Routledge
Taylor & Francis Group

Journal of Information Technology & Politics

ISSN: 1933-1681 (Print) 1933-169X (Online) Journal homepage: http://www.tandfonline.com/loi/witp20

#Polar Scores: Measuring partisanship using social media content

Libby Hemphill, Aron Culotta & Matthew Heston

JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
2016, VOL. 13, NO. 4, 365–377
http://dx.doi.org/10.1080/19331681.2016.1214093

WORKBENCH NOTES

#Polar Scores: Measuring partisanship using social media content

Libby Hemphill, Aron Culotta, and Matthew Heston

ABSTRACT
We present a new approach to measuring political polarization, including a novel algorithm and open source Python code, which leverages Twitter content to produce measures of polarization for both users and hashtags. #Polar scores provide advantages over existing measures because they (a) can be calculated throughout the legislative cycle, (b) allow for easy differentiation between users with similar scores, (c) are chamber-agnostic, and (d) are a generic approach that can be applied beyond the U.S. Congress. #Polar scores leverage available information such as party labels, word frequency, and hashtags to create an accessible, straightforward algorithm for estimating polarity using text.

KEYWORDS
Partisanship; polarization; political communication; politicians; social media; Twitter

Introduction

Existing political polarization measures such as DW-NOMINATE (Poole & Rosenthal, 1984) and issue partisanship (Baldassarri & Gelman, 2008) suffer from a number of limitations: they require a long time to generate data (in the case of legislative measures) or collect data (in the case of surveys); they don't allow for comparisons across chambers or groups; the underlying models are inaccessible making it difficult to explain similar scores; and the data behind them is difficult or expensive to collect. We present a new approach to measuring political polarization, including a novel algorithm (see Purpletag, 2016 for Python code used to calculate scores), that can be calculated cheaply and quickly and that empowers researchers to compare across groups.

Our approach uses Twitter activity to produce measures of polarization for both users and hashtags. First, the algorithm computes a polarization score—which we call a #Polar score—for each hashtag used by members of the U.S. Congress. This score is computed based on how aligned the hashtag is with a political party, using feature selection algorithms from the machine learning community. By aggregating these scores by user, topic, or time span, we can measure different aspects of polarization. We find that these scores correlate well with traditional measures of legislator polarization derived from voting records. However, this new approach provides several advantages in that it (a) can be calculated at any time throughout the legislative cycle, (b) allows users to easily differentiate speakers with similar scores, (c) is chamber-agnostic, potentially allowing for direct comparisons between members of the House and Senate, and (d) is a generic, language-independent approach that can be applied beyond the U.S. Congress.

Background

Social media and political communication

Social media, and Twitter in particular, are playing increasingly important roles in connecting people to political information (Himelboim, McCreery, & Smith, 2013), and politicians have taken to Twitter to provide information directly to their constituents (Hemphill, Otterbacher, & Shapiro, 2013). Nearly all members of Congress have Twitter accounts (often managed by either their Washington, DC, or campaign staff), many of which are highly active.

Twitter hashtags in political communication

A Twitter *hashtag* is a string of characters preceded by the # character (e.g., #obamacare). They are entered by the user along with the content of their message to indicate a keyword or topic associated with a tweet. In this way, hashtags provide

CONTACT Libby Hemphill libby.hemphill@iit.edu Department of Humanities, Illinois Institute of Technology, Chicago, IL 60616, USA.
Color versions of one or more of the figures in the article can be found online at http://www.tandfonline.com/WITP.

useful metadata for searching and browsing tweets. Hashtags are frequently used to organize political discussions, including elections (Gaffney, 2010), general political discussion (Small, 2011) and discussions of specific political groups and individuals (Romero, Galuba, Asur, & Huberman, 2011). Politicians' accounts use hashtags extensively—in our data, roughly 47% of all messages posted by politicians contain at least one hashtag, and every account used at least one hashtag. Tagging generally is an increasingly common activity in which users add keyword metadata to shared content (Golder & Huberman, 2006), and hashtags, much like tags in other systems such as bookmark-sharing services (Rader & Wash, 2008), mark individual tweets as relating to a topic or conversation.

Earlier studies of general Twitter use by Congress show increasing adoption of Twitter practices such as hashtagging (Golbeck, Grimes, & Rogers, 2010; Hemphill et al., 2013). These earlier studies indicate that Congress uses Twitter predominantly as a mechanism for providing information, especially about policy positions, and it is likely that hashtags play an important role in those efforts. For instance, hashtags operate as explicit *framing* attempts, where users leverage hashtags as keywords in efforts to control the rhetoric about a particular issue (Meraz & Papacharissi, 2013; Weber, 2013).

Frames are devices (e.g., metaphors, exemplars, catchphrases, depictions, and visual images) that help us organize our experiences, tools we use to make meaning of events (Entman, 1993). Frames and framing have received a great deal of attention in studies of political communication, especially in studies of news media. For instance, researchers have examined framing in discourses between news media and audiences in the Student New Left movement (Gitlin, 1980), anti-abortion protests (Pan & Kosicki, 1993), and the Iraq War (Entman & Rojecki, 1993).

Different political parties typically employ different frames within issue debates. For instance, Republicans frame abortion discussions around the baby or child and specific abortion procedures by using words such as "baby" and "procedure," while Democrats frame the same issue around women and choice by using words such as "women" and "right" (Monroe, Colaresi, & Quinn, 2009). Through careful word selection, communicators create frames that can influence audience's choices and behaviors (Scheufele & Tewksbury, 2007). Our measure assumes that politicians are using hashtags to accomplish framing and that different parties select different hashtags based on the frames they wish to establish.

Studies of politicians' and political communication use of Twitter in other countries suggests similar patterns (Graham, Jackson, & Broersma, 2014; Larsson & Moe, 2012; Small, 2011). Researchers also identify a specific political hashtag practice—wave-riding (Christensen, 2013) or hijacking (Jungherr, 2014; Weber, 2013)—in which users appropriate a hashtag used by their political opponents in an attempt to piggyback on the tag's attention and reorient the related narrative. Hijacking indicates that hashtags are used in framing attempts.

