A For Online Publication

A.1 Ideological segregation on Twitter

Our main analysis investigated a variety of channels through which individuals read the news, but it was limited to a particular opt-in sample of individuals. In this supplementary section, we augment our analysis by examining the news consumption habits of a nearly complete set of users on one specific social information channel, Twitter, one of the largest online social networks, and arguable the largest designed primarily for information discovery and dissemination, as exemplified by their instructions to users to "simply find the accounts you find most compelling and follow the conversations."[13]

The Twitter and toolbar datasets differ on two additional substantively important dimensions. First, Internet Explorer and Twitter users are demographically quite different. For example, whereas Internet Explorer users are believed to be, on average, older than those in the general Internet population, Twitter users skew younger. In particular, 27% of 18–29 year-olds use Twitter, compared to 10% of those aged 50–64 (Pew Research, 2013). Second, because of differing levels of information in the two datasets, in the toolbar analysis we examine the articles that an individual *viewed*, whereas with Twitter we look at the articles that were merely *shared* with that individual, regardless of whether or not he or she read the story. Thus, given these differences, to the extent that our results extend to this setting, we can be further assured of the robustness of our findings.

To generate the Twitter dataset, we start with the nearly complete set of U.S.-located individuals who posted a tweet during the two-month period March–April, 2013.[14] We focus on accounts maintained and used by an individual (as opposed to corporate accounts), and so further restrict to those that receive content from ("follow") between 10 and 1,000 users on the network. This process yields approximately 7.5 million individuals. Finally, similar to our restriction in the toolbar analysis, we limit to active news consumers, who received (i.e., followed individuals who posted)

[13] Twitter positions itself as a fully-customizable information portal, this quote comes from www.twitter.com/about.

[14] Twitter offers the option of "protected accounts," which are not publicly accessible. These accounts are rare and are not part of our study.

at least 10 front-section news articles and at least 2 opinion pieces.[15] In total, 1.5 million users meet all of these restrictions.

We begin our analysis by estimating the distribution of user polarity. In this setting, user polarity is the typical polarity of the articles to which a user is exposed (i.e., articles that are posted by an account the user follows), where we recall that the polarity of an article is the conservative share of the outlet in which it was published. Since users on Twitter often receive news by following the accounts of major news outlets rather than accounts of actual individuals (Kwak et al. 2010), and since these news outlets typically post hundreds of articles per day, individuals in our sample are generally exposed to large numbers of news articles—4,008 on average during the two-month time frame we study. As a consequence, data sparsity is not a serious concern, which in turn significantly simplifies our estimation procedure. Specifically, for each Twitter user, we estimate polarity by simply averaging the polarities of the articles to which he or she is exposed.

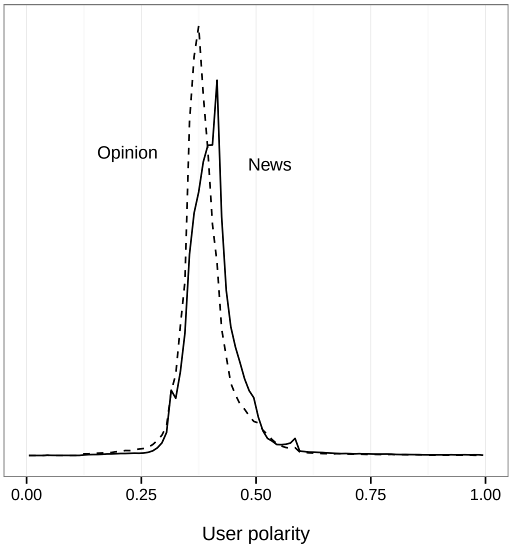
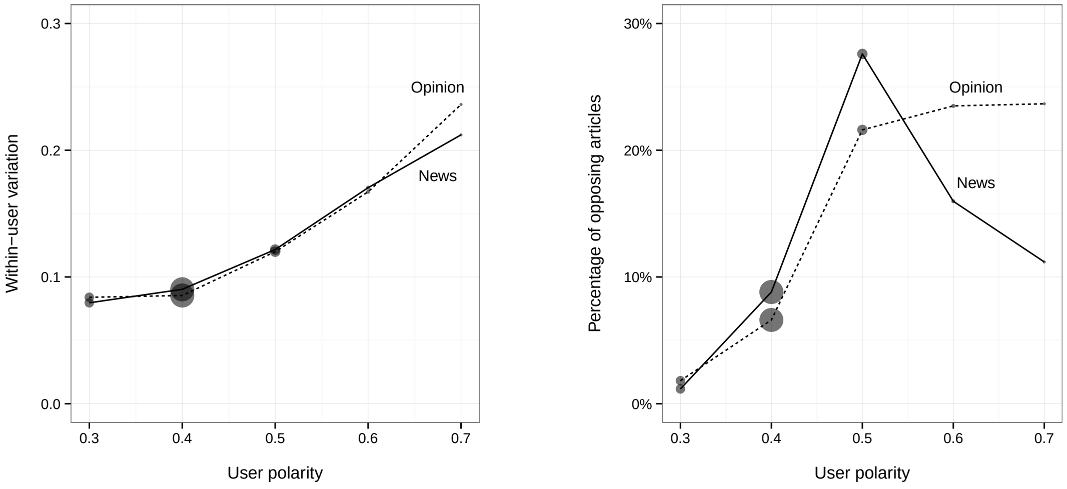


