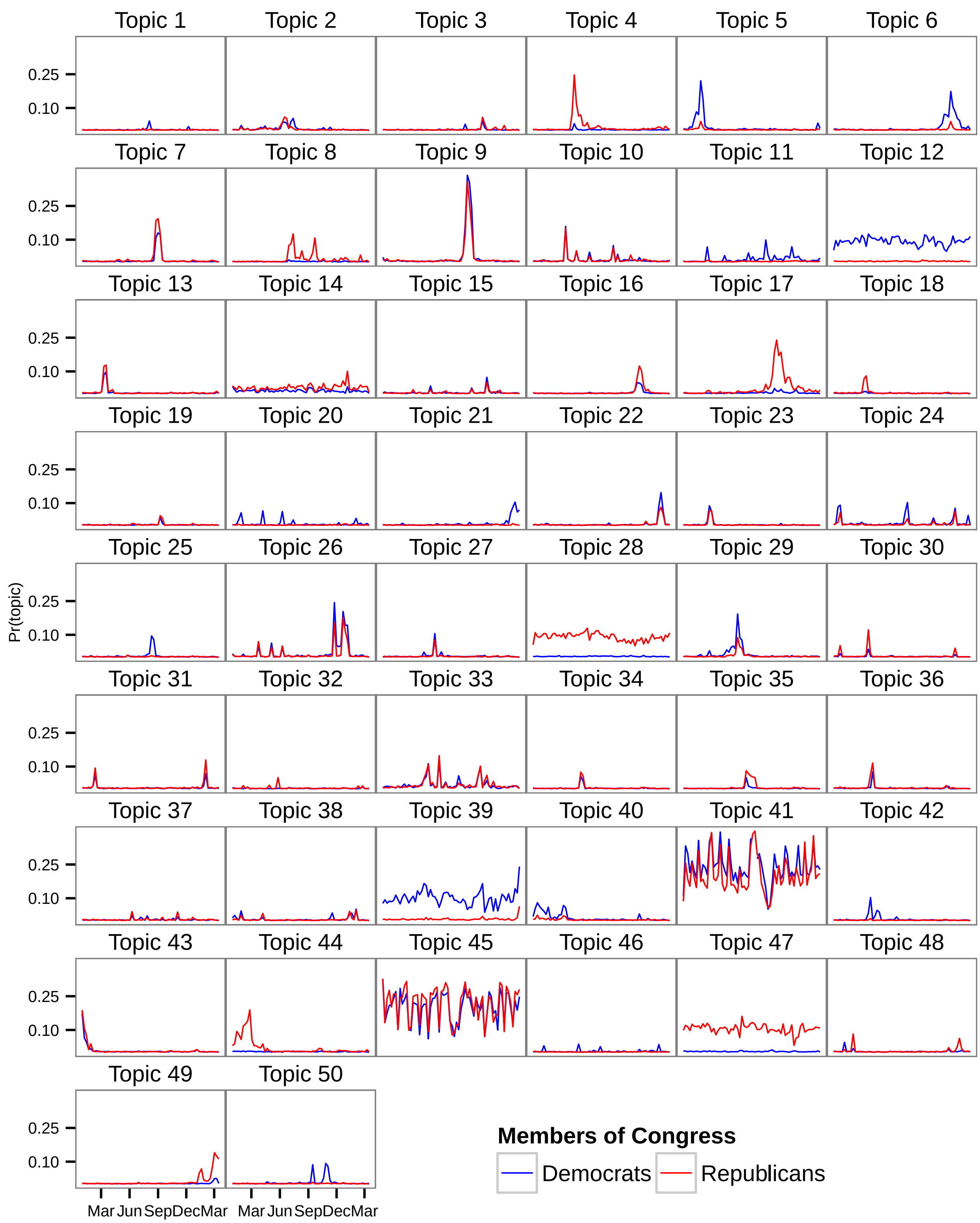


Figure B.3: Distribution of Topics over Time, by Party (Members of Congress)