Ideology and polarization measures

Roll call votes, monetary contributions, and, more recently, text, are the data most often used to create ideology and polarization measures. NOMINATE and its later iterations (e.g., DW-NOMINATE) (Poole & Rosenthal, 1984, 2007) are the most often used measure of polarization, and all NOMINATE measures are based on roll-call voting. NOMINATE, which stands for "nominal three-step estimation," is a family of multidimensional scaling methods for analyzing voting choices. DW-NOMINATE (the DW stands for "dynamic, weighted) is the latest measure; the scores produced fall between −1 (liberal) and 1 (conservative), and are the most commonly referenced measures of polarization in Congress. NOMINATE analysis has shown that party delegations are increasingly homogenous and polarized (Poole & Rosenthal, 1984). The original NOMINATE scores did not include information about uncertainty or bias, but recent updates include parametric bootstrapped standard errors (Lewis & Poole, 2004). All NOMINATE measures and variants that rely on roll-call votes face similar challenges in that they provide scores only for those already in office, are available only after legislative actions are taken, are available only yearly, cannot be compared between chambers, and are limited to the U.S. Congress.

Another variant of the NOMINATE method, PAC-NOMINATE (McCarty & Poole, 1998), uses contribution data to measure polarization. In this case, contributions to incumbents are treated as votes for him, while contributions to challengers are votes against the incumbent. PAC-NOMINATE assumes that contributions are earned based on policy positions and not the other way around—that is, a PAC gives money to a candidate because of where he stands and not in order to induce a stance. Violations of this assumption present challenges for this measure. For instance, if an organization attempts to influence a politician's stance by giving him money, PAC-NOMINATE treats that situation as though the politician is already aligned with the organization trying to buy him off. Also, as Bonica (2014) points out, PAC-NOMINATE is useful only in races with viable challengers, and he introduced an alternative measure of polarization based on campaign contributions. Bonica's CF scores account for both the scale (i.e., dollar amount) of the contribution and its occurrence and use all campaign contributions instead of just those made by PACs. This approach improves on roll-call vote measures by allowing us to estimate ideology before legislative actions are taken and to compare across legislative bodies, districts, and even types of politicians. However, these measures use proxies for positions in order to estimate ideology. Contributions are choices, but they aren't choices made by the politicians themselves. PAC variants of NOMINATE scores are also subject to timing limitations—they depend on campaign contributions and are less indicative in off-cycle time periods and for incumbents not actively campaigning.

A third kind of polarization measure that relies on political text is gaining ground. Studies that use "text-as-data" to measure partisan polarization have analyzed political speeches, party manifestos, and legislative bills. Classifying politicians based on their speech allows us to estimate their underlying ideology and to use that estimation to predict their future opinions (Yu, Kaufmann, & Diermeier, 2008) and to estimate ideological difference between politicians. For instance, we determine the extent to which individual words are used by opposing parties (Monroe et al., 2009). Researchers have recommended feature selection (Monroe et al., 2009), text classification (Yu et al., 2008), content analysis (Lowe, 2008), and scaling algorithms (Slapin & Proksch, 2008) as approaches for estimating ideology or polarization from political texts.

One commonly used approach—Wordscores—treats words within political texts as data and provides a language-independent technique for estimating policy positions within text and for comparing texts to one another (Laver, Benoit, & Garry, 2003). In a critique of Wordscores, Lowe argues that without an underlying statistical model, we can't know what assumptions Wordscores make about scores and words and so can't tell when they are a useful analysis tool (Lowe, 2008). Wordscores also rely on the existence of reference texts for each party and cannot be used in the absence of these texts (originally party manifestos). Our measure instead assumes there are exemplar authors who carry party labels rather than exemplar texts.

Ours is not the first project to leverage hashtags in estimating ideology. Prior research has used hashtags to estimate political ideology for public users, not just politicians (Conover et al., 2011; Weber, 2013) and treated political hashtags as "discursive clusters" (Bode, Hanna, Yang, & Shah, 2015) to study candidates' strategic rhetoric. Commercial Web sites such as WeFollow.org and Persecuting.us also use hashtags to estimate ideology. In both academic and commercial uses, hashtags have shown to be useful for detecting political affiliation of the users who adopt them.

In summary, the limitations of existing measures are that they cannot be calculated throughout the legislative or campaign cycles, that they do not afford cross-chamber comparisons, that they rely on expensive or difficult-to-collect data, and that they lack underlying models that explain their utility. We address these limitations by relying on data that is freely available and easily captured at any time using existing computational tools, allowing individuals in different groups to reside in the same data set, and detailing the underlying assumptions in the algorithm. In this way, we create a measure that is time-independent, population-agnostic, and transparent and that allows researchers to determine whether it's suitable for their needs. The #Polar scores we present use the text-as-data approach to produce two measures—one of political actors and one of political language—that improve on earlier measures by being timely and enabling comparison across a variety of actors and legislative bodies. We expect #Polar scores to correlate well with existing polarization measures for Congress but to improve our abilities to make comparisons between groups, to afford anytime calculation, and to easily differentiate users with similar scores by making the underlying features transparent.

Methods

We identified verifiable accounts associated with individual members of Congress (MoCs) and collected all tweets posted by those accounts using Twitter's REST API. We identified 12,677 different hashtags used by 473 different users in 55,244 different tweets between April 1, 2012, and September 30, 2012.[1] Note that not all tweets were actually authored by the politicians themselves. Rather, many politicians employ staffers and firms to manage their social media presence. We treat their social media accounts as brands of sorts rather than individual accounts. Regardless of who actually sends tweets, they are posted on a politician's behalf and in line with his team's messaging plan. Tweets function like many public statements politicians and their offices make—for example, press releases, speeches—as part of a politician's broader communication strategy.

Our goal is to develop a measure that indicates how polarized a hashtag is based on its usage. Our primary assumption is that a polarized hashtag is one that is highly predictive of the political party of its author. That is, if we observe that an MoC uses a particular hashtag, how easily can we predict her political affiliation? Phrased this way, our problem maps directly to the problem of *feature selection* from the machine learning and statistics literature (Guyon & Elisseeff, 2003). Below, we describe a methodology to apply feature selection algorithms to quantify the polarization of hashtags and, consequently, MoCs.