Figure A1: Distribution of individual-level polarity for Twitter users, where an individual's polarity score is the average conservative share of news outlets to which he or she is exposed, computed separately for descriptive news articles (solid line) and opinion pieces (dashed line).

[15] As with the toolbar analysis, articles were classified as front-section news and opinion according to the methods described in Section 1.



(a) Within-user variation.

(b) Opposing partisan views.

Figure A2: Within-user variation (a) and opposing partisan views (b) on Twitter, as a function of individual-level polarity. The sizes of the points indicate the relative number of individuals in each polarity bin, normalized separately for front-section news (solid line) and opinion (dashed line).

Figure A1 shows the resulting distribution of user polarity, where we separately plot the user polarity distribution computed for descriptive news articles (solid line) and opinion stories (dashed line). This plot illustrates two points. First, despite a slight leftward ideological skew relative to toolbar users, the bulk of Twitter users exhibit quite moderate news preferences. For example, 70% of Twitter users have polarity scores between 0.35 and 0.45, ranging from *The Huffington Post* to *CBS*. Second, segregation is correspondingly moderate, 0.10, and remarkably similar to our estimate from the toolbar data (0.11). Thus, despite the relative ease with which individuals may elect to follow politically extreme news publishers, and despite the worry that algorithmic recommendations of whom to follow could spur segregation, ideological segregation on Twitter looks very much like what we observe in direct navigation web browsing.

We investigate the exposure distribution further with two individual-level metrics: (1) within-user variation, defined as the standard deviation of the polarities of articles to which an individual is exposed; and (2) opposing partisan views, defined as the fraction of partisan articles from an individual's less preferred ideological perspective. The results are plotted in Figure A2, as a function of user polarity.

As indicated by Figure A2a, average within-user variation—averaged over all individuals in our sample of Twitter users—is 0.10, significantly higher than the 0.05 we observed for direct web browsing, but comparable to the 0.09 we found for articles obtained through aggregators (Figure 4a), consistent with the general view of Twitter as a custom aggregator. Further, as we saw before, within-user variation increases substantially as we move to the conservative end of the spectrum; that is, individuals who on average consume more conservative content also tend to consume content from a wider variety of ideological viewpoints.

We plot opposing partisan exposure in Figure A2b, restricting to individuals who are exposed to at least two partisan articles (as we required in the toolbar analysis). Average opposing partisan exposure is 11%, very close to the 10% we observe in the toolbar dataset—the vast majority of an individual's partisan views come from their preferred political side. However, a notable difference between the two datasets is that whereas in the toolbar data both left- and right-leaning individuals have little exposure to opposing views, on Twitter, right-leaning individuals have considerably more exposure to opposing views than left-leaning users. Though it is not entirely clear what is driving this effect, it is likely in part due to the overall leftward skew of Twitter, where it is thus harder for right-leaning individuals to isolate themselves from the majority view.