C Robustness check. Analysis using word counts

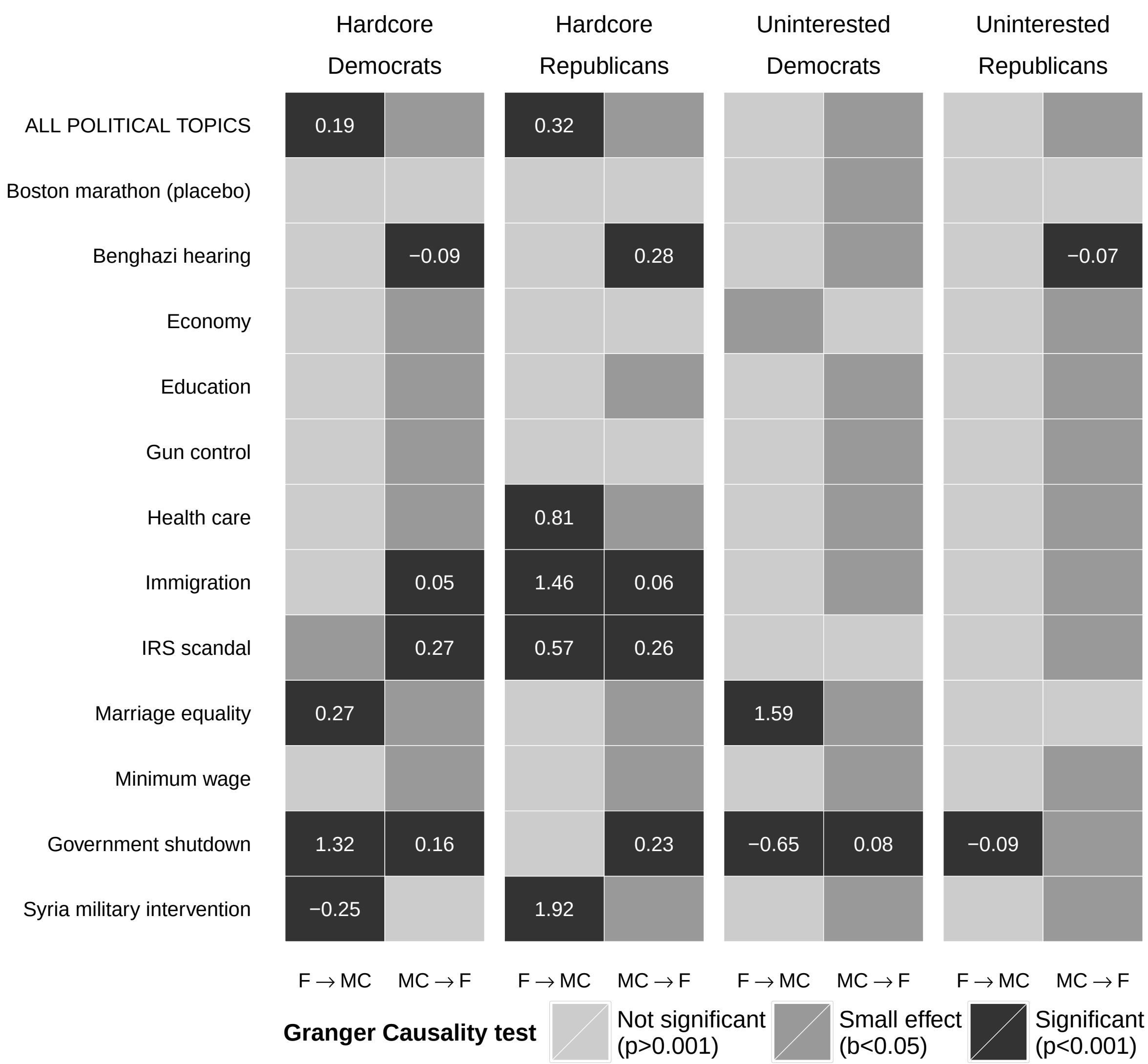
One important disadvantage of unsupervised topic models is that the resulting "topics" often do not perfectly overlap with existing political issues. Some topics refer to multiple issues, specially when they are discussed in Congress at the same time; and some relevant issues, such as economic policy, education, and immigration reform, might not appear because legislators only refer to them very sparsely. In order to show that our results are robust to different categorizations, here we replicate our main analysis using word counts instead. Informed by the topics we found in our estimation, and by our substantive knowledge about Congress, we prepared list of words associated with 12 political topics, including a "placebo" topic (the Boston Marathon attack) where we wouldn't expect to find any effect. Table C.1 lists the issues we considered and the words associated to each of them.