Feature selection models for estimating political ideology

Hashtags, like keywords, are evidence of intent to position one's self in relation to an issue or person. Because not all hashtags are necessarily used to position (e.g., #ff for "follow Friday"), we need a way to identify *positioning hashtags* (e.g., #obamacare, #aca). To do this, we assume that different political parties use different positioning strategies. It follows that hashtags whose usage differs significantly between parties are likely to be positioning hashtags. To quantify this, we turn to the *feature selection* literature of machine learning and statistics (Guyon & Elisseeff, 2003).

In machine learning, a *feature* is a measurable property of a phenomenon, and a *class* is the category to which a given observation belongs. *Classification* is the problem of estimating a function that accurately maps an observation to its proper class. In our case, hashtags are features, political parties are classes, and MoCs are observations. We represent each MoC as a binary vector indicating which hashtags he has used. For example, if only two hashtags, #tcot and #aca, are considered, then each MoC will be represented by a vector of length two, where the first element represents the presence of #tcot, and the second element represents the presence of #aca. Thus, a MoC who mentions only #tcot is represented by the vector {1,0}, while a MoC who mentions both hashtags is represented by {1,1}. One could also represent each MoC by a count vector, which considers the number of times an MoC has used a hashtag instead of just its presence or absence. However, doing so would allow one prolific user to bias the results. For example, Rep. Tim Griffin (R-AR) used the tag #ar2 in 967 different tweets. If a count vector were used, the feature selection algorithm would rank the #ar2 highly, since it is so predictive of the Republican Party.

Generically, feature selection algorithms determine which features (hashtags) are most useful for determining class (party). Feature selection algorithms typically proceed by analyzing a set of observations for which the classes are known and assigning a real-valued score to each feature, where a larger score means the feature is more predictive of class. We use the score assigned to each hashtag to quantify the likelihood that the hashtag was used with positioning intent. We compare three algorithms (see Guyon & Elisseeff, 2003 for mathematical details):

- **Information gain**: Computes the decrease in entropy of the class label distribution when a feature is included compared with when it is not.

- **Chi-squared**: Computes the chi-squared test statistic for the null hypothesis that the class label and feature value are independent.
- **Log odds ratio**: Computes the log of the odds of a feature appearing in one class divided by the odds of it appearing in the other class.

Evaluating selection algorithms

To determine which algorithm is most appropriate for our data, we follow the standard approach of evaluating each method by the party classification accuracy it produces, across a range of feature sizes. Here, the classification task is to predict the party of an MoC based on the set of hashtags that he or she has used. Average accuracies on held-out data are computed using k-fold cross-validation ($k = 10$). That is, given a labeled set of observations D, a feature selection algorithm F, and a maximum feature size m, we do the following:

- Split D into k equal-sized sets $D_1 \ldots D_k$
- For each set
- Construct $D_{\text{train}} = D \setminus D_k$; $D_{\text{test}} = D_k$
- Rank features in D_{train} according to F
- Retain the top m features
- Fit a classifier on D_{train} using only the selected m features
- Predict the class assignments for the held-out observations in D_{test}

We compute the average accuracy over the k sets D_{test} for each feature size m. Good feature selection algorithms should produce higher accuracies than bad algorithms across a range of values for m. Figure 1 displays the average accuracy (and standard error) for each algorithm using many feature sizes. For all results, we use a Naive Bayes classifier[2] (Hastie, Tibshirani, & Friedman, 2009).

The results indicate that we can classify MoCs by party with over 95% accuracy by examining the presence or absence of only 100 hashtags. The best result is 97.67% accuracy using 1,000 hashtags selected by chi-squared. Information gain and chi-squared feature selection strategies perform comparably, and both are superior to log odds. Averaged across all feature sizes, both chi-squared and information gain have an average accuracy of 95.44%, compared with 91.77% for log odds. Given chi-squared's performance and simplicity, we use it in subsequent experiments. Lowe and colleagues (2011) argue for the use of log odds ratios in scaling measures, but our results show that a chi-squared approach outperforms log odds.

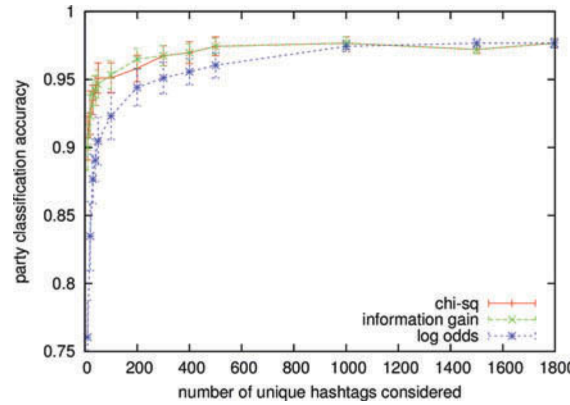

Figure 1. Comparing classification algorithms. The graph shows average accuracy (and standard error) for each algorithm using feature sizes in {10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 1500, 1797}.

Although many dimensionality reduction algorithms exist (e.g., LDA, Blei, Ng, & Jordan, 2003 or PCA, Jolliffe, 2002), these are primarily applied in an *unsupervised* way; that is, assuming there are no provided labels for each MoC. In this case, however, because we know the party affiliation of each MoC, we would like to find a reduced dimension that best reflects these labels. For example, simply running LDA on the MoC-hashtag matrix will likely find many topical groupings of hashtags (e.g., healthcare, immigration), but we are primarily interested in partisan grouping of hashtags, regardless of their topics. Although there exist supervised extensions to some of these algorithms (e.g., supervised LDA, McAuliffe & Blei, 2008), we additionally desire a methodology that is simple and transparent. Methods like PCA and LDA represent clusters as weighted combinations of hashtags, which may be difficult for humans to interpret. The advantage of using chi-squared feature selection is that we can assign a single value to each hashtag indicating its polarity score, and we can directly compare the values of hashtags to assess their relative polarity.