A.2 News and opinion classifier

To train the news and opinion classifiers, we require datasets consisting of a representative set of articles known to be front-section news, and another known not to be (i.e., a sampling of articles from the categories we wish to filter out, hereafter referred to as "non-news"); we likewise require labeled examples of descriptive versus opinion articles. To generate these sets we make use of the fact that many popular publishers indicate an article's classification in its URL (web address). For example, a prototypical story on *USA Today* (in this case, about U.S. embassy security) has the address `http://www.usatoday.com/story/news/world/2013/08/01/us-embassies-sunday-security/2609863/`, where "news/world" in the URL indicates the article's category. Identifying these URL patterns for 21 news websites, we are able to produce 70,406 examples of front-section news and opinion, and 73,535 examples of non-news. We use the same approach (looking for URLs with the word “opinion”) to generate a separate training dataset to distinguish between opinion pieces and descriptive news articles.

Given these training datasets, we next build a natural language model. We first compute the 1,000 most frequently occurring words in our corpus of articles, excluding so-called stop words, such as “and”, “the”, and “of”.[16] We augment this list with a set of 39 first and third person pronouns (Pennebaker, Francis, and Booth 2001; Pennebaker et al. 2007), since opinion pieces—unlike descriptive articles—are often written in the first person, and including such pronouns has been shown to improve performance (Glover et al. 2001). Each article is subsequently represented as a 1,039-dimensional vector, where the i-th component indicates the number of times the i-th word in our list appears in the article, normalized by the total number of words in the article. Using fractional scores rather than raw frequencies is a standard approach in natural language classification tasks for dealing with differences in article length (Manning and Schütze 1999). To retain the predictive power of the pronouns, quotations are removed from the articles before representing them as vectors of relative word frequencies.

Having defined the predictors (i.e., the relative frequencies of various popular words), and having generated a set of labeled articles, we now use logistic regression to build the classifiers. Given the scale of the data, we fit the models with the L-BFGS algorithm (Liu and Nocedal 1989), as implemented in the open-source machine learning package Vowpal Wabbit. Applying the fitted model to the entire collection of 4,127,140 articles in our corpus, we obtained 2,226,170 stories (46%) classified as front-section news or opinion, and of these 11% are classified as opinion. Note that as mentioned in the text, we use the classifier even for outlets that indicate the article category in the URL, which guards against differing editorial policies biasing the results.

The accuracy of our classifiers is quite high. When tested on a 10% hold-out sample of articles whose categories can be inferred from their URLs, the front-section news classifier obtains 96% accuracy, with 95% precision and 97% recall (where the positive class is news and the negative class is non-news). We also achieve good performance on a hand-labeled set of 100 randomly selected articles

[16] me, ive, myself, weve, we, wed, i, ill, were, well, mine, us, ourselves, lets, im, ours, our, my, id,shes, shed, himself, theyll, her, hes, theyve, them, hed, their, his, she, they, theyd, hers, shell, themselves, herself, him, and he.

from the full corpus: 81% accuracy, 84% precision, and 79% recall. Accuracy for the opinion classifier is high as well: 92% on a hold-out set of URL-labeled articles, with 96% precision and 76% recall (where the positive class is opinion and the negative class is descriptive news). On a randomly selected hand-labeled subset of 100 news articles accuracy is 88%, with 80% precision and 57% recall. Table 1 lists words with the highest positive and negative weights for both classifiers—the words accord with common intuition. While the overall performance of our classifiers is quite good, it is by no means perfect. However, we note that in many cases there is genuine ambiguity, even among human judges, as to what constitutes, for example, descriptive news versus opinion.

A.3 Measuring the political slant of publishers

Approaches for measuring the political slant of news outlets broadly fall into one of two categories: content-based and audience-based. Content-based approaches compare the entire body of published textual content from a source (rather than individual articles) to sources with known political slants. For example, Groseclose and Milyo (2005) use the co-citation matrix of newspapers and members of Congress referencing political think tanks. Similarly, Gentzkow and Shapiro (2010) use congressional speeches to identify words and phrases associated with a stance on a particular issue, and then tabulate the frequencies of such phrases in newspapers. Audience-based approaches, on the other hand, use the political preferences of a publication's readership base to measure political slant (Tewksbury 2005; Gentzkow and Shapiro 2011). Empirical evidence suggests that audience and content-based measures of slant are closely related (Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2006). In particular, Iyengar and Hahn (2009) show that individuals select media outlets based on the match between the outlet's and their own political positions, and moreover, it has been shown that outlets tailor their coverage to match the preferences of their base (DellaVigna and Kaplan 2007; Baum and Groeling 2008; Gentzkow and Shapiro 2010).

