Different Channel, Same Strategy?

Filling Empirical Gaps in Congress Literature

Andreu Casas
University of Washington
acasas2@uw.edu

David Morar
George Mason University
dmorar@gmu.edu

Abstract

Political scientists frequently study the public communications of members of Congress to better understand their electoral strategies, policy responsiveness, ability to influence public opinion and media coverage of Congress. However, different studies base their conclusions on different communication channels, including (among others) member websites, newsletters, press releases, and social media. These scholars have taken these individual sources as fully representative of the communication strategy of the elected official as a whole. However, what has not been asked is whether members communicate the same or different messages across these differing channels? In this paper we look at the members' press releases, Twitter, and Facebook messages sent from August to December 2014 in order to study to what extent their communication strategy is consistent across channels. We use an automatic semi-supervised method to classify the messages into political issues and to assess the validity of inferring broader communication patterns from a single source. Do members tend to highlight the same or different issues, and attempt to appeal to the same or different audiences across different communications channels? Does message consistency vary over time or context such as district competitiveness? Through this work we aim to give an empirical grounding to the implied notion that studying one communication channel is a proxy for studying all communication channels.

Paper to be presented at the 2015 Annual Meeting of the American Political Science Association (APSA), San Francisco, September 3-6, 2015.

Introduction

Congressional communication scholars study the communication strategy of members of Congress in order to better understand their interests, their policy positions and how political representation works (e.g. Mayhew 1974; Fenno 1978; Franklin 1991; Lipinski 2004; Groeling 2010; Carson et al. 2010; Grimmer 2013). However, although these scholars study the same phenomenon, they often use different data sources. For example, Fenno (1978) follows members of Congress to their district and uses close qualitative observations, Lipinski (2004) looks at newsletters, Groeling (2010) at TV coverage, and Grimmer (2013) at senators' press releases. These authors make the assumption that any of these data sources is a good proxy to study the communication strategy of political representatives as a whole. However, there are currently no empirical works showing that this is actually the case.

Moreover, ever since the rise of the Internet and social media platforms such as Twitter, Facebook, and Instagram, political representatives today can use a broader range of media channels to communicate their message to the public. In fact, recent research indicates that a large majority of representatives and senators are very active in social media (e.g. Straus et al. 2013) while also suggesting that the target audience of social media messages may be different than the audience of newsletters, TV messages or press releases (Barbera 2014). This begs the question *Do members of Congress adopt the same communication strategy independently of the media channel they use to communicate?*

In this paper we aim to fill this empirical gap in the literature by studying the issues members of Congress discussed in three media channels (press releases, Twitter, and Facebook) from August 1^{st} to December 31^{st} 2014. After collecting all the messages, we use a semi-supervised classifier (Nigam et al. 2006) to label all of them according to the Policy Agendas Project (PAP) topic-classification. The results show that members use the three media channels to discuss the same issues about 95% of the time. However, the difference increases as the number of press releases that a member issues decreases and the number of social media messages increases; suggesting that by only using press releases to measure a member's political agenda, researchers may be getting a slightly biased measurement as a result of having fewer data points. Moreover, despite the small size of the difference between the topics discussed in press releases compared to social media, the ideology of a member and the electoral competitiveness of her district predicts different levels of discordance between media channels.

The rest of the paper proceeds as follows. First we introduce some of the most relevant studies about the communication strategy of members of Congress and also about their social media presence and activity. Second we describe the data we collect and use in the study. Third we present the semi-supervised method we use to classify the messages into PAP topics. Then we present the results of the study and finally we conclude by discussing the implications of the findings.

Congressional Communications Research

Past research on Congressional communications (Mayhew 1974; Fenno 1978, Grimmer 2013; Lipinski 2004) use different methods and data sources to study how legislators present their work to their constituents. These works build on each other and compare their findings assuming that legislators adopt the same strategy independently of the media channel they use. For example, Fenno (1978) follows some members of Congress to their districts in order to test whether the perception that they have about their constituencies shape their behavior. Based on his qualitative observations, Fenno (1978) concludes that, since having a good relation with their constituency is important for the legislators' reelection, members of Congress adopt different communication strategies (*home styles*) depending on the nature of their constituencies. Following up on Fenno's work, in the first part of his book *Congressional Communication: Content and Consequences* Lipinski (2004) studies the communication strategy of 100 random members from 1991 to 1995 to test if Fenno's (1978:168) argument that "members run for Congress by running against it" is still valid in an era of stronger party discipline and costly campaigns. He uses a different method to measure the legislators' communication strategy. He collects and manually labels newsletters that these members mailed to their constituents and he ends up refuting Fenno's argument by showing that members of Congress are more likely to say positive things about Congress than they are to criticize it. In his analysis, Lipinski assumes that he is measuring the same phenomenon as Fenno. He argues that newsletters are a good measure of the legislators' messages to constituents because they write and send the messages themselves instead of having the mass media paraphrasing them. Similarly, Grimmer (2013) also builds on the work of Fenno (1978) and Mayhew (1974) to study whether different *home styles* correlate to different *electoral connection.* In other words, he studies if the electoral competitiveness of a district shapes the way legislators communicate to their constituents. Grimmer (2013) collects the press releases that all senators sent from 2005 to 2007 and measures the senators' strategy by looking at the issues they discuss in them.

However, it is theoretically possible that members use a different strategy when talking to constituents face to face than when sending them newsletters or press releases, and Lipinski (2004) and Grimmer do not directly compare the messages in newsletters and press releases to the members' *home style.* Fenno (1978) distinguishes different levels of constituencies (geographic, reelection, primary, and personal) and argues that members have the interest to adopt different behavior when dealing with them. Whereas it is easier for a legislator to know the type of constituent with whom she is having a face-to-face conversation, it is harder to know that when mailing a mass newsletter or press release. Thus, since the messages that members send using different media channels may have different target audiences, we argue that there is a need for further empirical knowledge on whether members adopt the same communication strategy across media channels or not.

Moreover, as more recent studies posit (Barbera 2014), and near-total adoption by members of Congress substantiates, social media is a significant part of the overall communication strategies. In the late 2000s, as social media began reaching a tipping point in adoption, and mainstream appeal (Larsson and Moe 2012, Marwick and boyd 2011, Case and King 2011, Williams and Gulati 2007), it also started gaining attention with Congressional members (Klotz 2007). Simply joining at first, and then starting to "tweet" and write Facebook statuses, members integrated social media into their communication strategy. This meant that it also popped up in the adjacent literature, which focused primarily on adoption (Druckman et al. 2007, Straus et al. 2013, Lassen and Brown 2011, Chi and Yang 2011 etc.). This initial foray into congressional use of social media was limited and mostly descriptive (Gulati and Williams 2013, Peterson 2012), and was primarily structured around the literature on diffusion of innovation. According to Zimmer and Proferes (2014, p.253) there have been 13 works in political science dealing with Twitter between 2007 and 2012, 2 in law, and 52 in communication, out of the total 382 that they found through bibliometric research. As a recent study (Pew 2015) notes, Facebook and Twitter have as their user base only a section of the overall population[1], one that is not necessarily an accurate representation of the general population. Based on this, it would be reasonable to believe that members of Congress adapt their message to each specific audience. Even more, according to Barbera (2014), the more active users, and thus, the ones that would constitute the target audience, are those that are more ideologically-extreme.

