Methodological Supplement

"Trying to understand Jeff Flake? We analyzed his Twitter feed — and were surprised."

Washington Post Monkey Cage, October 5, 2018

By: Gregory Eady, Jan Zilinsky, Jonathan Nagler, and Joshua Tucker

This document by Gregory Eady, October 4, 2018

Description. This document provides a brief description of a statistical method for measuring on social media (1) the ideology of social media *content* and (2) the ideology of *political actors* and *users*. It is an abbreviated version of a larger paper in a series on the topic.

Methodology

To ease understanding of our statistical notation and make the goal of our measurement strategy more concrete, we begin by introducing our data. We formally lay out our statistical model—with these data as our example—thereafter.

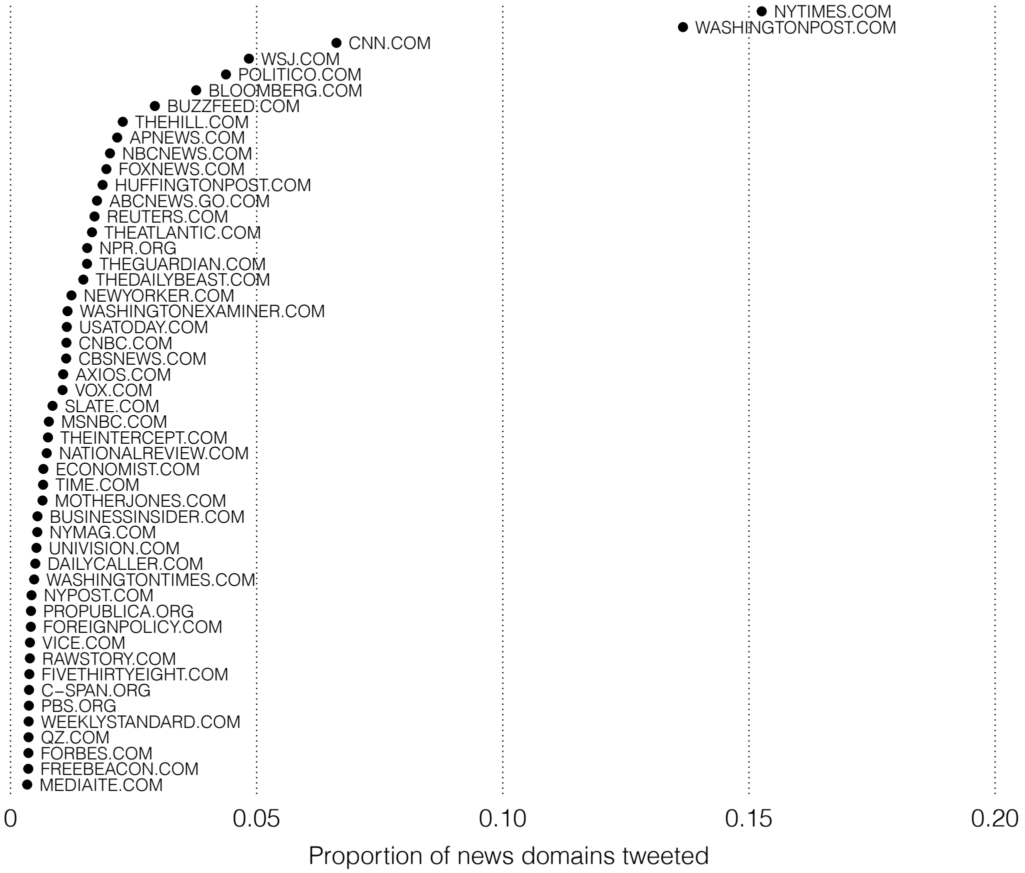
Data

The method that we introduce is platform-agnostic, and can be applied to any social media platform on which users and political actors alike share links to political news media. For our application, we use data from Twitter. Our reasons are threefold. First, the vast majority of Members of Congress in the US (99%) have Twitter accounts, many of which contain posts that link to media articles and commentary. These data allow us to validate our statistical model against well-known measures of the ideology of Members of Congress and, furthermore, to simultaneously test the extent to which the news that politicians share signals their political ideology. Second, and more pragmatically, Twitter's application programming interface (API) provides straightforward access to these data (3,200 of the most recent tweets per timeline), and those from all ordinary Twitter users who have not marked their Twitter profiles as private. Third, the shortened URLs included in Tweets (e.g. nyti.ms/2BUTshD) that are sent by users through Twitter's user interface are automatically unshortened (e.g. www.nytimes.com/2018/02/13/upshot/fake-news/...), reducing the number of URLs that require manual unshortening. We unshorten all remaining shortened links.[1]

