Ideological Segregation and the Effects of Social Media on News Consumption

Seth Flaxman
Carnegie Mellon University

Sharad Goel
Microsoft Research

Justin M. Rao
Microsoft Research

DRAFT

Abstract

Scholars and pundits have argued that the growth of social media and personalized web search could foster ideological segregation, a worry supported by laboratory experiments. We investigate the effects of such technological changes on news consumption by examining web browsing histories for 1.2 million U.S.-located users. We find that individuals indeed exhibit higher segregation when reading articles shared on social media or returned by search engines, a pattern driven by opinion pieces. However, opinion articles from social media and web search constitute only 2% of total news consumption. Rather, most people primarily read descriptive reporting, and do so by directly visiting a handful of mainstream, ideologically similar news outlets, resulting in little exposure to content from their less preferred side of the political spectrum, but also a moderate overall level of segregation. Consequently, while recent technological changes do appear to increase ideological segregation, the magnitude seems limited at this time.

Keywords: *media economics, information acquisition, media bias, online behavior, big data, confirmation bias*
JEL Codes: *D83, L86, L82*

1 Introduction

The Internet has dramatically reduced the cost to produce, distribute, and access diverse political information and perspectives. Online publishing, for example, circumvents much of the costly equipment required to produce physical newspapers and magazines. With the rise of social media sites such as Facebook and Twitter, individuals can now readily share their favorite stories with hundreds of their contacts, lowering the distribution costs of publishers (Bakshy et al., 2012; Goel et al., 2012b). And as web search engines and news aggregators become increasingly capable of generating personalized results, consumers can more easily find niche content tailored to their preferences (Agichtein et al., 2006; Das et al., 2007; Hannak et al., 2013; Speretta and Gauch, 2005).

These transformative effects of the Internet can be viewed as a boon for the democratization of ideas, as a wide spectrum of political positions now have the possibility of entering the public debate (Benkler, 2006). Several scholars and popular commentators, however, have raised concerns that instead of encouraging discussion, the proliferation of niche political perspectives and personalized recommendations promote so-called *echo chambers* or *filter bubbles* (Pariser, 2011; Sunstein, 2001, 2009), in which individuals are only exposed to like-minded others, leading to ideological segregation. Such segregation is a concern as it has long been argued that functioning democracies depend critically on voters who are exposed to and understand a variety of political views (Downs, 1957). Further, theoretical models have shown that segregation can lead to "electoral mistakes," especially if people fail to realize they are in an echo chamber (Bernhardt et al., 2008).

The worry that recent technological changes spur ideological fragmentation is not ungrounded speculation, rather it is supported by a number of findings in economics, psychology and sociology. In controlled experiments, people overwhelmingly opt to consume information that accords with their previously held views (Lord et al., 1984, 1979; Nickerson, 1998), and choose to read news articles from outlets that share their political opinions (Garrett, 2009; Iyengar and Hahn, 2009; Munson and Resnick, 2010). Survey evidence of blog readers (Lawrence et al., 2010) and cross-blog citations (Adamic and Glance, 2005; Herring et al., 2005) are consistent with this pattern. Social networks, moreover, have long been known to exhibit homophily (McPherson et al., 2001)—the tendency for contacts to be more similar than ran-

dom pairs of individuals—suggesting that social media sites expose individuals to largely congruent opinions. Furthermore, in laboratory studies people tend to share information that conforms to the group's majority opinion (Moscovici and Zavalloni, 1969; Myers and Bishop, 1970; Schkade et al., 2007; Spears et al., 1990), reinforcing the notion that individuals on social media sites are unlikely to express or discuss dissenting points of view.

Yet despite this seemingly compelling body of evidence, most metrics of political polarization in the general U.S. population have been relatively stable for the last several decades (Baldassarri and Gelman, 2008; Fiorina and Abrams, 2008; Prior, 2013).[1] Moreover, a casual perusal of popular news sites does not reveal a substantial proportion of niche or extreme political content. Indeed, the most popular destination for online news is *Yahoo! News*, a decidedly moderate outlet; and among the top 20 most popular news sites—which in aggregate account for three-quarters of total news traffic—the ideological spectrum ranges from the *New York Times* on the left to *Fox News* on the right, comparable to the ideological span of broadcast news.[2] Further, in a comprehensive study of news consumption, Gentzkow and Shapiro (2011) found that segregation in online news was similar to that of traditional, offline newspapers. Finally, Webster and Ksiazek (2012) find little evidence of audience fragmentation among major media outlets. We are thus left with a puzzle: How do we reconcile the considerable evidence that suggests current online conditions should promote ideological segregation with the apparent lack of significant fragmentation?