Results

First, we provide descriptive statistics of the data set, analyzing hashtag frequency and highlighting party differences. We then describe algorithmic approaches for estimating ideology using hashtags. In general, we found that Democrats and Republicans discuss similar issues but use different hashtags to do so. MoCs with extreme #Polar scores are individuals we would expect to engage in polarizing social media conversations because of their positions or behavior elsewhere. We also found strong correlations between #Polar measures and DW-NOMINATE scores.

General hashtag use

We first looked to see whether politicians use hashtags at all and provide an overview of their hashtag use. Table 1 presents two different measures of use:

- users/hashtag: how many users ever tweeted the hashtag
- total uses: a raw score of how many times a given hashtag was tweeted by any user.

Table 2 displays the most used tags along with the most popular hashtags for each major party. Distributions of users/hashtag and total uses are all heavily skewed. Most hashtags are used by only one user, used only once by any user, or used only once by anyone. Among the most used hashtags, those used by many users (high users/hashtag) and used in the most tweets (total uses) are general topic tags such as #JOBS and #SCOTUS that likely matter to broad constituencies. Parties use different hashtags to talk about the same issue. For instance, Republicans use #fullrepeal and #obamacare to talk about health care while Democrats use #ACA and #getcovered. General tags that describe political issues (#healthcare or #budget) are comparatively quite rare and have little impact on an MoC's #Polar score.

Table 1. Summary statistics of users/hashtag and total uses of hashtags.

	Mean	SD	Median	Min	Max
Users/hashtag	2.8	8.9	1	1	260
Total uses	9.0	68.9	1	1	3415

Comparing hashtag use between groups

As seen in Table 2, the issues Democrats and Republicans discuss overlap, but the hashtags they use to mark their conversations differ. The top of the Democrats' list includes healthcare (#ACA, for the Affordable Care Act), student loans (#DontDoubleMyRate), and employment (#JOBS). The Republicans' top issues are similar: employment (#4jobs), themselves (#tcot), and healthcare (#Obamacare).

As Table 2 shows, among the top 15 most frequently used hashtags in each party, only five appear on both lists, but overall tag frequency values for Republicans and Democrats were strongly correlated, $r(10{,}545) = 0.23$, $p < 0.001$. These analyses

Table 2. Most used tags along a number of measures of use.

Total users		Total tweets		Top tags (Democrats)		Top tags (Republicans)	
Tag	Users	Tag	Tweets	Tag	Count	Tag	Count
JOBS	260	4jobs	3415	ACA	783	4jobs	3413
SCOTUS	215	tcot	3070	DontDouble MyRate	634	tcot	3047
Obamacare	202	JOBS	2080	JOBS*	558	Obamacare	1926
gop	191	Obamacare	2054	VoteReady	539	smallbiz	1691
FF	186	smallbiz	1834	VAWA	502	JOBS*	1456
4jobs	165	gop	1383	gop*	462	stopthetaxhike	1260
smallbiz	153	stopthetaxhike	1268	EqualPay	425	FastAndFurious	1082
ACA	153	FastAndFurious	1115	FF*	421	ar2	1017
DontDoubleMyRate	149	FF	1027	hcr*	366	gop*	916
tcot	146	ar2	1017	p2	350	Energy	749
VAWA	142	ACA	865	FarmBill	332	FullRepeal	727
FullRepeal	132	Energy	857	netDE	331	FF*	604
Veterans	132	SCOTUS	745	Veterans	322	SCOTUS*	450
hcr	130	DontDoubleMyRate	739	SCOTUS*	254	Holder	408
Energy	126	FullRepeal	730	NJ	236	stribpol	385

Note. *Tag appears on both parties' top 15 lists.

suggest that Congress has converged on a set of hashtags. However, given the skew of the distributions for tag use, these raw counts of hashtag frequency overestimate a tag's popularity. In the next section, we present results from our algorithmic approaches to identifying ideological positioning through hashtags. Our approaches provide more robust means for comparing between groups than correlation allows.

#Polar-Hashtag scores

Table 3 lists the top 15 hashtags sorted by chi-squared value. Note that five of the top 15 do not appear on the list of most frequent hashtags in Table 2. This is because the chi-squared measure accounts for the relative frequency across classes, giving a clearer picture of framing relevance. It is interesting to note that #obamacare is the third most frequent hashtag used by Republicans, but is only the 15th-ranked hashtag according to chi-squared. Such phenomena reflect a hijacking or wave-riding process: the hashtag was started by and used early by Republicans but was then partially co-opted by Democrats, thereby diluting its #Polar score.

To calculate #Polar-Hashtag scores, we computed a signed version of chi-squared, in which positive values are predictive of Republican MoCs and negative values are predictive of Democratic MoCs, and the sign depends on the party for which the hashtag probability is larger. We use positive values for Republicans and negative for Democrats because the first dimension of DW-NOMINATE, the most common measure of political polarization, uses the same scale (Lewis & Poole, 2004). Thus, in Table 3 #4jobs has a signed value of +129.7 because relatively more Republicans use it, while #aca has a signed value of −111.0 because relatively more Democrats use it.

Table 3. Top 15 hashtags: Tags are ranked by their chi-squared results, and we indicate how many MoCs ever used the tag.

Hashtag	Chi-squared	# MoCs
4jobs	129.7	162
aca	111.0	150
fullrepeal	99.3	128
equalpay	86.1	80
tcot	84.3	140
dontdoublemyrate	75.7	144
stopthetaxhike	74.8	118
middleclasstaxcuts	62.7	53
lgbt	55.6	47
gopnmc	49.9	59
equalpayday	47.3	40
disclose	44.4	48
vawa	44.3	136
voteready	44.2	41
obamacare	43.1	194

Figure 2 displays the signed chi-squared score for each hashtag, along with the total number of distinct MoCs who use it. Overall, we see that many tags have small signed chi-squared values, even those such as #jobs that are used by many MoCs. The tag #jobs has a small chi-squared value because members of both parties frequently use it, highlighting the limitations of using raw frequency as an indicator of polarization. Republicans appear to prefer the tag #4jobs to the more ambiguous #jobs. Other tags such as #scotus, #veterans, and #medicare that do not take clear policy positions also appear near the midline. Tags used in discussions about contentious issues such as the Affordable Care Act and the Lily Ledbetter Fair Pay Act do show measurable signed chi-squared values. In discussing the Affordable Care Act, #aca is more likely used by Democrats while #obamacare and #fullrepeal are more likely used by Republicans. The Lily Ledbetter Fair Pay Act provides an interesting case because Democrats are likely to talk about it—as evidenced by the #equalpay and #equalpayday tags—but Republicans don't seem to talk about it at all. There is no clear counter tag with positive chi-squared value. The relative attention a given topic receives from the