As data, we begin by compiling a list of the Twitter accounts of all US Members of Congress; members of the executive and cabinet; and accounts associated with the Democratic and Republican parties. For each account, we collect the most recent 3,200 tweets from each political

[1] The unshortening of URLs is a time-consuming process even with multi-threaded code, a drawback to using link data. Social media providers, however, are beginning to provide these links to researchers in an unshortened format. Leon Yin from NYU's Social Media and Political Participation (SMaPP) lab also provides a Python package, `urlExpander`, for researchers to use for multi-threaded link shortening.

Figure 1: Fifty most tweeted national political commentary and news media domains as a proportion of all commentary/news domains



This graph shows the domains of the fifty most frequently tweeted domains as a proportion of all domains identified associated with national political news media.

actor's timeline. For each tweet, we then extract all URLs and unshorten all shortened links.[2] For analysis, we exclude URLs from quote tweets because URLs in these tweets are often used as a point of criticism or mockery, rather than support.

With the URLs for the political actors in hand, we categorize the links shared by these users by their domain. To do so, we define our population of political media websites as all sites that provide news or commentary about national politics, including sites from television media (e.g. cnn.com; foxnews.com), traditional print journalism (e.g. nytimes.com; wsj.com), and commentary (e.g. nationalreview.com; newrepublic.com). News and commentary sites include, furthermore, those that are generally considered highly partisan (e.g. dailycaller.com; thinkprogress.com), and thus sites expected to be on the extremes of the ideological spectrum—a useful check on the face validity of the statistical measurement approach that we introduce below.

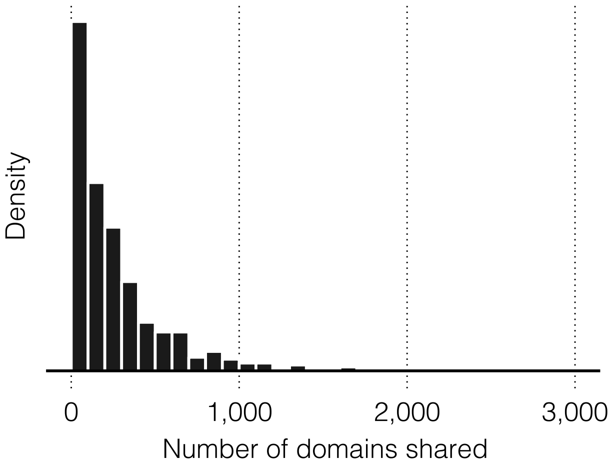
To show graphically the distribution of these data aggregated across all users, we present in Figure 2 the fifty most tweeted of these national political commentary and news domains as

[2] The vast majority of shortened links can be unshortened to retrieve the link that they point to.

Table 1: Example of social media user-media domain count matrix

	thinkprogress.org	nytimes.com	wsj.com	foxnews.com	breitbart.com	⋯
Ted Cruz (R)	0	37	50	80	34	⋯
Donald Trump (R)	0	2	5	29	6	⋯
Susan Collins (R)	0	5	1	1	0	⋯
Dianne Feinstein (D)	0	65	8	0	0	⋯
Cory Booker (D)	2	110	2	1	0	⋯
Kamala Harris (D)	8	165	10	0	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	

Figure 2: Number of news media domains shared by users on Twitter



This graph shows the density of the number of news media domains shared per user.

a proportion of all such domains. As the figure shows, the most frequently shared links are to well-known traditionally print news (e.g. New York Times, Washington Post, Wall Street Journal), and the major television media organizations (e.g. CNN, FOX News, NBC, ABC). By contrast, only a few periodicals dedicated to political commentary (e.g. The New Yorker, Foreign Policy, The Weekly Standard) find themselves among the most frequently shared domains.