As the use of social media progressively grew within the halls of Congress, so did the explorations of it. For instance, Shogan (2010) shows that information communication technologies have thoroughly impacted congressional comportment and customs. Studies started to focus on content as well (Mergel 2012), categorizing it (Chi and Yang 2011, Golbeck et al 2010), fitting it into patterns (Glassman et al 2013) and measuring frequencies (Lassen and Bode 2013). However, none of them put their results into context, and explain the broader picture of Congressional communication.

The current state of the literature covering social media use by members of Congress is interesting. The importance of social media in messaging has been established by studies like Barbera (2014), which demonstrate that a network analysis of Congressional accounts reveals a strikingly similar result to DW Nominate. While explicitly or implicitly making this point, some are directly citing literature and theory that was not written about, and has not been validated for social media (Straus et al. 2013, Barbera 2014, Esterling et al 2011). Others are arguing for the complexity of the use of social media (Barbera (2014)) without drawing any parallels. There have been cases of specific research looking at how Twitter is being used (Golbeck et al. 2010, Shapiro et al. 2012, Graf 2008), but they were also structured around the same old framework and typologies of understanding Congressional communications. These frameworks, being crafted in an age when communication between elected officials and the citizen was comparatively limited, may be out of date.

In sum, in this paper we aim to provide further empirical evidence on the assumption that legislators use similar communication strategies across media channels while bridging the gap between the traditional literature on congressional communications and how members use social media platforms.

[1] The PEW study puts Facebook at 62% of the entire adult population, and unsurprisingly 82% of the 18 to 29 internet user population, while only 48% of the 65+ crowd. Twitter, on the other hand, sits at 20% of the entire adult population, with the 18 to 29 year olds that use the internet reaching the highest overall percentage with 32%, and those 65+ the lowest with only 6%.

Data

We test whether members of Congress use a similar communication strategy across different media channels by studying their press releases, Twitter, and Facebook messages from August 1^{st} to December 31^{st} 2014. In particular we look at whether or not they use these three media channels to talk about the same issues. We believe that this five-month period is large enough (it represents about a fourth of the 113th Congress) to have a general sense of how members behave on average.

First, we collect the press releases of members of Congress using Derek Willis' Ruby gem `statement`[2]. This gem extracts press releases and statements by members of Congress. For each member and statement, the gem returns the title of the press release and also a link to the full text. However, for the purpose of this paper, since we are interested specifically in detecting the topics of the press releases, we collect only the titles (which we consider a summary of the press releases' topic) and not their full content. Second, we collect the messages sent by each member with a Twitter account using the Twitter REST API; and finally, we collect the member Facebook messages using Facebook's API. We build our own list of member Twitter and Facebook accounts and we complement it with the 'legislators-social-media' list in the `@unitedstates GitHub`[3]. In total, we collect 156,793 messages from 473 Members of Congress. In order to compare the member messages across channels, we need to use one media channel as a baseline. In this paper we use press releases (which is the most traditional and consolidated media channel) as such a baseline. For this reason we do not include any messages (Twitter or Facebook messages) from members that did not issue a press release during the period of analysis. In the following Table 1 we provide further information about the nature of the messages in the data set.

Table 1: Summary of collected messages

	n			members			mean		
	Total	D	R	Total	D	R	Total	D	R
Press Releases	22,330	12,827	9,503	473	226	247	47.21	56.76	38.47
Twitter	97,658	54,002	43,656	426	207	219	229.24	260.88	199.34
Facebook	36,805	14,508	22,297	321	147	174	114.66	98.69	128.15

Once we have the messages from the Members of Congress, the next step is to classify them into issues in order to compare whether those that were sent at the same time but using different media channels were about the same topics. To perform this classification, we use a semi-supervised algorithm.

The Method: Semi-Supervised Naive Bayes Classifier Using EM

Political and social scientists often use written documents as data sources to test their theories and hypotheses. In the past, when text-analysis was only a manual task, researchers had to either look at a small number of texts or obtain a amount of funding to hire numerous research assistants

[2]https://github.com/TheUpshot/statement

[3]`@unitedstates` is a `GitHub` account 'made by the public, used by the public' that contains reliable data about the U.S. Congress and their members, which is commonly used by social researchers.

(e.g. Baumgartner and Jones 1993, Adler and Wilkerson 2012). Thanks to improvements in computation, in the last decade computer and social scientists have developed numerous automated methods to study political text (for a review see Grimmer and Stewart 2013). Existing automated methods serve different purposes (e.g. ideological scaling and text classification). However, for the purpose of the paper, we focus here on automated methods that classify text-documents into topics.

There are two general types of automated methods to classify text into issues: supervised and unsupervised methods. On the one hand, supervised methods are a strong and useful resource because researchers can easily check the validity of any algorithm by comparing the predicted labels with the actual labels used to train the algorithm. On the other hand, one of the main weaknesses of supervised methods is that they do not use information from the unlabeled text and so they often lack adaptability. The predictive models do not account for relevant features in the unlabeled text. This decreases the algorithm's accuracy especially when the language of the labeled and unlabeled text differ. Fully unsupervised methods have higher adaptability than supervised methods because they generate predictive models using only features of the unlabeled text. However, contrary to supervised methods, it is harder to check the validity of unsupervised methods and the resulting classification is highly dependent to the number of clusters k used by the algorithm. Some researchers have solved this validity issue by manually validating the output of several models with different k specifications (e.g. Quinn et al. 2010, Grimmer 2010, Grimmer and Stewart 2013) or using an ensemble approach to look for topics that are consistent across different k specifications (Chuang et al. 2014). However, unsupervised methods have other weaknesses apart from validation. First, the resulting classifications are case-specific and they are hard to compare to other classifications and studies. Second, they do not take advantage of existing datasets of documents that have been manually labeled following contrasted classifications and criteria (e.g. Policy Agendas Project, Congressional Bills Project, and Party Manifesto Project).

In performing our classification we are interested in some of the advantages of supervised methods but also some of the advantages of unsupervised algorithms. We want to use an existing and contrasted topic classification (in this case the PAP codebook) so that other researchers can easily compare the findings of this study to others. But we also want to use a method that incorporates features from the unlabeled text into the generative model because we know that the language of our unlabeled messages could be highly different from the language of existing documents labeled according to the PAP classification (e.g. bills, hearings, news stories...). For example, a Twitter message and a bill title can be about the same topic but their language may be very different. In sum, for these multiple reasons we decide to use a semi-supervised algorithm.