Table C.1: Words Associated to Political Issues

Topic	Words
Boston attack (placebo)	prayers, thoughts, boston, marathon, victims, responders, tragedy, families
Benghazi hearing	benghazi, hearing, libya
Economy	unemployment, inflation, gap, growth, wages, salary, labor, prices, stock
Education	teachers, students, thankateacher, education, dontdoublemyrate, college, loans, teacher
Gun control	gun, control, rights, background, checks, newtown
Health care	obamacare, health, care, website, insurance, plans, medicare, medicaid, ACA
Immigration	immigration, deportation, citizenship, border, security
IRS scandal	IRS, targeting, groups, scandal
Marriage equality	DOMA, scotus, marriage, equality, loveislove
Minimum wage	minimum, wage, raisethewage, tenten, actontenten
Gov. shutdown	shutdown, gopshutdown, government, enough already, default, demandavote
Syria milit. intervention	syria, chemical, military, action, war, weapons

We computed the number of times per day each word was used by Members of Congress and their different follower groups, and then normalized them by the total number of words in their tweets in order to obtain a proportion. Using the same approach as in the main text of the paper, we estimated Granger causality tests with five lags. Figure C.1 shows our results, which are consistent with what we found using topic proportions: legislators appear to be responsive to their hardcore followers, but not the other way around; we didn't find any relationship for our placebo topic; and in general Members of Congress appear to be more responsive "hardcore" followers rather than those we characterized as "uninterested" in congressional politics.

Figure C.1: Results of Granger Causality Tests, by Topic, using Word Counts



D Topic validation

Our analysis assumes that tweets sent by Members of Congress are part of their communication strategy, and that, as such, can be considered representative of the political issues they deem relevant. Here we present additional evidence for this claim, showing that our estimates of the legislators' congressional agenda are very similar to those that would be obtained by analyzing their speeches.

Using the Capitol Words API, developed by the Sunlight Foundation, we scraped the text of all legislative speeches in the House and Senate for our period of analysis (January 1st, 2013 to April 10th, 2014). We then pre-processed the text using the same procedure described in Section 4. We will use this text as "ground truth" about the political issues legislators care about.

We conducted three different validation analyses. First, we grouped all speeches by legislator. We took the 50 legislators with the highest number of interventions in Congress, and computed their posterior topic distribution based on this text, applying the parameters from our analysis in Section 5.1. In other words, we implemented the same procedure that we used to estimate users' topic distributions, but using legislators' speeches. We also replicated this procedure using legislators' tweets, for the same 50 Members of Congress. This gave us an estimate of the proportion of time they spend discussing different political issues, based on both sources of information. We can now compare whether these distributions are similar to examine whether tweets are an accurate representation of the issues that each legislator debates on the floor of the House and Senate.

Figure D.1 displays the main results of our first analysis. For each legislator, we show the proportion of each topic based on their tweets (x-axis) and speeches (y-axis). Topics are labeled according to their number, and colored by whether they are "owned" by Democrats (in blue), Republicans (in red) or are bipartisan (in black). We also add the estimated Pearson's correlation coefficient. As we can see, in all cases there is a positive and large correlation between both topic distributions. This confirms that legislators that spend much time debating about a specific issue also tweet about that issue a lot.