We investigate this question by examining detailed web browsing records of 1.2 million anonymized U.S.-located Internet users. We have a nearly complete record of every web page viewed by these individuals over the three-month period between March and May of 2013, a total of 2.3 billion page views. In studying the news consumption habits of this sample of users, we encounter and address three methodological challenges. First, the vast majority of content on news sites concerns sports, entertainment, weather and other topics for which ideological segregation is not meaningful. We use tools from machine learning to identify national and world

[1] Congress, by contrast, has become notably more polarized over time (Poole and Rosenthal, 2001; Prior, 2013).

[2] Based on the Alexa ranking of news outlets (`http://www.alexa.com/topsites/category/Top/News`)

news stories, and to further separate out descriptive reporting from opinion pieces (which we refer to as "news" and "opinion", respectively). Second, we require a measure of each news outlet's ideological leaning. Here we follow past audience-based approaches (Gentzkow and Shapiro, 2011; Tewksbury, 2005) and rely on a site's *conservative share*, the fraction of its readership that supported the Republican candidate in the most recent presidential election. We develop a method to infer this metric, based on examining the relationship between geographic news site access patterns in our dataset and publicly-available county-level voting records. Finally, we need an estimate of ideological segregation, which we define as the average difference in the conservative shares of news outlets visited by two randomly selected individuals. To estimate segregation, we apply hierarchical regression models.

We find that segregation is marginally higher for descriptive news articles accessed via social media (0.12) than for those read by directly visiting a news outlet's home page (0.11). For opinion pieces, however, the pattern is more pronounced, with segregation moving from 0.13 for articles directly obtained from the publisher to 0.17 for socially recommended pieces to a striking 0.20 for articles found via web search. To help interpret these numbers, we note that 0.20 corresponds to the ideological distance between the centrist *Yahoo! News* and the left-leaning *Huffington Post* (or equivalently, *CNN* and the right-leaning *National Review*). But we also find that these more segregating socially recommended and search-based opinion stories account for only a small fraction (2%) of total news consumption; in fact, without separating out opinion from descriptive reporting, we would have largely missed the role of social and search play in shaping consumption. By comparison, directly accessed descriptive reporting comprises over 75% of consumption. As a consequence, the overall level of news segregation is relatively moderate (0.11), which approximately corresponds to the ideological distance between *USA Today* and the *Washington Post*.

The moderate level of segregation we observe could be the result of two qualitatively different individual-level consumption habits. On the one hand, a typical individual might regularly read a variety of liberal and conservative news outlets, but still exhibit a slight left- or right-leaning preference. On the other hand, individuals may choose to only read publications that are ideologically similar to one another, rarely reading opposing perspectives. We find strong evidence for the latter. In particular, users who visit left-leaning news outlets almost never ($< 5\%$ of

the time) read substantive news articles from conservative sites, and vice versa for right-leaning readers, an effect that is even more pronounced for opinion articles. This finding holds both for the majority of individuals who rely on one or two sites (e.g., who almost exclusively read articles from *The New York Times* or *Fox News*) and for those who visit several outlets. Thus, while most people typically consume centrist news content, the minority who read partisan articles are typically unaware of the other side of the political debate.

Our results are directionally consistent with worries that social media reinforce and spur ideological segregation, and moreover, we find a typical consumer reads articles from a cluster of highly ideologically similar news outlets. However, the relative dearth of socially recommended news stories, especially those in the opinion category, and the relatively centrist preferences of most individuals, lead to a moderate overall level of segregation. Moreover we do not observe the relatively extreme choice fragmentation observed in the laboratory. This is likely because laboratory studies in this literature tend to use particularly polarizing political issues, such as the death penalty or abortion rights, for which subjects tend to have a well-formed viewpoint. Our empirical findings suggest these types of stories are a very poor analog for descriptive news and while a better match for opinion content, are probably not representative in this case either.

We can only speculate as to why social media do not appear to be a dominant channel for circulating news, or why individuals do not seek out niche political sites. Perhaps it is because dominant social media platforms are used primarily for entertainment and interpersonal communication rather than for political discussion or advocacy.[3] Perhaps, even though it has grown increasingly easy to produce niche content, consumers simply do not have an appetite for extreme political perspectives.[4] Regardless, the net effect is that while the technological ingredients for ideological fragmentation are in place—and indeed appear to impact consumption—the consequences have thus far been avoided. If, however, the next generation of Internet users increasingly rely on social media to obtain news and opinion, then

[3]The meteoric rise of websites such as Instagram suggest that social sharing has qualitatively affected the consumption of photographs.

[4]Work in media economics, both theoretical and empirical, suggests that content creators respond to consumer preferences (Gentzkow and Shapiro, 2006; George and Waldfogel, 2006; Mullainathan and Shleifer, 2005), including their desired political slant (Baum and Groeling, 2008; Gentzkow and Shapiro, 2010, 2013).

our results suggest that would in turn lead to higher ideological segregation.