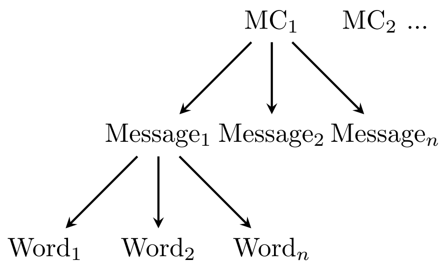
The Model

A method that satisfies this criteria is Nigam et al. (2006) *semi-supervised method using naive bayes and expectation-maximization (EM)*. In implementing this method we make the following assumptions, which are a combination of Nigam et al. (2006) and Grimmer (2010). We model the member communication strategy using a hierarchical structure (see Figure 1 for a graphical representation) with the Members of Congress (MCs) $i = \{1,\ldots,473\}$ at the top and we assume that messages are the result of a mixture of multinomials model θ with one mixture component for topic k. We also assume that during the period of analysis each MC decides the attention she wants to pay to each topic $k = \{1,\ldots,k\}$, $P(K_i|\theta)$. Then, we consider the way an MC distributes the attention among different topics represents the MC's agenda π_i (what Grimmer 2010 calls 'expressed agenda').

At the next level of the hierarchy there are the MCs' messages (press releases, Twitter and Facebook messages, etc.). Since we're looking at very short and specific messages, we assume that each message j is only about one topic and that the probability of a message j belonging to a topic k is given by the MC's agenda (π_i). We allow the attention that each MC i pays to each topic k to vary but the number of topics K is fixed across MCs in order to facilitate comparisons across them. Since each message can only be about one topic, the topic of a specific message (τ_{ij}) is a draw from a multinomial distribution.

$$\tau_{ijt}|\pi_{it} \sim Multinomial(1, \pi_{it}).$$

Figure 1: Graphical representation of the Hierearchical Model



Finally at the bottom of the hierarchical model is the content (the words) of the messages (y). We assume that the content of the message depends on its topic and so the words that compose the message (n_{ij}) are determined by a topic-specific parameter θ_k that contains information about the probability of each unique word to belong to the topic k.

$$y_{ij}|\tau_{ijk} = k, \theta_k \sim Multinomial(n_{ij}, \theta_k)$$

In sum, the full model that generates a message y_{ij} can be expressed as follows, where y_{ij} is a specific message j from a MC i, θ is a vector with the parameters of the mixture model that express a topic-probability distribution, n_{ij} is the words that compose the message, k is the topic of the message, and n_{ijw} is the number of counts of a specific unique word in the message.

$$P(y_{ij}|\theta) \propto P(n_{ij}) \sum P(k|\theta) \prod P(n_{ijw}|k;\theta)^{y_{iw}}$$

Estimation

First we estimate each parameter in θ using a fully supervised mode. We use labeled documents to calculate a Maximum Posteriori Estimate -MAP- (arg $\max_{\theta} P(\theta|X,Y)$) and find the values of θ that are most likely to generate the given documents (highest posterior probability). For each parameter in θ we assume the following Dirichlet prior distribution (a conjugate prior for multinomial distributions), with α initially set to 2 in order to favor a prior uninformative distribution.

$$P(\theta_{n_{ijw}|k}|\alpha) \propto \prod P(n_{ijw}|k)^{\alpha_w - 1}$$

We first use this prior and the labeled data to estimate the probability of each unique word belonging to each topic and also to estimate the probability of a topic being in a document (see equations 1.5 and 1.6 in Nigam et al. 2006). We use political text that has been labeled into 19 different topics according to the PAP classification. Because of the different nature of the unlabeled messages in our data set (press releases, Twitter, and Facebook messages), and because there is a possibility that the language used in them differs, we use data from three different PAP data sets as training documents in order to improve the accuracy of the predictive model: New York Times index from 1946 to 2008, Congressional Bills titles from the 112^{th} and 113^{th} Congress, (coded according to the PAP classification by the Congressional Bill Project), and quasi-sentences of State of the Union Speeches from 1946 to 2015. Out of all the available observations for each of the three data sets (49,201 for the NYT; 19,344 for the bills; and 22,020 quasi-sentences for the SOTU speeches), we only select 5,000 random observations for each of them to use as a training set. Previous work using supervised algorithms to predict PAP topics show that the accuracy of different algorithms stabilizes after few thousand training documents (Collingwood and Wilkerson 2012). However, we also end up removing from the training set those documents that have some special topics that only the NYT index have (PAP codes 23, 24, 26, 27, 29, 30, and 31) and also those documents that the coders labeled as non-informative (PAP codes -9, 0, and 99). As a result we use a training set of 13,009 labeled documents.

Then, we proceed with the semi-supervised labeling process. Since the new data (the 156,793 messages in our data set) has no labels, we cannot apply the equations we used in the supervised mode. Instead, we estimate a MAP using the EM technique (Dempster et al. 1977, Grimmer 2010). First, we start by implementing a fully supervised method and we use the parameters estimated in the previous naive Bayes model to predict the topic of each unlabeled message. Then we build a new generative model using both the labeled and unlabeled text, and by taking the predicted topics as 'true' topics. We then predict the unlabeled data again and we keep repeating the process (rebuilding the generative model and predicting) until the algorithm converges and the predictions stabilize (for a full description of the algorithm see Nigam et al. 2006). A fully unsupervised model usually uses random parameterization as starting point to find locally MAP. However in this case we use the labeled data to choose a more accurate and informative starting point.

Preprocessing

In order to improve the accuracy of the classifier we process the labeled and unlabeled text before using it to estimate political topics. We transform all the text to lower case, remove punctuation, remove stopwords, and we also stem all words using NLTK's Porter stemmer. We use as stopwords the list in the NLTK `python` module plus a set of uninformative keywords (because they appear very often in messages related to all issues) that help to speed up the computation: *congressman, congresswoman, senator, rep., congressmember, bill, authorize, extend, amend, provide, promote, require, implement, direct, reauthorize, prohibit.* As part of the punctuation-removal process we also remove all #hashtag symbols from Twitter and Facebook messages while keeping the text of the hashtag. For example, for the messages containing the hashtag #immigration we delete the symbol # but keep the word 'immigration' because it is a very informative feature that facilitates the topic modeling. However, in all Twitter and Facebook messages we remove the @at symbol that indicates that another user is being mentioned as well as the @user's name because it is an uninformative feature. Finally, we also remove all words with a length smaller than 3 because they usually are not informative and we also remove all Internet links.

Validation

One of the advantages of including a supervised component in the model is that it is easier to evaluate its validity by comparing the actual topics of the labeled text with the topics that the model predicts for them. We check how well the semi-supervised model predicts political topics by looking at the algorithm's precision and recall. Precision is the percentage of the total documents that the algorithm predicts as belonging to a specific topic that actually belong to it whereas recall is the percentage of the total documents that belong to a topic that the algorithm correctly predicts.