Our second validation analysis consisted on replicating Figure 5 in the main text, but using legislative speeches as sources of information about how the importance of different uses in the congressional agenda varies over time. In order to do so, we grouped all speeches by day, party, and chamber. We kept only those days in which at least two different legislators from each party spoke on the floor of each house. Using the text from the speeches, we then estimated the topic distribution for each party and day, averaging across chambers. Figure D.2 shows how these topic distributions vary over time, for a selection of topics related to Congressional politics. Our results using legislative speeches mirror those from tweets. In fact, the correlation between topic proportions from speeches and tweets over time is $\rho = 0.79$. More importantly, the differences in topic usage across parties are consistent regardless of the text source (e.g. see topics 4, 5, 6, 12, and 39) and spikes in specific topics on legislative speeches correspond to spikes on tweets (e.g. see the spikes in topics 7, 9, 29, and 49).

Finally, Table D.1 illustrates that the topics that are discovered in legislators' tweets are very similar to the topics that emerge from applying the same technique to their speeches on the floor. Here we show the top 10 scoring words associated to each topic, after running LDA on the corpus of legislative speeches, aggregated by day, party, and chamber, and choosing $K = 30$. After computing the top scoring words for each topic, we then linked them to the top scoring words for each topic in the Twitter corpus, based on how many words they have in common. This table shows a very clear mapping between these topics and those we found in our earlier analysis, such as the government shutdown (row 5), the sequester and budget (row 7), unemployment benefits (row 10), immigration reform (row 2), the possibility of a military intervention in Syria (row 15), the IRS scandal (16), etc.

Overall, these three analyses indicate that tweets can be a reliable source of information about the importance that Members of Congress attribute to different political issues, whose validity is equivalent to that of legislative speeches.

Figure D.1: Comparing Topic Distributions from Tweets and Speeches, by Legislator