Here we use an audience-based measure of news outlet slant. Specifically, we estimate the fraction of each news outlet's readership that voted for the Republican candidate in the most recent presidential election (among those who voted for one of the two major-party candidates), which we call the outlet's *conservative share.*

Thus, liberal outlets have conservative shares less than about 50%, and conservative outlets have conservative shares greater than about 50%, in line with the usual left-to-right ideology spectrum. To estimate the political composition of a news outlet's readership, we make use of geographical information in our dataset. Specifically, each webpage view includes the county in which the user resides, as inferred by his or her IP address. With this information, we then measure how the popularity of a news outlet varies across counties as a function of the counties' political compositions, which in turn yields the estimate we desire.

More formally, as a first approximation we start by assuming that the probability any user u_i views a particular news site s is solely a function of his or her party affiliation. Namely, for a fixed news site s, we assume Democrats view the site with probability p_d and Republicans view the site with probability p_r.[17] Reparameterizing so that $\beta_0 = p_d$ and $\beta_1 = p_r - p_d$, we have

$$\mathbb{P}(u_i \text{ views } s) = \beta_0 + \beta_1 \delta_r(u_i) \tag{5}$$

where $\delta_r(u_i)$ indicates whether user u_i is a Republican. Though our ultimate goal is to estimate β_0 and β_1, we cannot observe an individual's party affiliation. To circumvent this problem, for each county C_k we average (5) over all users in the county, yielding

$$\frac{1}{N_k} \sum_{u_i \in C_k} \mathbb{P}(u_i \text{ views } s) = \beta_0 + \beta_1 \frac{1}{N_k} \sum_{u_i \in C_k} \delta_r(u_i) \tag{6}$$

where N_k is the number of individuals in our sample who reside in county C_k.

While the left-hand side of (6) is observable—or at least is well approximated by the fraction of users in our sample that visit the news site—we cannot directly measure the fraction of Republicans in our sample (i.e., the sum on the right-hand side of (6) is not directly observable). To address this issue, we make a further assumption that our sample of users is representative of the county's voting population, a population for which we can estimate party composition via the 2012 election

[17] As discussed later, by "Democrats" we in fact mean those who voted for the Democratic candidate in the last presidential election, and similarly for "Republicans."

returns. We thus have the following model:

$$P_k = \beta_0 + \beta_1 R_k \tag{7}$$

where P_k is the fraction of toolbar users in county C_k that visit the particular news outlet s, and R_k is the fraction of voters in county C_k that supported the Republican candidate, Mitt Romney, in the 2012 U.S. presidential election. To estimate the parameters β_0 and β_1 in (7), we fit a weighted least squares regression over the 2,654 counties for which we have at least one toolbar user in our sample, weighting each observation by N_k (i.e., the number of people in our dataset in county C_k).

Clearly, (7) is only an approximation of actual behavior, with our specification ruling out the possibility that a generally liberal outlet is disproportionately popular in conservative counties. In particular, our model ignores the impact of local news coverage, with individuals living in the outlet's county of publication visiting the site regardless of its political slant. Addressing this local effect, we modify our generative model to include an additional term. Namely, outside a news outlet's local geographic region, we continue to assume that Democrats visit the site with probability p_d, and Republican's visit the site with probability p_r, and we use (7)—fit on all non-local counties—to estimate p_r and p_d. Inside the local region we assume individuals visit the site with probability p_ℓ, irrespective of their political affiliation, and we estimate p_ℓ to be the empirically observed fraction of local toolbar users who visited the news outlet.