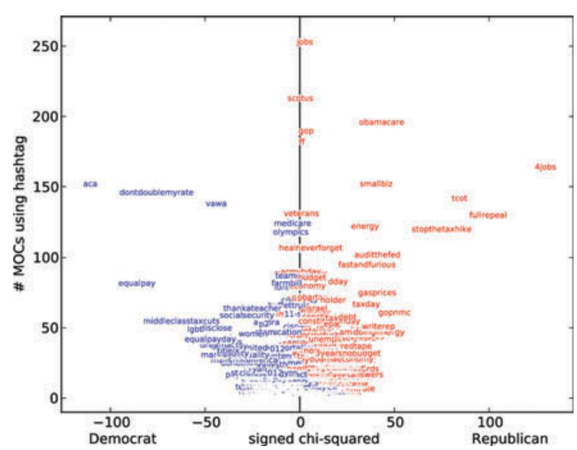

Figure 2. Hashtags' signed chi-squared values and number of MoCs who used them. Red tags have positive signed chi-squared values (more likely used by Republicans), blue negative (more likely used by Democrats).

parties necessarily influences the #Polar-Hashtag scores. For instance, if an MoC talks mostly about women's issues, and all the tags about women's issues are highly associated with Democrats, then we believe it is correct to assign this MoC a strongly liberal score.

The general topic of health care provides an interesting example of what happens when both parties talk about a particular issue. A person who talks a lot about health care using #obamacare, #fullrepeal, and other conservative tags will be assigned a strong conservative score. Similarly, someone using the tags #ACA or #getcovered will be assigned a strong liberal score. Only someone who uses tags from both sides (e.g., #ACA and #fullrepeal) will receive a low polar score. In practice, using tags from both sides is rare, and thus health care appears as a polarized topic overall. We did not compute general topic scores, though if one wished to assign #Polar scores by topic, one could simply restrict the analysis to tags on a particular topic, producing MoC-topic polarity scores.

Our approach allows us to calculate polarization scores for tags at any point in time. Figure 3 shows how the #Polar-Hashtag score for the tag #getcovered changed over time, for instance. The hashtag #getcovered was used mostly by Democrats to encourage constituents to purchase health insurance under the Affordable Care Act. Calculating #Polar-Hashtag scores over time allows us to compare #getcovered with a related hashtag used by Republicans to disparage the ACA: #trainwreck (see Figure 4). The hashtag #getcovered became popular only near the end of 2013

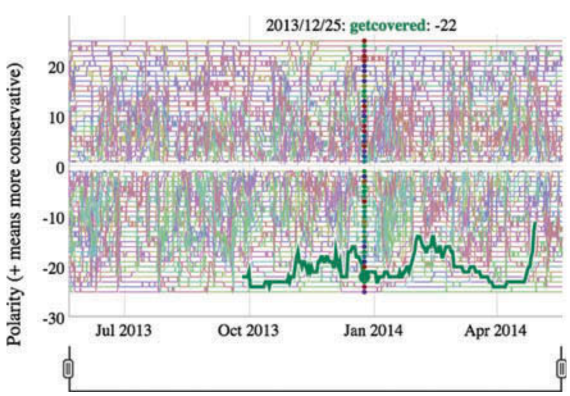

Figure 3. #Polar-Hashtag scores for #getcovered over time. Graph includes 100 most conservative and 100 most liberal tags for each day. Detail is shown for December 25, 2013, when #getcovered had a #Polar-Hashtag score of −22, near the middle of its score for the period between May 2013 and May 2014.

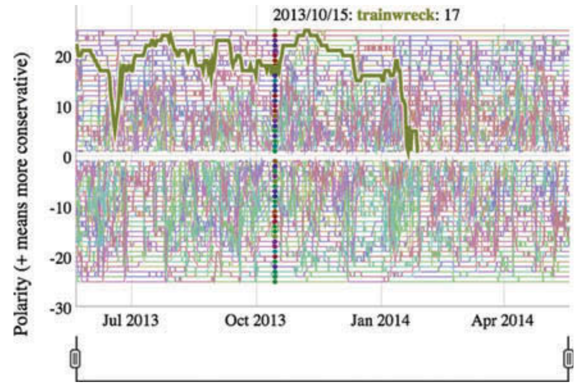

Figure 4. #Polar-Hashtag scores for #trainwreck over time. Graph includes 100 most conservative and 100 most liberal tags for each day. Detail is shown for October 15, 2013, when #trainwreck had a #Polar-Hashtag score of 17.

while #trainwreck had already been popular for months. Then we see that #trainwreck disappears during the first quarter of 2014 while #getcovered is used further into the year. This kind of day-to-day or even month-to-month comparison of polarization is not possible using existing popular polarization measures such as NOMINATE.

#Polar-User scores

We computed the aggregated #Polar-User score by summing together the signed scores for each unique hashtag they used. Figure 5 shows the aggregated signed chi-squared scores for each MoC. The "U" shape in the graph indicates that people with extreme signed chi-squared results also use many different hashtags. Although more sophisticated statistical methods may be used here to combine hashtag scores into MoC scores, we use a sum for simplicity and transparency. In the next section we provide empirical evidence that this simple approach aligns well with existing measures.