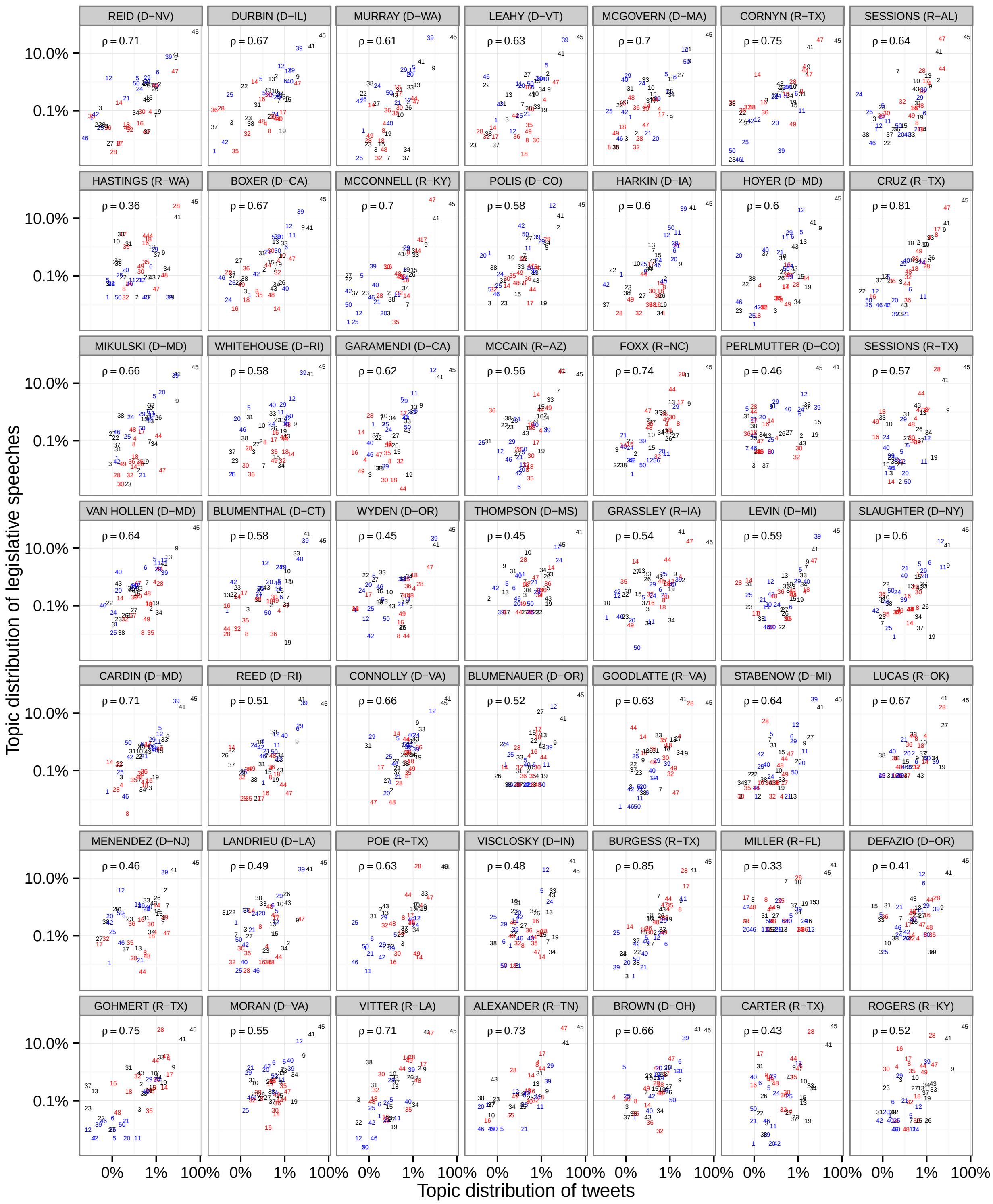


Figure D.2: Distributions of Topics over Time, from Legislative Speeches

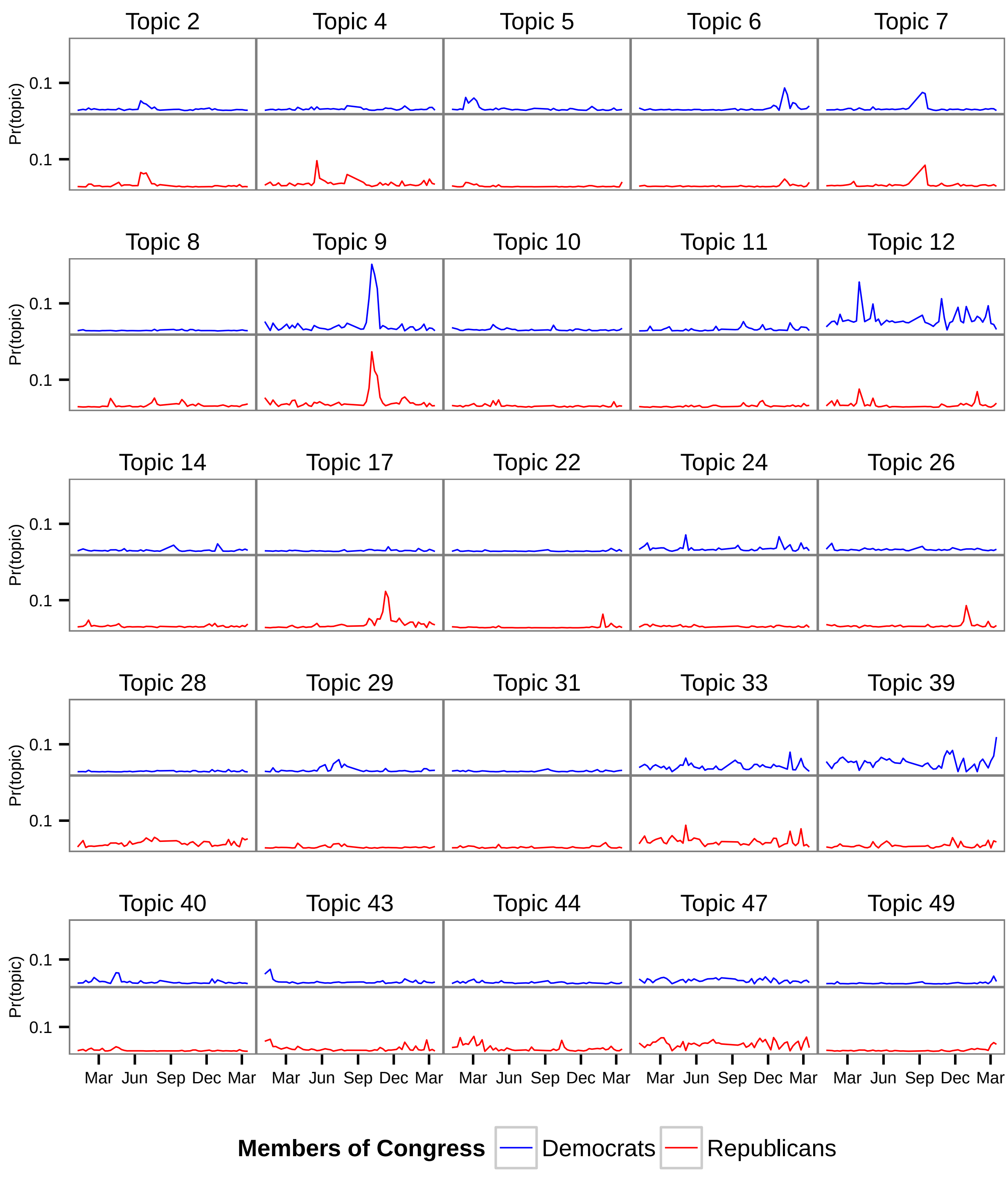


Table D.1: Top 20 Topics from Legislators' Speeches and Their Association to Topics from Tweets

	Top words associated with topic	TW
1	"will, people, can, one, going, time, just, get, now, think"	45
2	"community, today, school, service, speaker, years, many, also, rise, state"	41
3	"senator, senate, bill, president, will, committee, consent, ask, unanimous, work"	45
4	"speaker, bill, time, gentleman, act, thank, committee, support, house, today"	45
5	"government, house, shutdown, senate, republicans, people, get, republican, let, open"	9
6	"budget, health, care, cuts, will, republican, country, economy, jobs, need"	12
7	"budget, tax, spending, debt, president, taxes, year, money, government, programs"	44
8	"health, care, obamacare, insurance, law, president, people, affordable, coverage, plan"	17
9	"senate, president, court, majority, circuit, judge, nominees, rules, senator, nomination"	45
10	"unemployment, jobs, benefits, job, work, wage, people, insurance, minimum, americans"	6
11	"service, army, honor, air, sergeant, medal, war, marine, colonel, military"	33
12	"energy, climate, coal, gas, oil, change, jobs, natural, epa, states"	45
13	"speaker, government, law, people, freedom, religious, states, constitution, house, administration"	45
14	"immigration, bill, border, security, system, amendment, people, reform, law, immigrants"	2
15	"iran, nuclear, weapons, syria, president, military, israel, united, world, war"	7
16	"president, irs, government, senator, question, think, law, administration, power, constitution"	4
17	"farm, bill, food, program, farmers, snap, agriculture, programs, nutrition, crop"	27
18	"water, federal, projects, project, amendment, secretary, chairman, resources, corps, lands"	39
19	"veterans, care, health, legislation, bill, military, program, medicare, benefits, mental"	33
20	"amendment, chairman, chair, acting, gentleman, time, balance, defense, yield, funding"	45

Note: "TW" indicates the topic from the tweets that is most closely associated to each topic from legislative speeches. For example, the fifth topic from the speeches ("government, house, shutdown…") can be mapped to topic 9 in the tweets corpus ("Government Shutdown") based on its top words. In order to find the associations, we extracted the top 100 words from each topic in the speeches and then found the topic from tweets that have the most words in common among its respective 100 words.