2 Data and Methods

Our primary analysis is based on web browsing records collected via the Bing Toolbar, a popular add-on application for the Internet Explorer web browser. Upon installing the toolbar, users can consent to sharing their data via an opt-in agreement, and to protect privacy, all records are anonymized prior to our analysis. Each toolbar installation is assigned a unique identifier, giving the data a panel structure. While it is certainly possible that multiple members of a household share the same browser, we follow the literature by referring to each toolbar installation as an "individual" or "user" (Athey and Mobius, 2012; De los Santos et al., 2012; Gentzkow and Shapiro, 2011).

Based on these toolbar records, we analyze the web browsing behavior of 1.2 million U.S.-located users for the three-month period between March and May of 2013, making this one of the largest studies of web content consumption to date. To ensure this set of users was reasonably active, we drew a random sample of all toolbar users who viewed at least ten webpages during the first week of March 2013. For each user, we have a time-stamped collection of URLs opened in the browser, along with the user's geographic location, as inferred via their IP address. In total, our dataset consists of 2.3 billion distinct page views, with a median of 991 page views per individual.

As with nearly all observational studies of individual-level web browsing behavior, our study is restricted to individuals who voluntarily share their data, which likely creates selection issues. These users, for example, are presumably less likely to be concerned about privacy. Moreover, though our panelists did not report any demographic information, it is generally believed that Internet Explorer users are on average older than the Internet population at large. Instead of attempting to re-balance our sample using difficult-to-estimate and potentially incorrect weights, we acknowledge these shortcomings and note throughout where they might be a concern. When appropriate, we also replicate our analysis on different subsets of the full dataset, increasing the likelihood our results extend beyond the particular sample of users we study. As a further robustness check, we replicate our analysis

on the set of U.S.-located users on the social network Twitter.

2.1 Identifying News and Opinion Articles

We select an initial universe of news outlets (i.e., web domains) via the Open Directory Project (ODP, dmoz.org), a collective of tens of thousands of editors who hand label websites into a classification hierarchy. As of June 2013, 7,923 distinct domains were included in the four primary ODP news categories: news, politics/news, politics/media, and regional/news. Since the vast majority of these news sites receive relatively little traffic, to simplify our analysis we restrict to the one hundred domains that attracted the largest number of unique visitors from our sample of toolbar users.[5] This list of popular news sites includes every major national news source (e.g., *The New York Times*, *The Huffington Post*, and *Fox News*), well-known blogs (e.g., *Daily Kos* and *Breitbart*), and many regional dailies (e.g., *The Seattle Times* and *The Denver Post*). The complete list is given in the Appendix.

Our focus in this paper is on the consumption of U.S. and international text-based news and opinion, corresponding to the content that generally appears in the front section and opinion pages of newspapers. However, the bulk of articles on general news websites do not fall into these categories, but rather relate to sports, weather, lifestyle, entertainment, and similar, largely apolitical categories. Since articles from these categories are much less likely to reflect the political slant of the outlet, our first aim is to filter them out. Given the wide variety of blogs and traditional news outlets that we consider, which stories qualify as "front-section news" or opinion is not immediately obvious in the browsing records. We address this problem from a machine learning perspective, classifying each article based on the words that appear in it.

We build two binary classifiers using large-scale logistic regression: the first selects front-section news and opinion pieces from the universe of articles in the sample; the second starts from the set of articles chosen in the first step, and then separates out descriptive reporting from opinion pieces. To achieve these aims, we require training 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

[5] This list has high overlap with the current Alexa rankings of news outlets (`http://www.alexa.com/topsites/category/Top/News`).

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. embassies closing due to security concerns) has the link `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". We augment this list with the complete set of 39 first and third person pronouns (Pennebaker et al., 2007, 2001), since opinion pieces—unlike descriptive articles—are often written in the first person. 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 4.1 million articles in our corpus, we obtain 1.9 million stories (46%) classified as front-section news or opinion, and of these 11% are classified as opinion. Note that we use the classifer even for outlets that indicate the article category in the URL, which guards against differing editorial policies biasing the results.

Front-section news & opinion (+) vs. "non-news" (−)

Positive	Negative
contributed, democratic	film, today
economy, authorities,	pretty, probably
leadership, read	personal, learn
republican, democrats	technology, mind
country's, administration	posted, isn't

Opinion (+) vs. descriptive news (−)

Positive	Negative
stay, seem	contributed, reporting
important, seems	said, say
isn't, fact	spokesman, experts
actually, reason	interview, expected
latest, simply	added, hers

Table 1: Most predictive words for classifying articles as either news or non-news, and separately, for separating out descriptive news from opinion.