Finally, we approximate the conservative share $p(s)$ of a news outlet s as the estimated fraction of Republicans that visit the site normalized by the total number of Democratic and Republican visitors. Specifically,

$$p(s) = \left[N_\ell r_\ell p_\ell + p_r \sum_{k\,:\,C_k \text{ non-local}} N_k r_k\right] \Bigg/ \left[N_\ell p_\ell + \sum_{k\,:\,C_k \text{ non-local}} N_k(r_k p_r + (1 - r_k)p_d)\right]$$

where N_k is the number of people in our dataset in county C_k, $p_d = \beta_0$, $p_r = \beta_0 + \beta_1$, r_k is the two-party Romney vote share in county C_k (i.e., the number of Romney supporters divided by the total number of Romney and Obama supporters, excluding third party candidates), and parameters subscripted with ℓ indicate values for the outlet's local county of publication. This entire process is repeated for each

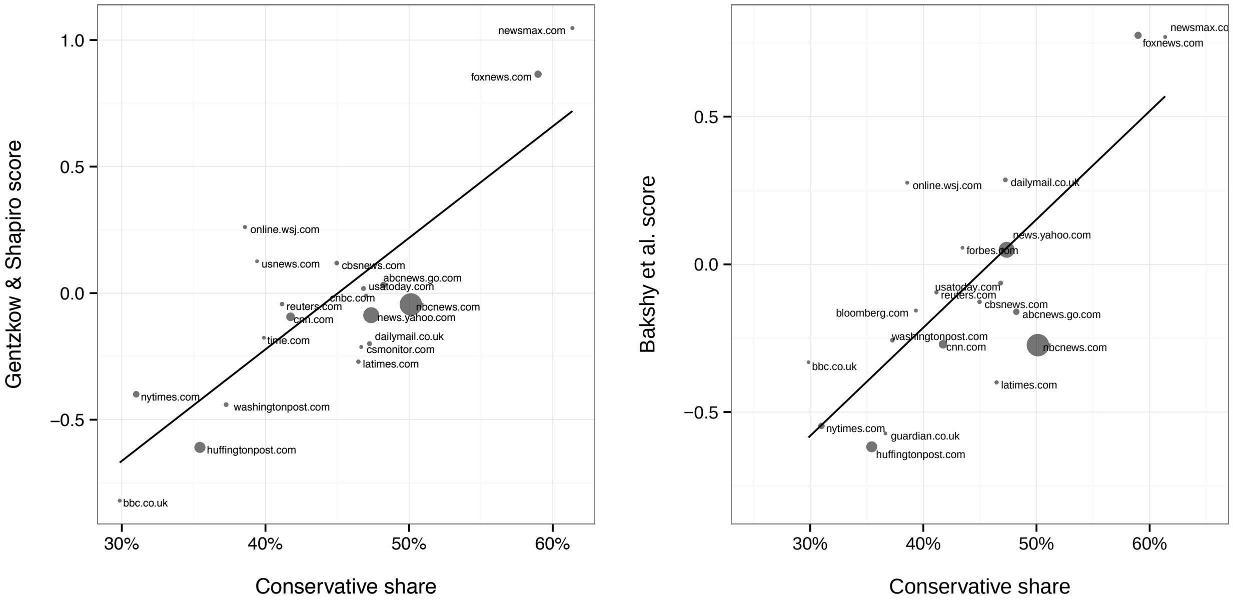


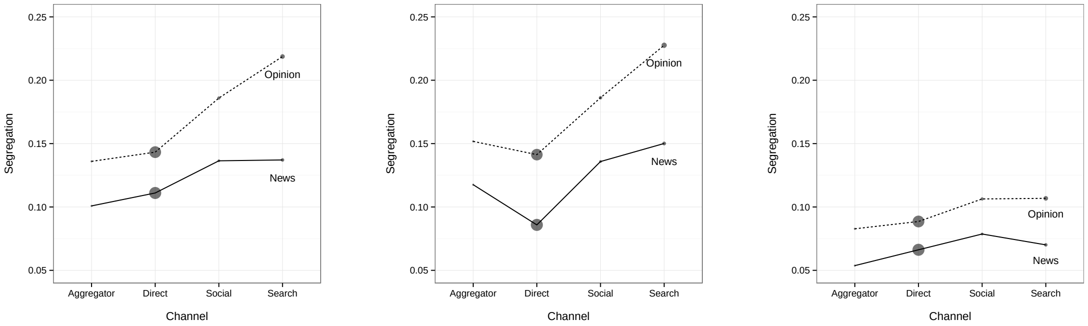
Figure A3: A comparison of our estimate of conservative share of an outlet's audience to two alternate measure of ideological slant as estimated by Gentzkow and Shapiro (2011) on left, and Bakshy et al. (2015) on right, where point sizes are proportional to popularity. Among these 20 publications, the correlation of our estimates with Gentzkow and Shapiro is 0.82, and with Bakshy et al is 0.77.

of the 100 news outlets in our dataset.