Comparing #Polar Scores to existing partisanship measures

We compared #Polar Scores with two existing measures based on voting behavior (DW-NOMINATE) and text (Wordfish). DW-NOMINATE scores are based on roll-call voting records and are often used in analyses of political polarization (Lewis & Poole, 2004), and here we compare them to our signed chi-squared measure. The first dimension of DW-

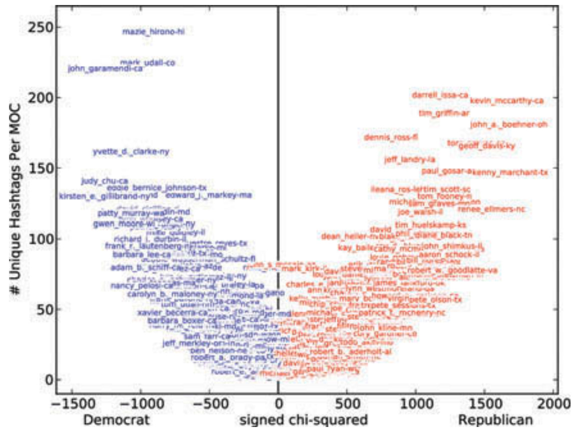

Figure 5. Aggregated signed chi-squared scores for individual MoCs. Individuals are labeled by their name and the state they represent. MoCs are colored red if they are Republicans, blue if they are Democrats.

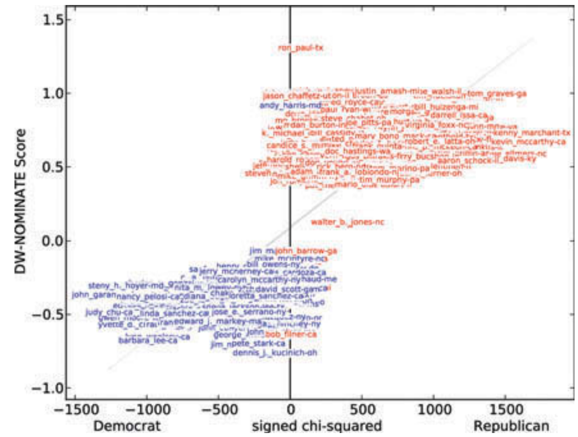

Figure 7. Comparing signed chi-squared and DW-NOMINATE results (House). Individuals are labeled by their name and the state they represent. MoCs are colored red if they are Republicans, blue if they are Democrats.

NOMINATE roughly maps to the liberal–conservative continuum. Figures 6 and 7 plot our signed chi-square value (x-axis) against the first dimension of the DW-NOMINATE score (y-axis). DW-NOMINATE scores are not comparable across chambers, so we include figures for both the Senate (Figure 6) and House (Figure 7).

We find a strong correlation between signed chi-squared and DW-NOMINATE scores in both the House, $r(331) = 0.80$, $p < 0.001$, and the Senate, $r(76) = 0.83$, $p < 0.001$. DW-NOMINATE scores vary little among both Democrats and Republicans. We see that most politicians are about as polarized in their rhetoric on Twitter as we would expect based on how polarized they are in their voting records. For instance, Rep. Marchant (R-TX) and Sen. Gillibrand (D-NY) had the highest and lowest signed chi-squared scores. Both also had high and low DW-NOMINATE scores, demonstrating their consistent conservative and liberal voting records. They talk and vote along the same polarized lines. However, we find that discrepancies between DW-NOMINATE and chi-squared scores can provide more nuanced insight into messaging strategies of MoCs, allowing us to differentiate candidates with similar DW-NOMINATE scores. For instance, Sens. Casey (D-PA) and Shaheen (D-NH) have nearly identical DW-NOMINATE scores (−0.345 and −0.341, respectively) but very different signed chi-squared scores (−85 and −591). That tells us that, on average, their voting records look similar, but their rhetoric is very different. Sen. Shaheen is much more polarizing in her language than her voting record suggests. She uses hashtags such as #aca, #equalpay, #dontdoublemyrate, #vawa, and #lgbt, all of which are predominantly used by Democrats. On the other hand, Sen. Casey uses a few strongly Democratic hashtags (e.g., #dontdoublemyrate), but also uses some hashtags associated more with Republicans (#dday, #usarmy), resulting in his more moderate chi-squared score.

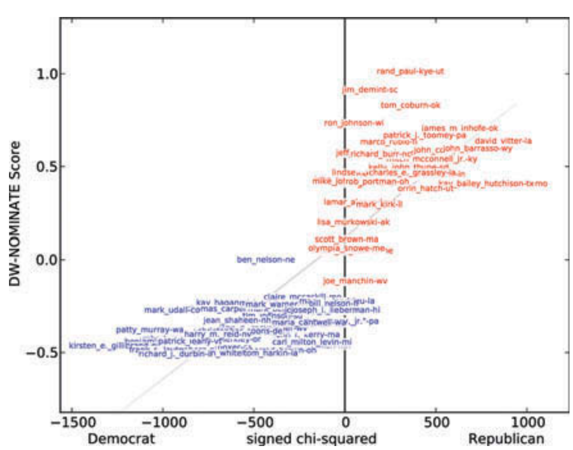

Figure 6. Comparing signed chi-squared and DW-NOMINATE results (Senate). Individuals are labeled by their name and the state they represent. MoCs are colored red if they are Republicans, blue if they are Democrats.

Table 4 lists the eight MoCs who talked and voted most differently: seven Democrats and one

Table 4. MoCs whose chi-squared and DW-NOMINATE signs differ, indicating that they tweet and vote differently

MoC	Signed chi-squared	DW-NOMINATE	Rank diff	Unique hashtags
Rep. John Barrow (D-GA)	111.9	−0.086	6	35
Sen. Joe Manchin (D-WV)	58.4	−0.128	7	6
Rep. Jason Altmire (D-PA)	33.3	−0.137	7	16
Rep. Sandy Levin (D-MI)	12.9	−0.337	7	29
Rep. Larry Kissell (D-NC)	9.9	−0.161	8	5
Rep. Ruben Hinojosa (D-TX)	1.1	−0.323	8	1
Rep. Bob Filner (D-CA)	1.1	−0.654	8	1
Rep. Andy Harris (R-MD)	−0.5	0.900	8	1

Republican. The "Rank Diff" column shows how different a user's chi-squared and DW-NOMINATE scores are based on their rank order under each metric; a negative rank difference indicates a user is more Republican in his talk than in his voting record, and a positive rank difference indicates a user is more Democratic in his talk than in his voting record. We included "Unique Hashtags Used" because some MoCs used just a few hashtags, but those hashtags were very polarizing. Reps. Hinojosa (D-TX), Filner (D-CA), and Harris (R-MD), for instance, seem to use different rhetoric than their voting records would suggest, but they each used only one hashtag. In each case, that particular hashtag was more often used by members of the other party. When we remove MoCs who used just one hashtag, no Republicans appear on the list of people who tweet differently from how they vote, and even those Democrats who remain on the list don't have large differences between their ranks according to signed chi-squared and DW-NOMINATE.