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 and opinion classifier obtains 92% accuracy, and on a hand-labeled set of 100 randomly selected articles from the full corpus, we see 81% accuracy. Furthermore, the fitted model is relatively interpretable, as indicated in Table 1, which lists the words with the largest positive weights (indicating a story is likely front-section news or opinion) and the largest negative weights (indicating a story is likely not news). Accuracy for the opinion classifier is high as well: 96% on a hold-out set of URL-labeled articles, and 88% on a randomly selected subset of articles classified as front-section news or opinion. Table 1 also lists words with the highest positive and negative weights for the opinion classifier.

In addition to separating out descriptive news from opinion, we examine ideological segregation as a function of an article's subjectivity. We measure subjectivity with the Subjectivity Lexicon,[6] introduced by Riloff and Wiebe (2003) and refined in two subsequent papers (Wiebe et al., 2004; Wilson et al., 2005). The Subjectivity Lexicon was built by hand labeling sentences in news articles, and then using natural language processing and machine learning techniques to score individual words (separately by part of speech). Ultimately, the lexicon scores a word as either objec-

[6]Available for download at `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`.

tive, weakly subjective or strongly subjective. For example, the word "unhealthy" is rated as weakly subjective, whereas the synonym "sickly" is rated as strongly subjective.

To compute each article's subjectivity, we assign a value of 0 to objective words and 1 to both weakly and strongly subjective words, and we then average the subjectivity scores of the words in the article. Several variants of determining an article's subjectivity are discussed in Liu (2010), such as limiting to words in the first paragraph of the article, and using various weighting schemes. The simple procedure we employ, however, tends to work adequetely in our setting. In particular, on a hand-labeled set of 100 front-section news and opinion articles rated as either objective, weakly subjective or strongly subjective, the Pearson correlation between the human and the algorithmic (i.e., based on the Subjectivity Lexicon) ratings was 0.49 (the Spearman correlation was 0.41).

2.2 Measuring the Political Slant of Publishers

Algorithmically measuring the ideological leanings of news articles is known to be a difficult problem. In the absence of human ratings, there are no existing methods to reliably assess an article's slant with both high recall and precision.[7] Since our sample has over 1.9 million articles classified as either front-section news or opinion, human labeling is not feasible. We thus follow the literature (Gentzkow and Shapiro, 2010, 2011; Groseclose and Milyo, 2005) and focus not on the slant of individual articles but on the slant of news outlets, ultimately assigning articles the polarity score of the outlet on which they were published. By doing so, we clearly lose some signal. For instance, we mislabel liberal op-eds on generally conservative news sites, and we mark neutral reporting of a breaking event as having the overall slant of the outlet. Nevertheless, such a compromise is common in the literature, and where possible, we attempt to mitigate any resulting biases.

[7]High precision is possible by focusing on the use of highly polarizing words such as "Obamacare," but the recall of this method tends to be very low, meaning most pieces of content are not rated. There are semi-supervised methods that have successfully extended a relatively small number of human ratings to a larger set of news articles via clustering algorithms (Zhou et al., 2011), but these approaches assume individuals read ideologically similar content, leading to potential tautologies in our analysis. Even with human ratings, the wide variety of sites we investigate—ranging from relatively small blogs to national newspapers—exhibit correspondingly diverse norms of language usage, making any content-level assessment of political slant quite difficult.

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 (Gentzkow and Shapiro, 2011; Tewksbury, 2005). Empirical evidence suggests that audience and content-based measures of slant are closely related. 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 (Baum and Groeling, 2008; DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2010). Theoretical models also support this relationship between audience and content-based measures (Gentzkow and Shapiro, 2006; Mullainathan and Shleifer, 2005).

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 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.[8] Reparameterizing so that $\beta_0 = p_d$ and $\beta_1 = p_r - p_d$, we have

$$\mathbb{P}(u_i \text{ views } s) = \beta_0 + \beta_1 \delta_r(u_i) \tag{1}$$

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 (1) over all users in the county, yielding

$$\frac{1}{N_k} \sum_{u_i \in C_k} \mathbb{P}(u_i \text{ views } s) = \beta_0 + \beta_1 \frac{1}{N_k} \sum_{u_i \in C_k} \delta_r(u_i) \tag{2}$$

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

While the left-hand side of (2) 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, so the right-hand side of (2) 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 returns. We thus have the following model:

$$P_k = \beta_0 + \beta_1 R_k \tag{3}$$

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 (3), 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, (3) 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

[8] 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."

Publication	Conservative share	Publication	Conservative share
BBC	0.30	L.A. Times	0.46
New York Times	0.31	Yahoo! News	0.47
Huffington Post	0.35	USA Today	0.47
Washington Post	0.37	Daily Mail	0.47
Wall Street Journal	0.39	CNBC	0.47
U.S. News & World Report	0.39	Christian Sci. Monitor	0.47
Time Magazine	0.40	ABC News	0.48
Reuters	0.41	NBC News	0.50
CNN	0.42	Fox News	0.59
CBS News	0.45	Newsmax	0.61

Table 2: For the 20 most popular news outlets, each outlet's estimated conservative share (i.e., the two-party fraction of its readership that voted for the Republican candidate in the last presidential election), with larger values corresponding to more conservative publications.