As shown in Figure 1 in the main text, our outlet scores are highly correlated with those reported by Pew. Figure A3 similarly shows that our scores are inline with those estimated by Gentzkow and Shapiro (2011), and also with scores by Bakshy, Messing, and Adamic (2015).

A.4 Sensitivity analysis

We carried out three robustness checks to confirm our main findings regarding segregation by channel (Figure 3). First, instead of a hierarchical Bayesian model, we estimated segregation by simple averaging. Specifically, for each of the eight consumption categories (e.g., opinion × direct), we: (1) subset the data to include only individuals who read at least one article in that category; (2) estimated the polarity for each such user by averaging the outlet scores for each article they read in that category; and (3) computed the sample standard deviation of the user polarity dis-



(a) Model-free. (b) Broader sample. (c) Segregation via median.

Figure A4: Robustness checks for segregation estimates: (a) shows results based on a model-free estimate of user polarity (i.e., simple averaging); (b) shows results when our sample is extended to include all users who viewed at least one news article; and (c) plots segregation defined as the median absolute distance between users (as opposed to the root mean squared distance). In all three cases, opinion has higher segregation than descriptive news, as do social media and web search, consistent with our main findings.

tribution. The results, shown in Figure A4a, are qualitatively similar to our main findings. Opinion has higher segregation than descriptive news, as do social media and web search.

Second, we apply this model-free approach to a broader set of users, namely those who read at least 1 front-section news article, for a total of 573,809 users. The results are shown in Figure A4b. We again see patterns qualitatively similar to our main findings.

Finally, as a third robustness check, we defined and estimated segregation as the median absolute distance between users (as opposed to the root mean squared distance). To do so, we started with the model-free estimates of user-level polarity, sampled pairs of random users, computed the absolute value of the distance between them, and then computed the median of this distribution of distances. The results are shown in Figure A4c. As expected, the magnitude of the segregation estimates are smaller. Nevertheless, the directional results are consistent with our primary findings.

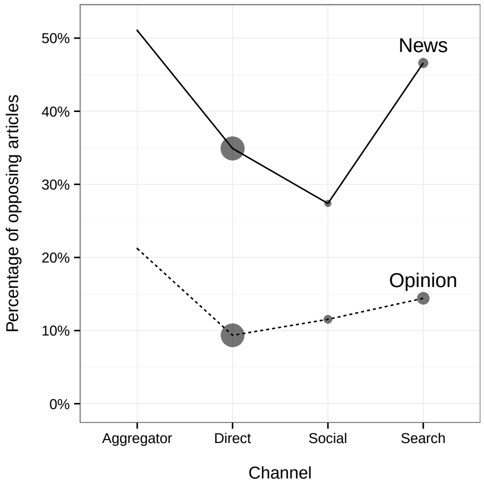


Figure A5: An alternative measure of "opposing exposure," defined as the proportion of articles read by partisans that are moderate or come from the opposing side.

A.5 An alternative measure of ideological isolation

Here we consider a measure of ideological isolation that quantifies the extent to which partisans are exposed to any cross-cutting sources, including exposure to moderate news outlets. Specifically, for individuals who predominantly read left-leaning articles, we measure the proportion of articles they view that are moderate or right-leaning. And for individuals who predominantly read right-leaning articles, we measure the proportion of their articles that are moderate or left-leaning. The results are shown in Figure A5. As must necessarily be the case, opposing exposure is larger under this measure than under the stricter one we use in the main text. Interesting, however, for nearly all channels (the lone exception being aggregators for descriptive news), partisans get the vast majority of their news from either right-leaning or left-leaning sources, with relatively little exposure to even ideologically moderate sources.

A.6 Additional tables and figures

Table 5: Conservative shares for the top 100 news outlets, ranked by share.