Table 2 shows how the algorithm performs in predicting the labels of the PAP's text we use as training documents. The results show that the average precision of the model is 78% and the overall recall is 73%. Although the precision and recall for some specific topics is quite low (e.g. the precision for the topic *Macro Economy* or the recall for *Science and Tech, Social Welfare, Housing, and Rights*), on average the algorithm produces satisfactory labels. Moreover, as the confusion matrix in Appendix I indicates, the imprecisions of the method are close to be randomly distributed since they are not systematically biased towards any specific topic. This means that the results of this study are not likely to be biased as a result of the algorithm's imprecisions.

Table 2: The algorithm's performance in predicting the labeled PAP text

	n	precision	recall
1.Macro Economy	1042	0.56	0.81
2.Rights	355	0.97	0.39
3.Health	919	0.69	0.87
4.Agriculture	225	0.99	0.38
5.Labor & Employment	428	0.93	0.50
6.Education	594	0.90	0.80
7.Environment	311	0.93	0.62
8.Energy	429	0.89	0.70
9.Immigration	128	0.98	0.41
10.Transportation	400	0.93	0.66
12.Law & Crime	725	0.82	0.71
13.Social Welfare	326	0.98	0.33
14.Housing	252	0.95	0.37
15.Finance	1220	0.77	0.86
16.Defense	1231	0.72	0.73
17.Science & Tech	266	0.96	0.32
18.Foreign Trade	788	0.97	0.74
19.Int Affairs	1720	0.61	0.88
20.Gov Operations	1171	0.65	0.78
21.Land & Water	479	0.85	0.82
	13009	0.78	0.73

Variable of Interest

To test whether the communication strategy of members of congress is similar across different media channels we build a variable that measures topic-discrepancy between the issues discussed in press releases on the one hand and political topics discussed in Twitter and Facebook messages on the other hand. For each member, month and media channel we count the number of messages related to each PAP topic. Then we use the counts to calculate the relative attention that each member paid to each topic each month in each media channel. Afterwards, for each member, topic, and month we calculate the absolute difference between the relative attention paid to that topic in press releases and the relative attention paid to the same topic in Twitter and in Facebook. Finally, for each member and month we average those absolute differences across topics and we use that measure (discrepancy scores) as a key variable of interest. We obtain a general discrepancy score (IV) that indicates the topic differences between press releases on one hand and Twitter and Facebook messages on the other one; and we also obtain partial scores (IVt and IVf) for the differences between press releases and Twitter, and press releases and Facebook. r discrepancy scores mean that the topics of a member's messages were different across media channels. The following table illustrates the process.

Table 3: Example of how the *discrepancy score* is calculated for a specific MC $i = 1$ and for $k = 3$ number of topics.

		PR			T			F			IV	IVt	IVf
	Month	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3			
MC_1	8	.2	.5	.3	.1	.3	.6	.3	.3	.4	.33	.2	.13
...	...												
MC_1	12	...	...	...	...	...	...	...	...	...	...	...	...

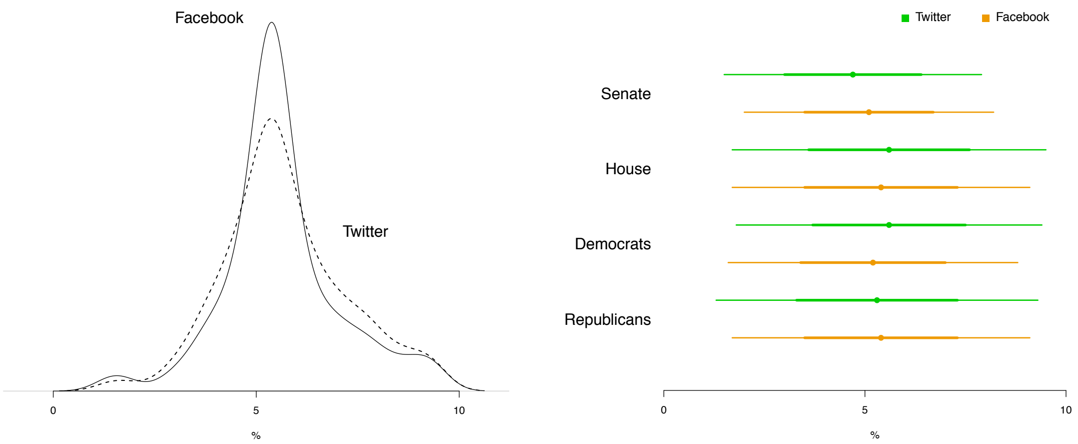
Results

A first look to the variable of interest indicates that on average members of Congress talk about the same issues in press releases, Twitter, and Facebook messages. However, this issue-correlation is not perfect. In particular, the results show that, on average and compared to how members distribute their attention across topics in press releases, each month they distribute their attention differently in about 5% of the Twitter messages and also about 5% of the Facebook messages (see Figure 2a). This means that, on average, if a member of Congress issues 100 press releases and sends 100 Twitter messages and 100 Facebook messages, 95 of the messages in each social media platform are about the same topic as 95 press releases while 5 of them are not. Moreover, these differences are constant across different types of members. For example, when we subset all the members of Congress in different groups by party (Republicans-Democrats) and by chamber (House-Senate) we observe practically the exact same differences (see Figure 2b).

Figure 2: How different the topics of Twitter and Facebook messages are from the topics of Press Releases?

(a) Overall differences

(b) Differences by party and chamber



Hence, the difference between the issues discussed in press releases and in social media does not seem to be relevant. However, how is that difference distributed? Is there any systematic component related to that difference? In other words, is there any specific topic that captures most of the 5% differential attention in press releases or social media?

Figures 3, 4, and 5 show the average percentage of attention that Republicans and Democrats paid to each PAP topic for the whole period of analysis (from August 1^{st} to December 31^{st} 2014) in press releases, Twitter and Facebook. The figures show that the attention to each issue is very similar across the three media channels. There are only some small differences when it comes to the attention that members paid to *Macro Economy*, *Government Operations*, and *Defense* and *International Affairs*. For example, Democrats paid a little more attention to macro economic issues (about 6%) in press releases than they did in social media (about 14% in Twitter and 15% in Facebook) and they paid less attention to topics related to government operations in social media (9% in Twitter and 11% in Facebook) than in press releases (17%).

However, the topic with the st difference of attention between press releases, Twitter and Facebook is the combination of *Defense* and *International Affairs*[4]. Members of Congress talk about these two issue in 25.5% of their Press Releases while they do it in 34% of their Twitter messages and 39.5% of their Facebook messages. One potential explanation for these differences is that they are the result of a bad algorithm's performance. In fact, the algorithm scores the lowest precision scores when it comes to the topics *Macro Economy*, *International Affairs*, and *Defense* (see Table 2). However this is unlikely because such error should equally affect the classification of text in press releases, Twitter and Facebook messages, and also because we are here considering defense and international affairs as one macro-topic and each of these two topics collect a part of the classification errors of the other category (see confusion matrix in Appendix I).