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 (3)—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 of the 100 news outlets in our dataset.

Table 2 lists estimated conservative shares for the 20 news outlets attracting the

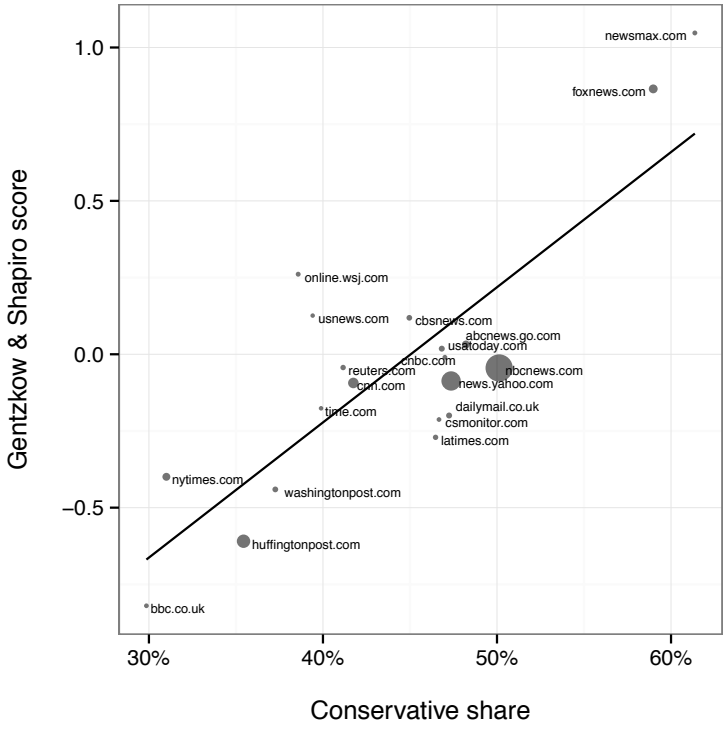


Figure 1: For the 20 most popular news outlets, a comparison of each outlet's estimated conservative share to an alternate measure of its ideological slant as estimated by Gentzkow and Shapiro (2011), where point sizes are proportional to popularity. Among these 20 publications, the correlation between the two scores is 0.82.

most number of unique visitors in our dataset, ranging from the *BBC* and *The New York Times* on the left to *Fox News* and *Newsmax* on the right. While our measure of conservative share is admittedly imperfect, the list does seem largely consistent with commonly held beliefs on the slant of particular outlets.[9] Furthermore, as shown in Figure 1, our ranking of news sites is quite similar to the Gentzkow and Shapiro (2011) list based on 2008 audience data in which users' party affiliations were explicitly collected.[10] Among the top 20 domains, we find a correlation of 0.82

[9]One exception is *The Wall Street Journal*, which we characterize as left-leaning even though it is generally thought to be politically conservative. We note, however, that the most common audience and content-based measures of slant also characterize the paper as relatively liberal (Gentzkow and Shapiro, 2011; Groseclose and Milyo, 2005). Moreover, as a robustness check, we repeated our analysis after omitting *The Wall Street Journal* from our dataset, and found that *none* of our substantive results changed.

[10]The measure from Gentzkow and Shapiro (2011) to which we compare is not precisely conser-

between the two rankings, and across the full set of 41 sites appearing in both lists, the correlation is 0.40. Conservative shares for our full list of 100 domains are given in the Appendix.

2.3 Inferring Consumption Channels

We define and investigate four channels through which an individual can discover a news story: direct, aggregator, social, and search. Direct discovery means a user directly and independently visits a top-level news domain such as `nytimes.com` (e.g., by typing the URL into the browser's address bar, accessing it through a bookmark or performing a "navigational search," explained below), and then proceeds to read articles within that outlet. The aggregator channel refers to referrals from *Google News*—one of the last remaining popular news aggregators—which presents users with links to stories hosted on other news sites.[11] We define the social channel to include referrals from Facebook, Twitter, and various web-based email services. Finally, the search category refers to news stories accessed as the result of web search queries on Google, Bing and Yahoo!.

Given only a time series of webpage views for an individual, it is impossible to perfectly infer the discovery channel. For example, if a user visits a social media site and a news aggregator in two separate browser windows, and then views a news story, we cannot unambiguously determine the channel through which the individual found the article. To mitigate this problem, we employ the following simple heuristic: define the referrer of a news article to be the most recently viewed URL that is a top-level domain (e.g., `nytimes.com` or `facebook.com`, but not a specific story link, such as `nytimes.com/a-news-story`). We then use the referrer to classify the discovery channel. For example, if the referrer is a news domain, such as `foxnews.com`, then the channel is "direct", whereas the channel is "social" if the referrer is, for instance, `facebook.com`. This heuristic is based on two key assumptions: first, users do not type in the "long URL" web addresses assigned to individual articles, but rather are directed there via a previous visit to a top-level domain and a subsequent chain of clicks; and second, top-level domains are not typically shared or posted via email,

vative share, but is closely related.