	Domain	Publication Name	Conservative Share
1	timesofindia.indiatimes.com	Times of India	0.04
2	economist.com	The Economist	0.12
3	northjersey.com	North Jersey.com	0.14
4	ocregister.com	Orange Country Register	0.15
5	mercurynews.com	San Jose Mercury News	0.17
6	nj.com	NewJersey.com†	0.17
7	sfgate.com	San Francisco Chronicle	0.19
8	baltimoresun.com	Baltimore Sun	0.19
9	courant.com	Hartford Courant	0.22
10	jpost.com	Jerusalem Post (EN-Israel)	0.25
11	prnewswire.com	PR Newswire	0.27
12	sun-sentinel.com	South Florida Sun Sentinal	0.27
13	nationalpost.com	National Post (CA)	0.28
14	thestar.com	Tornoto Star	0.28
15	bbc.co.uk	BBC (UK)	0.30
16	wickedlocal.com	Wicked Local (Boston)	0.30
17	nytimes.com	New York Times	0.31
18	independent.co.uk	The Independent	0.32
19	philly.com	Philadelphia Herald	0.32
20	hollywoodreporter.com	Hollywood Reporter	0.33
21	miamiherald.com	Miami Herald	0.35
22	huffingtonpost.com	Huffington Post	0.35
23	guardian.co.uk	The Guardian	0.37
24	washingtonpost.com	Washington Post	0.37
25	online.wsj.com	Wall Street Journal	0.39
26	news.com.au	News.com (AU)	0.39
27	dailykos.com	Daily Kos	0.39
28	bloomberg.com	Bloomberg	0.39
29	dailyfinance.com	Daily Finance	0.39
30	syracuse.com	Syracuse Gazette	0.39
31	usnews.com	US News and World Report	0.39
32	timesunion.com	Times Union (Albany)	0.40
33	time.com	Time Magazine	0.40
34	reuters.com	Reuters	0.41
35	telegraph.co.uk	Daily Telegraph (UK)	0.41
36	businessweek.com	Business Week	0.42
37	cnn.com	CNN	0.42
38	politico.com	Politico	0.42
39	theatlantic.com	The Atlantic	0.42
40	nationaljournal.com	National Journal	0.43
41	alternet.org	Alternet	0.43
42	ajc.com	Atlanta Journal Constitution	0.44
43	forbes.com	Forbes	0.44
44	seattletimes.com	Seattle Times	0.44
45	rawstory.com	The Raw Story	0.44
46	newsday.com	News Day	0.44
47	cbsnews.com	CBS	0.45
48	rt.com	Russia Today	0.45
49	theepochtimes.com	The Epoch Times	0.46
50	latimes.com	Los Angleles Times	0.47

	Domain	Publication Name	Conservative Share
51	csmonitor.com	Christian Science Monitor	0.47
52	realclearpolitics.com	Real Clear Politics	0.47
53	usatoday.com	USA Today	0.47
54	cnbc.com	CNBC	0.47
55	dailymail.co.uk	The Daily Mail (UK)	0.47
56	mirror.co.uk	Daily Mirror (UK)	0.47
57	news.yahoo.com	Yahoo! News	0.47
58	abcnews.go.com	ABC News	0.48
59	upi.com	United Press International	0.48
60	chicagotribune.com	Chicago Tribune	0.49
61	ap.org	Associated Press	0.50
62	nbcnews.com	NBC News	0.50
63	suntimes.com	Chicago Sun-Times	0.51
64	freep.com	Detriot Free Press	0.52
65	azcentral.com	Arizona Republics	0.53
66	tampabay.com	Tamba Bay Times	0.54
67	orlandosentinel.com	Orlando Sentinel	0.54
68	thehill.com	The Hill	0.57
69	nationalreview.com	The National Review	0.57
70	news.sky.com	SKY	0.58
71	detroitnews.com	Detroit News	0.59
72	express.co.uk	The Daily Express (UK)	0.59
73	weeklystandard.com	The Weekly Standard	0.59
74	foxnews.com	Fox News	0.59
75	washingtontimes.com	Washington Times	0.59
76	jsonline.com	Milwaukee Journal Sentinel	0.61
77	newsmax.com	Newsmax	0.61
78	factcheck.org	factcheck.org	0.62
79	reason.com	Reason Magazine	0.63
80	washingtonexaminer.com	Washington Examiner	0.63
81	ecanadanow.com	E Canada Now	0.63
82	americanthinker.com	American Thinker	0.65
83	twincities.com	St. Paul Pioneer Press	0.67
84	jacksonville.com	Florida Times Union	0.67
85	opposingviews.com	Opposing Views	0.67
86	chron.com	Houston Chronicle	0.67
87	startribune.com	Minneapolis Star Tribune	0.68
88	breitbart.com	Breitbart	0.70
89	star-telegram.com	Ft. Worth Star-Telegram	0.74
90	stltoday.com	St. Louis Post-Dispatch	0.75
91	mysanantonio.com	San Antonio Express News	0.77
92	denverpost.com	Denver Post	0.80
93	triblive.com	Pittsburg Tribune-Review	0.85
94	sltrib.com	Salt Lake Tribune	0.85
95	dallasnews.com	Dallas Morning News	0.86
96	kansascity.com	Kansas City Star	0.93
97	deseretnews.com	Deseret News (Salt Lake City)	0.94
98	topix.com	Topix	0.96
99	knoxnews.com	Knoxville News Sentinel	0.96
100	al.com	Huntsville News/Mobile Press Register/Birmingham News	1.00