Among those remaining Democrats who tweet differently than they vote are some of Congress's most conservative Democrats. For instance, Rep. John Barrow's (D-GA) signed chi-squared ranking is lower than we would expect because he used hashtags such as #jobs, #NoShowNoPay, and #CutTheFleet. #Jobs was used far more often by Republicans. #NoShowNoPay and #CutTheFleet both refer to bills aimed at cutting spending.[3] It is not surprising to see him talk this way because Rep. Barrow is widely recognized as a conservative Democrat and has a mixed voting record that accounts for his nearly zero DW-NOMINATE score. Sen. Manchin and Reps. Altmire and Levin are similarly conservative compared to their Democratic colleagues, and we expect them to use some Republican frames.

As an additional comparison with prior work, we compute scores for each MoC using Wordfish, a scaling algorithm that estimates policy positions from text (Slapin & Proksch, 2008). Although Wordfish was originally designed to analyze party manifestos, here we apply it to the MoC-hashtag frequency matrix, using the same subset of hashtags selected by #Polar. When we compare the resulting Wordfish scores with DW-NOMINATE, we find a comparable correlation in the Senate (.81 vs. .83 for #Polar-User); however, the correlation in the House is poor (.22 vs. .80 for #Polar-User). Examining the most polarized hashtags according to Wordfish, we find that hashtags used by a small number of users have an outsized influence on the results (c.f. the #ar2 discussion above). We suspect this problem stems in part from Wordfish's assumption of a Poisson event model (as opposed to Bernoulli), and in part from the fact that Wordfish does not use the MoC's party affiliation when estimating parameters. Whereas #Polar uses party affiliation to guide the scores assigned to each hashtag, Wordfish instead assumes that the discovered dimension corresponds to the left–right political spectrum. In these data, it appears that the discovered dimension is strongly associated with location; indeed, eight of the top ten most "conservative" hashtags and seven of the top ten most "liberal" hashtags discovered by Wordfish are indicators of location (e.g., #ar2, #ny24, #az08, #wagov). By using the existing party labels and a chi-square approach, #Polar scores are able to temper the influence of such locative tags.

Likely confounds and considerations

Although #Polar scores correlate well with existing measures and exhibit face validity, a number of confounds likely influence their calculation. First, time clearly influences what MoCs talk about. Our own prior work (Hemphill et al., 2013) shows that MoCs do more political positioning on Twitter around primary elections than around general elections, for instance. MoCs seeking reelection may deploy less polarizing language around general elections as a strategy for appealing to the median voter (Downs, 1957). We provide time period and lag options in #Polar score calculations to allow researchers to control for time effects, and the ability to adjust these time parameters enables us to detect changes in rhetorical strategy in the first place.

Second, general topic influences #Polar scores. We see evidence that different parties use different hashtags, keywords, and phrases to discuss similar political issues, and this kind of framing also influences the eventual #Polar scores. Whereas other polarization measures estimate or aggregate topics (e.g., Slapin & Proksch, 2008), we do not collapse data into topics. Every tweet posted by an MoC can be annotated with metadata such as party, chamber, tenure, and so on. Instead of trying to estimate the latent polarity in individual tweets, we use the party labels as proxies for partisanship. In this way, hashtags that are most affiliated with parties are most affiliated with polarity. Researchers could cluster hashtags into topics, either manually or algorithmically, and then limit #Polar score calculations within those topics. Our measure assumes that the choices of both general topics and specific hashtags are motivated by underlying ideology, and therefore, the tags themselves reflect partisanship.

MoCs often use locative tags, and they may also influence #Polar scores even though they are not explicitly political. Our measure includes controls for the number of MoCs using a given hashtag that limit the influence of locative hashtags. For instance, #ar2 is used frequently, but only by a small number of MoCs, and its overall score reflects its limited popularity.

Researchers should make cross-chamber comparisons with caution. Membership in a specific chamber of Congress likely influences what particular users will talk about, but the constraints are much lower for Twitter than for role-call voting. DW-NOMINATE and PAC-NOMINATE scores should not be compared across chambers because the House and Senate vote and debate different legislation. However, on Twitter, users can choose topics freely and can discuss a more diverse set of political issues than those on which they can vote. A couple methods of collapsing data could be useful for facilitating cross-chamber comparisons: joint tags and joint topics. In the first, joint tags, researchers could limit analysis to just those tags that are used by members of both the House and Senate. In the second, joint topics, researchers could assign tags to general topics (e.g., health care, immigration) and then compare members of both chambers by topic.

Last, while we have used #Polar scores to estimate points in a two-party (liberal–conservative) system, both classification and feature selection algorithms easily generalize to multiparty settings with multiple classes. For political systems with more than one ideological dimension, our method would use political parties as a surrogate for the combination of dimensions. Parties still occupy a single point in a two-dimensional scale, and the party labels thus carry information about position along both axes.

Conclusion

#Polar scores leverage available information such as party labels, word frequency, and hashtags to create a readily accessible, straightforward algorithm for estimating polarity using political speech. Our approach provides several specific advantages over other ideology estimation approaches. First, #Polar-User scores can be calculated at any time and for various time periods. DW-NOMINATE scores are determined by a member's entire voting history and are calculated only once per session. As of May 7, 2014, for instance, DW-NOMINATE scores are available only through the 112th Congress and for 525 current members of Congress. #Polar-User scores can be calculated for various time periods. By setting different start and end points for posts to include in calculating the #Polar-User scores, we provide a tool for analyzing changes in a politician's rhetoric over time. Because DW-NOMINATE is cumulative, it does not easily allow for this kind of longitudinal comparison.

Second, #Polar-User scores can readily differentiate between members who have nearly identical scores because the underlying features—the tags used and their frequency—are transparent. When using #Polar-User scores, researchers can easily identify the specific hashtags that explain differences between user's scores, allowing them to differentiate between similar users and to check the face validity of scores.