[11] Most former news aggregators have switched to either producing their own original content, as in the case of *Yahoo! News*, or hosting stories primarily from a single news site, such as AOL directing traffic to their subsidiary, *The Huffington Post*.

social media or aggregators. The second assumption clearly does not hold for web search engines, as many individuals use a search engine simply to navigate to a news publisher's front page. That is, they perform a web search for the publication's name instead of directly typing the outlet's URL into the browser. Since this type of "navigational search" query is widely regarded as a shortcut to typing in the URL directly (Broder, 2002), we define it as a direct view. In contrast, an article view is attributed as coming from search if a user directly accesses it from the search results page.

Note that even when referring pages can be perfectly inferred, there is still genuine ambiguity in how to determine the channel. For example, if an individual follows a Facebook link to a *New York Times* article and then proceeds to read three additional articles at that outlet, are all four articles "social" or just the first? Our inference method addresses this attribution problem by taking the middle ground. Any subsequent article-to-article views (e.g., clicks on a link from a story to a recommended, related story) are classified as "social," whereas an intermediate visit to the news outlet's front page (which we observe as a visit to a top-level domain) results in subsequent views being classified as "direct."

2.4 Limiting to Active News Consumers

As recent studies have noted, only a minority of individuals regularly read online news. For example, a 2012 survey by Pew Research showed that 39% of adults claimed to have read online news in the previous day,[12] a finding supported by observational studies of browsing behavior (Goel et al., 2012a). Because our aim is to understand the preferences and choices of individuals who actively read front-section news and opinion, we limit to the even smaller subset of the population who have read at least 10 substantive news articles (i.e., excluding stories in sports, entertainment, and other apolitical categories) in the three-month timeframe we consider, and who have additionally read at least two opinion pieces. This first requirement of having read at least 10 substantive news articles reduces our initial sample of 1.2 million individuals to 173,450; and the second requirement of having read at least two opinion pieces further reduces the sample to 50,383. Our primary analysis

[12] http://www.people-press.org/2012/09/27/in-changing-news-landscape-even-television-is-vulnerable/

focuses on this 4% of our sample who are active news consumers. Though this subgroup comprises a small fraction of our sample, it is both a natural subpopulation to consider, and arguably one that has a disproportionate impact on political outcomes and policy decisions, a point we return to in the discussion.

3 Ideological Segregation

3.1 Overall Segregation

We begin by defining and estimating the overall ideological segregation in our sample of users. Recall that the conservative share of a news outlet—which we also refer to as the outlet's *polarity*—is the estimated fraction of the publication's readership that voted for the Republican candidate in the most recent presidential election. Thus it is an audience-based measure of the publication's ideological leaning. To (informally) define segregation, we first define the polarity of an *individual* to be the typical polarity of the news content that he or she reads. We then define segregation to be the expected distance between the polarity scores of two randomly selected users. Our definition of segregation is in line with past work (Gentzkow and Shapiro, 2011; White, 1986),[13] and intuitively captures the idea that segregated populations are those in which individuals are, on average, exposed to disparate points of view. However, due to sparsity in the data, this measure of segregation is not entirely straightforward to estimate. In particular, under a naive inference strategy, noisy estimates of user polarities would inflate the estimate of segregation. We thus estimate segregation via a hierarchical Bayesian model (Gelman and Hill, 2007), as described next.

We define the polarity score of an *article* to be the polarity score of the news outlet in which it was published.[14] Now, let X_{ij} be the polarity score of the j-th

[13] One difference is that in traditional measures of residential segregation, individuals are modeled as belonging to one of several discrete groups (e.g., based on race); in our setting, however, individuals lie on a continuous polarity spectrum.

[14] While this is standard practice, it precludes, for example, a conservative outlet that sometimes publishes liberal editorials. Ideally, the classification would be done at the article level, but there are no known methods for reliably doing so.

article read by user i. We model

$$X_{ij} \sim N(\mu_i, \sigma_d^2) \tag{4}$$

where μ_i is the latent polarity score for user i, and σ_d is a global dispersion parameter (to be estimated from the data). To mitigate data sparsity, we further assume the latent variables μ_i are themselves drawn from a normal distribution. That is,

$$\mu_i \sim N(\mu_p, \sigma_p^2). \tag{5}$$

To complete the model specification, we assign weak priors to the hyperparameters σ_d, μ_p and σ_p. Ideally, we would perform a fully Bayesian analysis to obtain the posterior distribution of the parameters. However, for computational convenience, we use the approximate marginal maximum likelihood estimates obtained from the `lmer()` function in the `R` package `lme4` (Bates et al., 2013).