[4]Here we consider these two PAP topics together because the automatic methods that we use to classify the messages have trouble distinguishing between the two. This is because most of the keywords related to defense are also relevant keywords when talking about international affairs. For example, a word like *Iraq* can be more related to defense if we talk about the military conflict between the country and the U.S. but it can also be related to international affairs if we talk about providing humanitarian aid to the country.

Figure 3: Press Releases

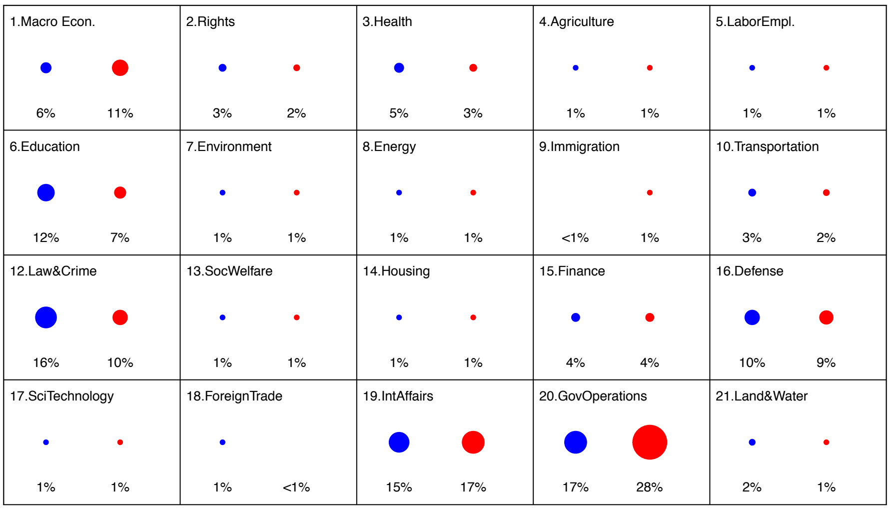


Figure 4: Twitter

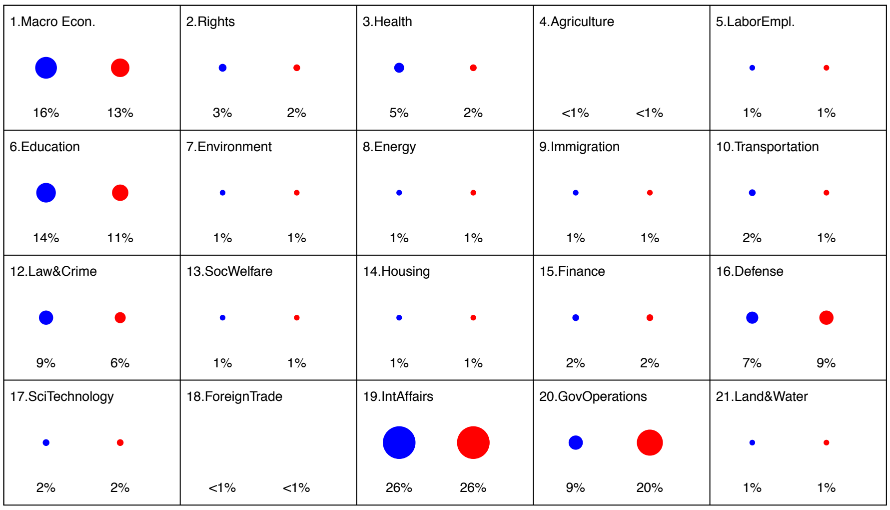
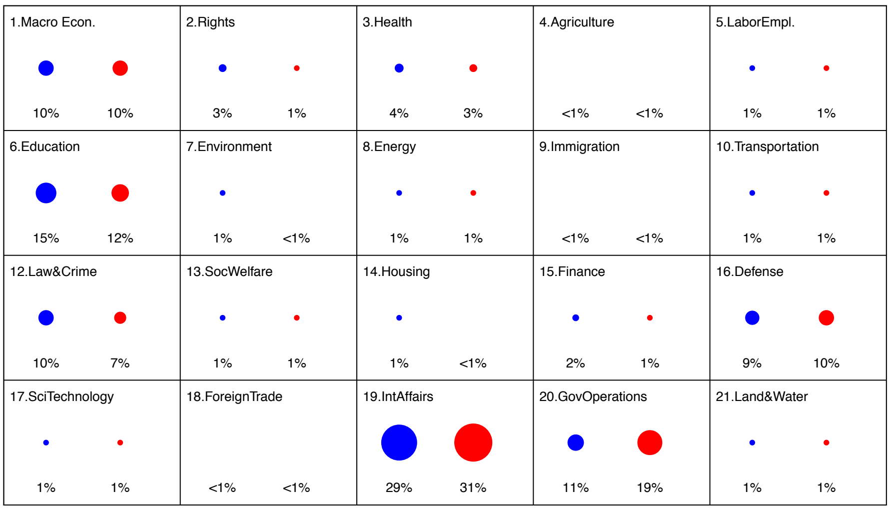


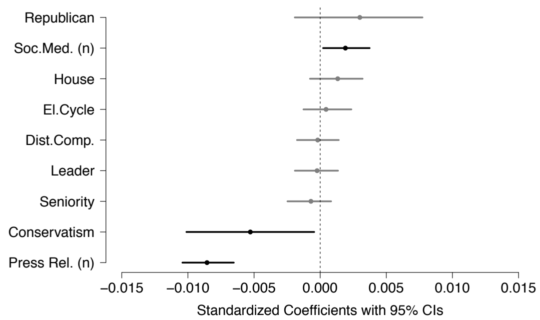
Figure 5: Facebook



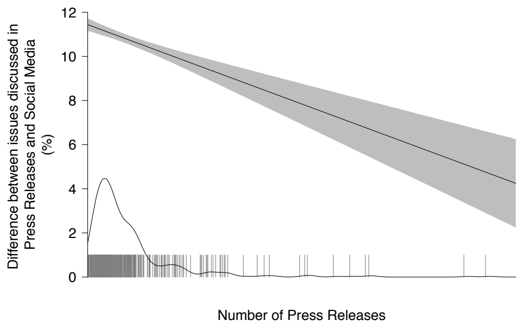
A second possible explanation is that, since members of Congress send more Twitter and Facebook messages than they issue press releases (see Table 1), when we use press releases to measure the political agenda of a member we get a slightly biased or a different measurement as a result of having fewer data points. To check the validity of this explanation we model the difference in attention between press releases and social media (the key variable of interest) using covariates that are frequently used by congressional scholars when explaining the communication strategy of members (e.g. Fenno 1978, Lazarus and Steigerwalt 2009, Grimmer 2013, Casas and Wilkerson 2014): the chamber a member serves in (*House*), the party a member represents (*Republican*), the number of days a member has been in office (*Seniority*), the member's ideology (*Conservatism*) measured using DW-NOMINATE scores, whether a member has a leadership role in one of the chambers[5] (*Leader)*, the competitiveness of the district[6], whether the member is in an election cycle (=1 for all House representatives and a third of the senators), and we also add the number of press releases, Twitter messages and Facebook messages that each sent on average every month.