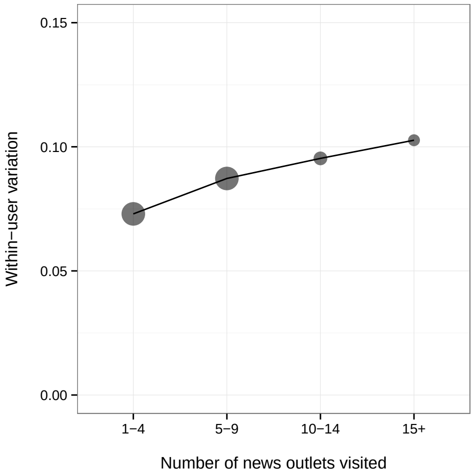


Figure A6: For a typical individual, within-user variation (i.e., standard deviation) of the conservative share of news outlets he or she visits, as a function of the number of outlets visited.

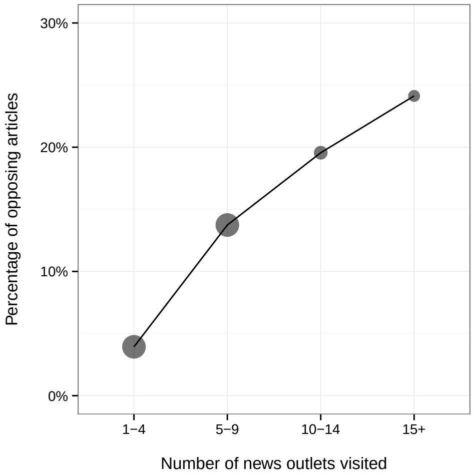


Figure A7: For a typical individual, fraction of partisan articles that are on the opposite side of the ideological spectrum from those he or she generally reads, as a function of the number of news outlets visited.

		News				Opinion			
		aggregator	direct	search	social	aggregator	direct	search	social
News	aggregator	0.0026							
	direct	0.0007	0.0058						
	search	0.0008	0.0033	0.0075					
	social	0.0010	0.0043	0.0042	0.0075				
Opinion	aggregator	0.0018	0.0013	0.0011	0.0010	0.0085			
	direct	0.0007	0.0064	0.0039	0.0050	0.0024	0.0089		
	search	0.0011	0.0038	0.0068	0.0048	0.0030	0.0057	0.0199	
	social	0.0008	0.0043	0.0048	0.0072	0.0030	0.0064	0.0089	0.0135

Table 6: Variance-covariance matrix for the model used to estimate ideological consumption by channel and subjectivity type, as described in Eqs. (3) and (4).

		News				Opinion			
		aggregator	direct	search	social	aggregator	direct	search	social
News	aggregator								
	direct	0.17							
	search	0.18	0.51						
	social	0.23	0.65	0.56					
Opinion	aggregator	0.39	0.18	0.14	0.12				
	direct	0.15	0.89	0.48	0.61	0.28			
	search	0.16	0.35	0.56	0.4	0.23	0.43		
	social	0.13	0.49	0.47	0.71	0.28	0.58	0.54	

Table 7: Correlation matrix for the model used to estimate ideological consumption by channel and subjectivity type, as described in Eqs. (3) and (4).