Having specified the model, we can now formally define segregation, which we do in terms of the expected squared distance between individuals' polarity scores. Namely, we define segregation to be $\sqrt{\mathbb{E}(\mu_i - \mu_j)^2}$. After simple algebraic manipulation, our measure of segregation further reduces to $\sqrt{2}\sigma_p$.[15] Higher values of this measure correspond to higher levels of segregation, with individuals more spread out across the ideological spectrum.

Figure 2 shows the distribution of polarity scores (i.e., the distribution of μ_i) for users in our sample. We find that most individuals are relatively centrist, with two-thirds of people having polarity scores between 0.41 and 0.54. In other words, the majority of individuals are ideologically centered at publications ranging from *Reuters* on the center-left to *The Orlando Sentinel* on the center-right. Overal segregation is estimated to be 0.11, which means that for two randomly selected users, the ideological distance between the publications they typically read is on par with that between the centrist *NBC News* and the left-leaning *Daily Kos* (or equivalently, *ABC News* and *Fox News*). Thus, though we certainly find a degree of ideological segregation, it would seem to be relatively moderate, and largely in line with the

[15]Though the distributional assumptions we make are standard in the literature (Gelman and Hill, 2007), our modeling choices of course affect the estimates we obtain. As a robustness check, we note that a naive, model-free estimation procedure yields qualitatively similar, though ostensibly less precise, results.

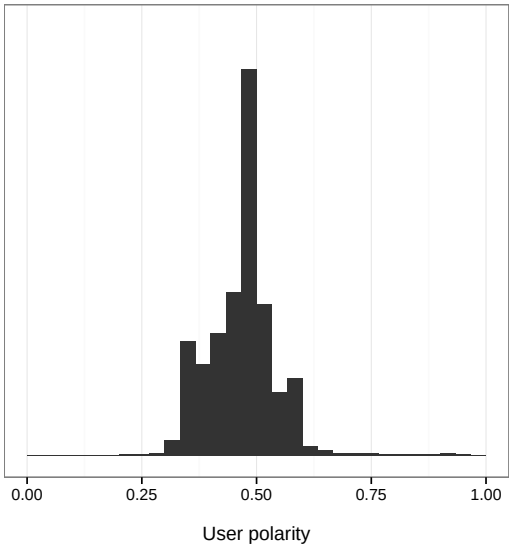


Figure 2: The distribution of individual-level polarity, where each individual's polarity score is the (model-estimated) average conservative share of the news outlets he or she visits.

most recent past assessment, based primarily on 2006 data (Gentzkow and Shapiro, 2011). Notably, given the interim rise of social media and personalization—and the accompanying predictions of ideological fragmentation—it is perhaps surprising that this would be the case, an issue we investigate in detail below.

3.2 Segregation by Channel and Article Subjectivity

When measuring segregation across various distribution channels and levels of article subjectivity, the data sparsity issues we encountered above are exacerbated. For example, even among active news consumers, relatively few individuals regularly read news articles from both aggregators and social media sites. And when we further segment articles into opinion and descriptive news, it compounds the problem. However, the polarity of consumption for a user across channels should be correlated; for example, the opinion pieces one reads from Facebook are likely ideologically related to the articles one reads from Google News. There is thus opportunity to improve our estimates by "sharing strength" across channels and subjectivity levels, and accordingly to jointly estimate the segregation parameters of interest. Joint estimation with weak priors also mitigates channel selection issues.

As discussed in Section 2, we consider four consumption channels (aggregator, direct, web search and social media), and two subjectivity classes (descriptive reporting and opinion pieces) for each channel. Thus, in total we seek to measure segregation along eight subjectivity-by-channel dimensions. Let X_{ijk} denote the polarity of the j-th article that user i reads in the k-th subjectivity-by-channel category, where we recall that the polarity of an article is defined to be the conservative share of the site on which it was published. Generalizing our hierarchical Bayesian framework, we model

$$X_{ijk} \sim N(\mu_i^k, \sigma_d^2) \tag{6}$$

where μ_i^k is the k-th component in the latent 8-dimensional polarity vector $\vec{\mu_i}$ for user i, and σ_d is a global dispersion parameter. As before, we deal with sparsity by further assuming a distribution on the latent variables $\vec{\mu_i}$ themselves. In this case, we model each individual's polarity vector as being drawn from a multivariate normal:

$$\vec{\mu_i} \sim N(\vec{\mu_p}, \Sigma_p) \tag{7}$$

where $\vec{\mu_p}$ and Σ_p are global hyperparameters. The full Bayesian model is analyzed by assigning weak priors to the hyperparameters and computing posterior distributions of the latent variables, but in practice we simply fit the model with marginal maximum likelihood.