Figure 6: OLS regression predicting the *"Difference between the distribution of attention in press releases and the distribution of attention in social media"*.

(a) Standardized Coefficients + 95% CIs (the effect of 1sd deviation increase)



(b) Predicted values for the Dependent Variable given different values of press releases issued by month (95% CIs).



The result of fitting an ordinary least-squares regression[7] to the variable of interest (Figure 6) shows how, in fact, the average number of press releases that a member of Congress issues every month has a significant effect on the difference between how a member distributes their attention across topics in press releases compared to how the member does it in Twitter and Facebook. The fewer press releases a member issues (or the more social media messages sent), the r the difference between the topics discussed in them compared to social media. In particular, the standardized coefficient for the number of press releases in Figure 6a (*Press Rel. (n)*) indicates that if a member increases the average number of monthly press releases from 9 to 21, the difference between the attention-distribution in press releases and the attention-distribution in social media decreases

[5] We consider that a member performs a leadership role when he/she is a chair or ranking member of a committee or subcommittee, or he/she is the Speaker of the House, the President Pro Tempore of the Senate, or the Majority or minority leader or whip of any of the chambers.

[6] We use Grimmer's (2013) methods to calculate the competitiveness of the district; which is to calculate the absolute difference between the national electoral share of Obama in the '12 elections and the electoral share that Obama obtained in the district or state of each member of Congress. We also reverse the sign of the values so that as the values increase, the higher the competitiveness of the district: $-|Obama'12_{National} - Obama'12_{district_i}|$

[7] Given that the dependent variable is a proportion from 0 to 1, we also modeled the data using a Beta regression. However, since the results of that regression were very similar to the OLS, we report here the results of the OLS regression in order to facilitate the interpretation of the results.

around .9% (.7% - 1.1%) percentage points. As Figure 6b indicates, this 1% change is quite substantial given that, compared to the topics members discuss in press releases, they usually talk about different topics in 10% of the messages in social. These results suggest that, depending on what media source researches use to measure the political agenda and the communication strategy of members of Congress, they may obtain slightly different results. In particular, these results also suggest than, since the correlation between how members distribute their attention in press releases and how they distribute it in social media converges as a member issues more press releases, social media messages may be a better source when measuring the communication strategy of members of Congress. Moreover, the coefficient for conservatism (statistically significant at the .1 level) also suggests that as more conservative a member is, the more similar that member distributes the attention in press releases and social media. In fact, a similar model with an interaction effect between ideology (*Conservatism*) and the number of press releases indicates that more conservative members are more likely to issue more press releases and so to have a more similar discourse across all media channels (see Model 1b in Appendix II).

Finally, another potential explanation is that a member-specific component explains why members distribute their attention slightly differently in press releases compared to social media messages. The *Conservatism* coefficient in the model in Figure 6 suggests that this may actually be the case. To explore this possibility a little further, we pay closer attention to how the covariates in the previous model explain the differences in attention related to the most salient issue: *Defense* and *International Affairs*.

On one hand, the literature on the communication strategy of members of Congress suggests that those members running in districts or states that are highly competitive are less likely to discuss their policy positions and also less likely to focus on issues that are highly ideological (e.g. Mayhew 1974, Fenno 1978, Carson et al. 2010, Grimmer 2013). On the other hand, the literature on issue-ownership (e.g. Petrocik 1996; Petrocik et al. 2003) shows that issues such as defense and international affairs benefit the electoral expectations of Republicans more than Democrats' when they are salient. Moreover, whereas press releases have a broad target (all media outlets), Barbera's (2014) findings suggest that the target audience of social media messages of members of Congress are individuals with extreme ideological positions and core supporters of the party and the member. Drawing from this literature one can argue that members from competitive districts may have the interest to keep a less ideological profile in press releases (to appeal to moderate voters and increase their chances of being reelected) while being more clear about their policy positions in social media (to appeal to their party base). In particular, members (especially Republicans) running in competitive districts may be interested in talking more about defense and international affairs in social media than in press releases.

To test this argument we model the difference between the relative attention that members paid to the PAP topics defense and international affairs in press releases and the mean of the relative attention they paid to the same topics in Twitter and Facebook. We use the same covariates of the model in Figure 6. When we include all the members of Congress in the model (Figure 7a) we observe again that the number of press releases and the number of social media messages that a member sends has a significant effect in explaining the difference in attention between media channels. As the number of social media messages increases, a member pays a r relative attention to those issues in social media than in press releases; and as the number of press releases increases, the opposite effect happens. Moreover, we find evidence to support the argument that members running in competitive districts are more likely to talk more about defense and international affairs in social media than they do in press releases. As district competitiveness increases, the r the difference between the relative attention in press releases and social media. Since this is a macro issue that is more electorally relevant for Republicans than for Democrats, we also run another model including only Republican members of Congress (Figure 7b). The results are very similar but in this case the effect of district competitiveness is even more substantial and significant. Figure 8 reports the predicted values of this second model given different values of district competitiveness. Smaller values of district competitiveness indicate that a member is running in a safe electoral district, while values closer to 0 mean that the member is running in a highly competitive district where Obama obtained in 2012 an electoral share equal to the electoral share he obtained at the national level. Hence, Figure 8 shows that members running in districts where Obama obtained very good results in 2012 talk more about defense and international affairs in social media than they do in press releases.

Figure 7: OLS regression predicting the *"Attention that members paid to Defense and International Affairs in press releases minus the attention they paid to those issues in social media"*.

(a) ALL MEMBERS. Standardized Coefficients + 95% CIs (the effect of 1sd deviation increase)

(b) REPUBLICANS. Standardized Coefficients + 95% CIs (the effect of 1sd deviation increase)