Third, #Polar-User scores provide options (e.g., limiting tags used in scoring, aggregating topics) for comparing across chambers. Therefore, we may use #Polar-User scores to evaluate claims such as "The House is more polarized than the Senate" and "The House has become increasingly more polarized." Like Wordscores, #Polar-Hashtag scores are language- and institution-independent and can be used to estimate the ideology of any speaker based on the text produced on Twitter. #Polar-Hashtags require less text and fewer computational resources than Wordscores, however, and are the most readily available tool for estimating ideology using publicly available text.

Notes

1. We provide both the data (Culotta, Hemphill, & Heston, 2015) and Python code (Purpletag, 2016) for developing #Polar scores.
2. We also used tested logistic regression and support vector machine classifiers, but neither resulted in higher accuracy, so we omit them from further discussion.
3. "NoShowNoPay" refers to a bill that would cut Congressional pay for missing votes, and "CutTheFleet" refers to a bill co-sponsored by Rep. Barrow and Rep. Richard Hanna (R-NY) that reduces the number of vehicles the federal government owns.

Funding

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1525662 and IIS-1526674.

Notes on contributors

Libby Hemphill is an associate professor of Information and Communication Studies at the Illinois Institute of Technology. Her research interests are social media and civic engagement, political communication in social media, and computational approaches to social media analysis.

Aron Culotta is an assistant professor of Computer Science at the Illinois Institute of Technology. He investigates socially beneficial applications of natural language processing, machine learning, and text mining algorithms.

Matthew Heston is a PhD student in Technology and Social Behavior at Northwestern University. He is interested in multimethod research in human–computer interaction and computer-mediated communication.

References

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408–446. doi:10.1086/590649

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bode, L., Hanna, A., Yang, J., & Shah, D. V. (2015). Candidate networks, citizen clusters, and political expression strategic hashtag use in the 2010 midterms. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 149–165. doi:10.1177/0002716214563923

Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science*, *58*(2), 367–386. doi:10.1111/ajps.2014.58.issue-2

Christensen, C. (2013). Wave-Riding and Hashtag-Jumping. *Information, Communication & Society*, *16*(5), 646–666. doi:10.1080/1369118X.2013.783609

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 89–96). Menlo Park, CA: The AAAI Press.

Culotta, A., Hemphill, L., & Heston, M. (2015). Data used to develop #Polar scores. doi:10.6084/m9.figshare.3409927.v1

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

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

Entman, R., & Rojecki, A. (1993). Freezing out the public: Elite and media framing of the U.S. anti-nuclear movement. *Political Communication*, *10*(2), 155–173. doi:10.1080/10584609.1993.9962973

Gaffney, D. (2010). *#iranElection: Quantifying online activism.* Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Web Science Trust. Retrieved from: http://journal.webscience.org/295/2/websci10_submission_6.pdf

Gitlin, T. (1980). *The whole world is watching: Mass media in the making & unmaking of the new left*. Berkeley: University of California Press.

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, *61*(8), 1612–1621. doi:10.1002/asi.21344

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198–208. doi:10.1177/0165551506062337

Graham, T., Jackson, D., & Broersma, M. (2014). New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns. *New Media & Society*. doi:10.1177/1461444814546728

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second). New York, NY: Springer.

Hemphill, L., Otterbacher, J., & Shapiro, M. (2013). What's Congress doing on Twitter? *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13 (pp. 877–886). doi:10.11452441776.2441876

Himelboim, I., McCreery, S., & Smith, M. A. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 40–60. doi:10.1111/jcc4.12001

Jolliffe, I. (2002). *Principal component analysis*. New York, NY: Springer-Verlag.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239–259. doi:10.1111/jcom.12087

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747. doi:10.1177/1461444811422894

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *The American Political Science Review*, *97*(2), 311–331. doi:10.1017/S0003055403000698

Lewis, J. B., & Poole, K. T. (2004). Measuring bias and uncertainty in ideal point estimates via the parametric bootstrap. *Political Analysis*, *12*(2), 105–127. doi:10.1093/pan/mph015

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, *16*(4), 356–371. doi:10.1093/pan/mpn004

Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, *36*(1), 123–155. doi:10.1111/j.1939-9162.2010.00006.x

McAuliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 121–128). Retrieved from http://papers.nips.cc/paper/3328-supervised-topic-models.pdf

McCarty, N. M., & Poole, K. T. (1998). An empirical spatial model of congressional campaigns. *Political Analysis*, *7*(1), 1–30. doi:10.1093/pan/7.1.1

Meraz, S., & Papacharissi, Z. (2013). Networked gatekeeping and networked framing on #Egypt. *The International Journal of Press/Politics*, *18*(2), 138–166. doi:10.1177/1940161212474472

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2009). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403. doi:10.1093/pan/mpn018

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication*, *10*, 55–75. doi:10.1080/10584609.1993.9962963

Poole, K. T., & Rosenthal, H. (1984). The polarization of American politics. *The Journal of Politics*, *46*(4), 1061–1079. doi:10.2307/2131242

Poole, K. T., & Rosenthal, H. (2007). *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Purpletag [source code]. (2016). Retrieved from https://github.com/casmlab/purpletag/releases/tag/v1.0. doi:10.5281/zenodo.53888

Rader, E., & Wash, R. (2008). Influences on tag choices in del.icio.us. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (pp. 239–248). San Diego, CA: ACM.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. *WWW '11 Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 113–114). doi:10.1145/1963192.1963250

Scheufele, D. A., & Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, *57*(1), 9–20. doi:10.1111/j.1460-2466.2006.00326.x

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722. doi:10.1111/ajps.2008.52.issue-3

Small, T. A. (2011). What the hashtag? *Information, Communication & Society*, *14*(6), 872–895. doi:10.1080/1369118X.2011.554572

Weber, I. (2013). "Political polarization of web search queries and Hashtags" by Ingmar Weber, with Martin Vesely as coordinator. *SIGWEB Newsletter* (Summer), 1–10. doi:10.1145/2501187.2501191

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, *5*(1), 33–48. doi:10.1080/19331680802149608