To aggregate these data, for all users $i = 1, \ldots, N$ and media domains $m = 1, \ldots, M$, we generate an $N \times M$ user-domain count matrix. In other words, each cell in the matrix simply represents the number of times that a given user i has tweeted a story from media organization m. To see this by example, Table 1 presents a sub-matrix of our data for six well-known Republicans and Democrats (rows) and five media domains (columns). As these data appear to suggest, Republican politicians appear more likely to tweet links to media stories right of center (foxnews.com; breitbart.com) than they are those left of center (thinkprogress.com; nytimes.com); Democrats, more likely to tweet links to media stories to the left than they are those to the right.

Statistical model

To estimate the ideology of (1) the media organizations from which social media users and political actors share political news, and (2) the ideology of those users and actors themselves, we develop a Bayesian item-response theory (IRT) model for the news media URLs that are shared on social media. Consistent with the data described above, let y_{img} denote the count of the media domain $m = 1, \dots, M$ shared by a user or political actor $i = 1, \dots, N$ who is affiliated with the group $g \in$ {D, R, U} (Democrat, Republican, Unaffiliated). Concretely, y_{img}, in other words, denotes the value of a single cell in Table 1, where the columns represent the media organizations m, and the rows represent the users/actors i affiliated with group g.

To model these data in a way consistent with our spatial assumption of news sharing, we introduce two latent variables as our primary quantities of interest. The first, ϑ_{ig}, denotes the ideology of user i (affiliated with group g); the second, ζ_m, the ideology of a given media organization m. As shorthand, we refer to both of these parameters as *media scores*, making clear by context whether we are referring to individual users or media organization domains. We then model the data, y_{img}, as arising from a negative binomial (count) distribution as follows:

$$y_{img} \sim \text{NegBin}(\pi_{img}, \omega_i) \tag{1}$$

$$\pi_{img} = \exp(\alpha_i + \gamma_m - ||\vartheta_i - \zeta_m||^2 + \mathbf{x}_i'\boldsymbol{\beta}), \tag{2}$$

where α_i denotes a user-specific intercept, γ_m denotes a domain-specific intercept, and $\mathbf{x}_i$ and $\boldsymbol{\beta}$ denote vectors of user-specific covariates and parameters respectively. Lastly, ω_i denotes a user-specific dispersion parameter. Substantively, α_i represents the relative extent to which a given user shares news in general, and γ_m the relative extent to which a given media domain is shared by users within the sample. The term containing our quantities of interest, $-||\vartheta_i - \zeta_m||^2$, captures the notion that the larger the distance between the ideology of a given user (ϑ_i) and a given media organization (ζ_m), the less likely that user is to share links to that organization's news content. Lastly, although dispersion parameters are rarely given substantive interpretation, ω_i represents an important quantity: the extent to which a given user's sharing behavior is predictable.[3] This parameter, in other words, captures the extent to which a given user shares media consistent with his or her ideology, or shares information from sources more broadly across the ideological spectrum.

For estimation in a Bayesian framework,[4] we place priors on each group of parameters, and set constraints as necessary to identify the model. In particular, the user- and domain-specific intercepts are each given common distributions, $\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$ and $\gamma_i \sim \text{Normal}(0, \sigma_\gamma)$ respectively. We use group-level information about users, $g \in \{D, R, U\}$ (Democrat, Republican, Unaffiliated), by placing separate common prior distributions on the parameters denoting the ideology, ϑ_{ig}, of politicians who are members of the Democratic and Republican parties, $\vartheta_{iD} \sim \text{Normal}(\mu_\vartheta^{(D)}, \sigma_\vartheta^{(D)})$ and $\vartheta_{iR} \sim \text{Normal}(\mu_\vartheta^{(R)}, \sigma_\vartheta^{(R)})$ respectively.[5] For identification, the prior distribution for the ideology of ordinary users is set, as is common in typical IRT models, as $\vartheta_{iU} \sim \text{Normal}(0, 1)$.[6] The parameters denoting the ideology of media organizations, are given weakly informative prior distribution, $\zeta_m \sim \text{Normal}(0, 5)$. Finally, the dispersion parameters, ω_i,

[3] Dispersion parameters and (un)predictability in statistical and machine learning models are increasingly being given important substantive interpretations for theory testing. One important early case is that by Lauderdale (2010). Recent work in this vein includes that by Gentzkow, Shapiro and Taddy (2017), Peterson and Spirling (2018), Bertrand and Kamenica (2018), and Eady and Loewen (2018).