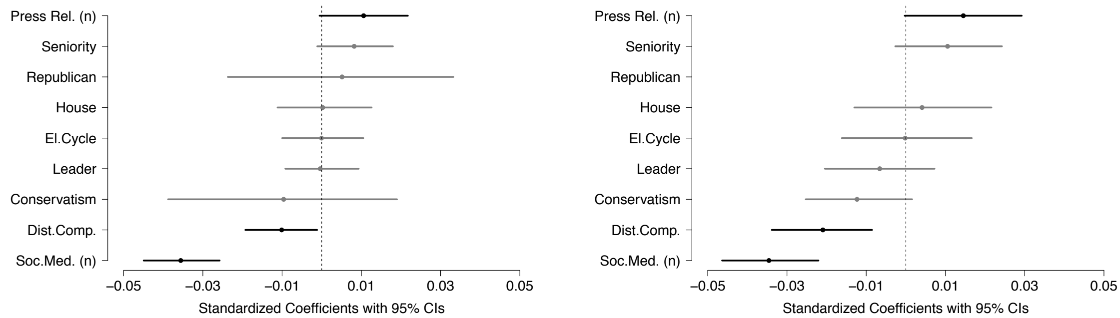
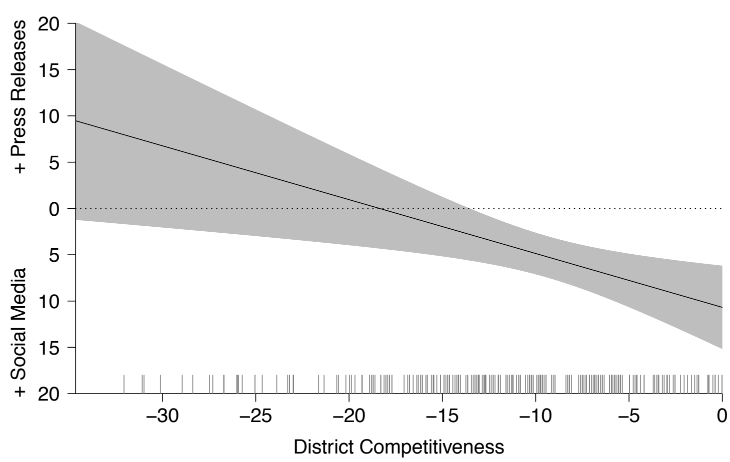


Figure 8: Predicted values of the model in Figure 7b given different values of district competitiveness (95%)



Discussion

Scholars studying the communication strategy of members of Congress often use different data sources (e.g. newsletters, TV coverage, and press releases) while assuming they are all a good proxy for the members' strategy as a whole. However, no empirical evidence exists indicating that this is actually the case. In this paper we compared the issues members of Congress discussed in three different media channels: press releases, Twitter, and Facebook. We find that members usually talk about the same issues in these three media channels (about 90-95% of the time) but we also find that this difference increases when the members issue a small number of press releases, a number of social media messages and, for specific topics, when the member represents a highly competitive districts.

These findings provide some empirical evidence to support the assumption that, on average, members of Congress use a very similar communication strategy independently of the media channel they use. However, the findings also show that researchers may obtain a slightly different measurement of the members strategy depending on the data source they use. For example, since members usually issue less press releases than they send social media messages, researchers using press releases to measure the individual agenda of a member may conclude that the member discussed a slightly smaller number of topics than if they measured the agenda of the same member including social media data. Whether these small differences may have substantive implications for congressional communication research is still to be determined. To have a clearer sense of the implications of such differences, future research should replicate some of the past studies in the field (e.g. Grimmer 2013) using multiple data sources and see if past findings still hold.

In the paper we also show that the strategies that members of Congress adopt across different media channels may be more divergent depending on the specific issues they discuss. We show that members (specially Republicans) representing highly competitive districts are more likely to talk about defense and international affairs in social media than they are in press releases. Since there is reason to believe that the target audience of social media messages is not the same as the audience of press releases (Pew 2015; Barbera 2014), drawing on the literature on congressional communication (e.g. Fenno 1978; Grimmer 2013) and issue-ownership (e.g. Petrocik et al. 2003) one could argue that members from competitive districts prefer to talk less about issues 'owned' by the party in press releases (in order to appeal moderate voters) while emphasizing them in social media (to appeal to core party members). We only provide here some evidence supporting this argument, but future research should look at a r number of Republican and Democratic 'owned' issues to see if, when discussing them, members from competitive districts distribute their attention differently in press releases than they do in social media.

References

E. Scott Adler and John D. Wilkerson. *Congress and the Politics of Problem Solving*. Cambridge University Press, New York, NY, 2012.

Pablo Barbera. Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, pages 1–16, 2014.

Frank Baumgartner and Bryan Jones. *Agendas and Instability in American Politics*. University of Chicago Press, Chicago, 1993.

Jamie L. Carson, Gregory Koger, Matthew J. Lebo, and Everett Young. The Electoral Costs of Party Loyalty in Congress. *American Journal of Political Science*, 54(3):598–616, 2010.

Andreu Casas and John Wilkerson. A Delicate Balance: Republican Party Branding during the 2013 Government Shutdown. *Paper presented at the annual meeting of the Midwest Political Science Association, Chicago*, 2014.

Carl J. Case and Darwin L. King. Twitter Usage in the Fortune 50: A Marketing Opportunity? *Journal of Marketing Development & Competitiveness*, 5(3):94–103, 2011.

Feng Chi and Nathan Yang. Twitter Adoption in Congress. *Review of Network Economics*, 10 (1), January 2011.

Jason Chuang, John D. Wilkerson, Rebecca Weiss, Dustin Tingley, Brandon M. Stewart, Margaret E. Roberts, Forough Poursabzi-Sangdeh, Justin Grimmer, Leah Findlater, Jordan Boyd-Graber, and Jeff Heer. *Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations*. Montreal, Canada, 2014.

Loren Collingwood and John Wilkerson. Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, 9(3):298–318, 2012.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rdin. Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society, Series B*, volume 39, pages 1–38, 1977.

J. N. Druckman, M. J. Kifer, and M. Parkin. The Technological Development of Congressional Candidate Web Sites: How and Why Candidates Use Web Innovations. *Social Science Computer Review*, 25(4):425–442, November 2007.

Richard F. Fenno. *Home style: House members in their districts*. HarperCollins, New York, 1978.

Charles H. Franklin. Eschewing Obfuscation? Campaigns and the Perception of U.S. Senate Incumbents. *The American Political Science Review*, 85(4):1193, 1991.

Jennifer Golbeck, Justin M. Grimes, and Anthony Rogers. Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, pages 1612–1621, 2010.

Joseph Graf. New media-The cutting edge of campaign communications. Campaigns on the Cutting Edge. In Richard J. Semiatin, editor, *Campaigns on the Cutting Edge*. CQ Press, Washington, DC, 2008.

Justin Grimmer. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, 18(1):1–35, January 2010.

Justin Grimmer. Appropiators Not Position Takers: The Distorting Effects of Eletoral Incentives on Congressional Representation. *American Journal of Political Science*, 57(3):624–642, 2013.

Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, July 2013.

Tim Groeling. *When Politicians Attack: Party Cohesion in the Media.* Cambridge University Press, New York, 2010.

G. J. Gulati and C. B. Williams. Social Media and Campaign 2012: Developments and Trends for Facebook Adoption. *Social Science Computer Review*, 31(5):577–588, October 2013.

R. J. Klotz. Internet Campaigning for Grassroots and Astroturf Support. *Social Science Computer Review*, 25(1):3–12, February 2007.

A. O. Larsson and H. Moe. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, 14(5):729–747, August 2012.

D. S. Lassen and A. R. Brown. Twitter: The Electoral Connection? *Social Science Computer Review*, 29(4):419–436, November 2011.