[4] All models in this article are fit using the Bayesian inference engine Stan (Carpenter et al., 2017), which provides flexibility in fitting these models to link data from other social media platforms and the inclusion of user-level covariates.

[5] Uniform prior distributions are placed on the hyperparameters $\mu_\vartheta^{(\cdot)}$ and $\sigma_\vartheta^{(\cdot)}$.

[6] Setting the prior $\vartheta_{iU} \sim \text{Normal}(0, 1)$ resolves the problem of additive aliasing caused by the fact that the likelihood is invariant to adding a constant to the parameters ϑ_{ig} and ζ_m.

are given a common distribution $\omega_i \sim \text{InvGamma}(\omega_a, \omega_b)$.[7]

To identify the model, we need to address the problem of reflection invariance (Bafumi et al., 2005), which refers to the fact that the likelihood is invariant to multiplication of the parameters ϑ_{ig} and ζ_m by -1. We need, in other words, to fix the direction of the scale such that higher values of ϑ_{ig} and ζ_m indicate either liberal or conservative. Bafumi et al. (2005) propose a number of ways to achieve identification. Here, we constrain the signs of five pairs of the parameters of $\vartheta_i g$, such that those of well-known liberal Members of Congress are less than those of well-known conservative Members of Congress. In other words, we identify the models such that higher values of ϑ_{ig} (and ζ_m) indicates conservatism.[8]

To validate our statistical measurement approach, we investigate the extent to which Members of Congress reveal their ideology through the sharing of political news on social media. Our measurement strategy allows us investigate this because one of the key benefits of using data from social media news sharing is that it results in a behavioral measure of ideology among both ordinary citizens *and* politicians. Thus, whereas past research has relied on the following and endorsement of politicians by ordinary social media users, i.e. a perceptual measure of ideology, the data concerning the content shared by politicians themselves permits measurement from the behavior of politicians themselves.

As validation, we compare estimates of ideology from social media news sharing to those obtained using legislators' roll-call voting record. To do so, we compare media scores for Members of Congress to estimates of their ideology from the well-known DW-NOMINATE scoring procedure (Poole and Rosenthal, 1985; Boche et al., 2018). The results show that the correlation between media score estimates and those from DW-NOMINATE is extremely high, $\rho = 0.96$, when examining the data in aggregate. The correlation is also very high in distinguishing the ideology of legislators within parties ($\rho^{Senate(D)} = 0.68$, $\rho^{Senate(R)} = 0.62$, $\rho^{House(D)} = 0.58$, $\rho^{House(R)} = 0.55$).

References

Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.

Bertrand, Marianne and Emir Kamenica. 2018. "Coming Apart? Cultural Distances in the United States Over Time." Working paper, June.

Boche, Adam, Jeffrey B. Lewis, Aaron Rudkin and Luke Sonnet. 2018. "The New Voteview.com: Preserving and Continuing Keith Poole's Infrastructure for Scholars, Students and Observers of Congress." *Public Choice* 176(1):17–32.

Carpenter, Bob, Andrew Gelman, Matt D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1):1–32.

Eady, Gregory and Peter Loewen. 2018. "Where in the World is Donald Trump? Measuring Citizen Uncertainty in Candidate Ideology." Working paper, July 23.

[7] The hyperparameters ω_a and ω_b are given improper uniform priors Uniform$(0, \infty)$.

[8] Although technically fixing the order of a single pair of parameters ϑ_{ig} is sufficient for identification, model fitting can be difficult without further constraints to aid estimation. For estimation, the starting parameters for Democratic Members of Congress are set to -2 and those for Republicans, $+2$.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2017. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." Working paper, May.

Lauderdale, Benjamin E. 2010. "Unpredictable Voters in Ideal Point Estimation." *Political Analysis* 18(2):151–171.

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1):120–128.

Poole, Keith and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.