Jeffrey Lazarus and Amy Steigerwalt. Different Houses: The Distribution of Earmarks in the U.S. House and Senate. *Legislative Studies Quarterly*, 34(3):347–373, August 2009.

Daniel Lipinski. *Congressional communication: content & consequences.* University of Michigan Press, Ann Arbor, 2004.

A. Marwick and d. boyd. To See and Be Seen: Celebrity Practice on Twitter. *Convergence: The International Journal of Research into New Media Technologies*, 17(2):139–158, May 2011.

David R. Mayhew. *Congress: The Electoral Connection.* Yale University Press, New Haven, 1974.

Ines Mergel. Connecting to Congress: The use of Twitter by Members of Congress. *Journal of Policy Advice and Political Consulting*, 3:108–114, 2012.

Kamal Nigam, Andrew McCallum, and Tom Mitchell. Semi-Supervised Text Classification Using EM. In A. Chapelle, O. Zien, and B. Scholkopf, editors, *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006.

Rolfe Daus Peterson. 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, December 2012.

John R. Petrocik. Issue Ownership in Presidential Elections, with a 1980 Case Study. *American Journal of Political Science*, 40(3):825–850, 1996.

John R. Petrocik, William L. Benoit, and Glenn J. Hansen. Issue Ownership and Presidential Campaigning, 1952-2000. *Political Science Quarterly*, 118(4):599–626, December 2003.

Pew. Mobile Messaging and Social Media. Technical report, Pew Research Center, 2015.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, 54(1):209–228, January 2010.

Matthew Shapiro, Libby Hemphill, and Jahna Otterbacher. Doing What I Say: Connecting Congressional Social Media Behavior and Congressional Voting. *Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL*, 2012.

Colleen J. Shogan. Blackberries, Tweets, and YouTube: Technology and the Future of Communicating with Congress. *PS: Political Science & Politics*, 43(02):231, April 2010.

Jacob R. Straus, Matthew Eric Glassman, Colleen J. Shogan, and Susan Navarro Smelcer. Communicating in 140 Characters or Less: Congressional Adoption of Twitter in the 111th Congress. *PS: Political Science & Politics*, 46(01):60–66, 2013.

Christine B. Williams and Girish J. Gulati. Social Networks in Political Campaigns: Facebook and the 2006 Midterm Elections. *Paper presented at the Annual Meeting of the American Political Science Association*, 2007.

Michael Zimmer and Nicholas J. Proferes. A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3):250–261, May 2014.

Appendix I: Confusion Matrix illustrating the algorithm's performance

Real Label	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	17	18	19	20	21	Total
1.Economics	839	0	13	0	2	1	0	2	0	1	3	0	0	25	26	0	1	81	47	1	1042
2.Rights	12	137	28	0	1	6	1	0	0	0	17	0	0	8	21	2	0	62	59	1	355
3.Health	36	0	797	0	0	3	0	0	1	1	7	1	0	7	17	0	2	26	20	1	919
4.Agriculture	38	0	24	86	0	0	4	2	0	2	2	0	0	17	6	0	2	27	11	4	225
5.Labor	77	1	31	0	213	5	0	1	0	2	2	0	0	12	13	0	0	41	30	0	428
6.Education	25	1	11	0	3	473	0	1	0	0	6	0	0	12	14	0	0	32	15	1	594
7.Environment	22	0	8	0	0	2	194	8	0	1	3	0	1	10	4	0	1	18	14	25	311
8.Energy	25	0	8	0	0	1	1	302	0	0	2	0	0	14	16	0	5	35	11	9	429
9.Immigration	4	0	9	0	3	3	0	0	53	0	12	0	0	5	6	0	1	22	10	0	128
10.Transportation	23	0	5	0	1	0	2	3	0	264	6	0	0	23	31	0	0	16	23	3	400
12.Law&Crime	28	0	34	0	0	8	0	0	0	3	518	0	0	12	26	0	1	42	49	4	725
13.Soc.Welfare	52	1	72	0	0	7	0	0	0	1	6	106	0	9	12	0	0	38	22	0	326
14.Housing	24	0	22	0	0	2	0	1	0	1	4	0	93	28	17	0	0	29	28	3	252
15.Banking	63	0	15	0	0	3	1	2	0	1	4	0	1	1050	12	0	1	28	34	5	1220
16.Defense	47	0	13	0	1	0	0	1	0	0	7	0	1	8	895	0	0	231	24	3	1231
17.Science	20	0	14	0	2	3	1	4	0	0	2	0	0	43	16	86	0	43	29	3	266
18.Fgn.Trade	38	0	4	0	0	2	0	2	0	0	3	0	0	38	11	0	582	92	15	1	788
19.Int.Affairs	35	0	13	1	0	3	0	5	0	0	19	0	0	27	62	1	2	1520	31	1	1720
20.Gov.Opperations	72	1	30	0	2	1	1	1	0	4	7	1	0	22	31	1	1	75	917	4	1171
21.Land&Water	21	0	7	0	1	0	3	4	0	3	0	0	2	2	7	0	0	16	21	392	479
Total	1501	141	1158	87	229	523	208	339	54	284	630	108	98	1372	1243	90	599	2474	1410	461	13009

Appendix II: Models in Figure 6 and Figure 7

Table 4: Models in Figure 6

	Model 1		Model 1b	
	pe	se	pe	se
Intercept	0.107***	(0.005)	0.11***	(0.005)
Republican	0.006	(0.006)	0.005	(0.006)
Dist. Comp.	-0.00002	(0.0001)	-0.00002	(0.0001)
Seniority	0.000	(0.000)	0.000	(0.000)
Conservatism	-0.009*	(0.005)	-0.004	(0.005)
Press Rel. (n)	-0.001***	(0.0001)	-0.001***	(0.0001)
Election Cycle	0.002	(0.004)	0.001	(0.004)
Soc.Med. (n)	0.00003*	(0.00002)	0.00003	(0.00002)
House	0.003	(0.003)	0.003	(0.003)
Leader	-0.001	(0.002)	-0.001	(0.002)
Conservatism x Press Rel. (n)			-0.001	(0.0001)

Table 5: Models in Figure 7

	All Members		Republicans	
	pe	se	pe	se
Intercept	0.322***	(0.03)	0.326***	(0.043)
Republican	0.009	(0.034)		
Dist. Comp.	-0.001*	(0.001)	-0.003***	(0.001)
Seniority	0.000	(0.000)	0.000	(0.000)
Conservatism	-0.017	(0.031)	-0.059	(0.039)
Press Rel. (n)	0.001	(0.001)	0.002	(0.001)
Election Cycle	-0.001	(0.025)	0.0001	(0.037)
Soc.Med. (n)	-0.001***	(0.0001)	-0.001***	(0.0001)
House	0.002	(0.018)	0.012	(0.028)
Leader	-0.001	(0.012)	-0.013	(0.017)