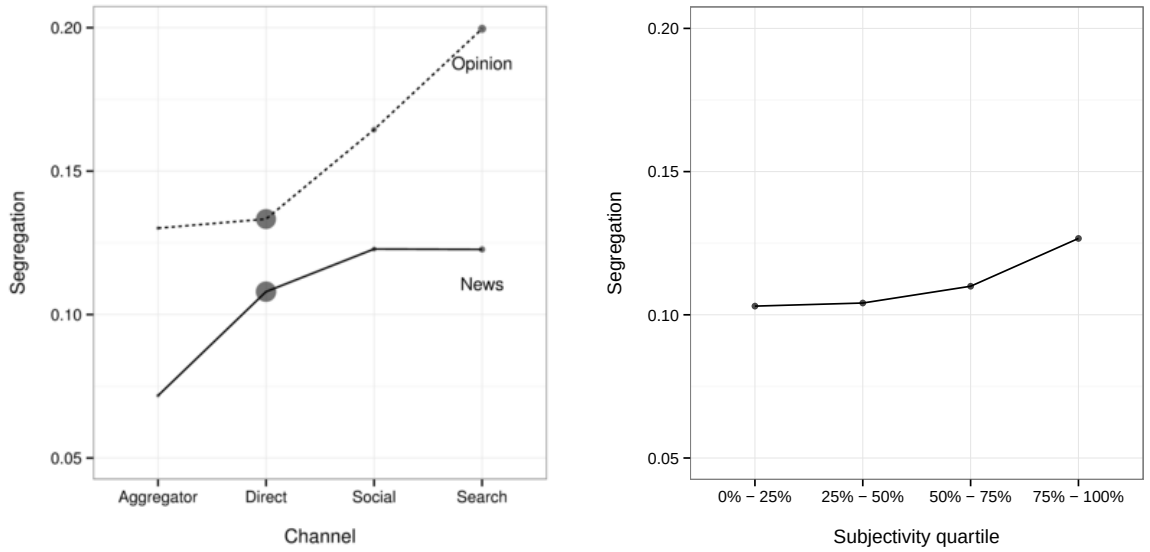
As with the analysis in Section 3.1, the diagonal entries of the covariance matrix Σ_p yield estimates of segregation for each of the eight subjectivity-by-channel categories. In particular, letting σ_k^2 denote the k-th diagonal entry of Σ_p, segregation in the k-th category is $\sqrt{2}\sigma_k$. Table 3 lists these diagonal parameter estimates.[16] The off-diagonal entries of Σ_p measure the relationship between categories of one's ideological exposure. For example, after normalizing Σ_p to generate the corresponding correlation matrix, we find the correlation between social media-driven descriptive news and opinion is 0.71 (i.e., an individual's ideological exposure is quite similar across these two categories). The full correlation matrix is included in the Appendix.

To help visualize these model estimates, Figure 3(a) plots segregation across

[16] Given the large sample size, all estimates are statistically significant well beyond conventional levels.

	Front-section news		Opinion	
Consumption channel	μ_p	σ_p	μ_p	σ_p
Aggregator	0.44	0.051	0.44	0.092
Direct	0.47	0.076	0.47	0.094
Social	0.46	0.087	0.47	0.12
Search	0.46	0.087	0.46	0.14

Table 3: Fitted parameters for the Bayesian model used to estimate ideological consumption by channel and subjectivity type, as described in Eqs. (6) and (7). The column μ_p indicates the corresponding entry of $\vec{\mu_p}$, and the column σ_p indicates the corresponding diagonal entry of the model-estimated covariance matrix Σ_p.



(a) Segregation for various channels through which news articles are read, for both descriptive news (solid line) and opinion (dotted line). Point sizes indicate the relative fraction of partisan traffic attributed to each source, normalized separately within the news and opinion lines.

(b) Segregation as a function of article subjectivity (as estimated by word usage), with the most objective articles appearing in the left-most bin, and the most subjective in the right-most bin.

Figure 3: Estimates of ideological segregation across consumption channels and subjectivity types, where segregation is defined as the expected difference in polarity between two randomly selected individuals.

the four consumption channels, for both opinion and descriptive news. The size of the markers is proportional to total consumption within the corresponding channel,

normalized separately for opinion and descriptive news. To ground the scale of the y-axis, we note that among the top 20 most popular news outlets, conservative share ranges from 0.30 for the liberal BBC to 0.61 for the conservative Newsmax.

Figure 3(a) indicates that social media is indeed associated with higher segregation than direct browsing. For descriptive news this effect is subtle, with segregation increasing from 0.11 for direct browsing to 0.12 for articles linked to from social media; and for opinion stories. However for opinion pieces, the effect is more pronounced, rising from 0.13 to 0.16. It is unclear whether this increased segregation is the effect of active algorithmic filtering of the news stories appearing in one's social feed (Pariser, 2011), the result of ideological similarity among one's social contacts (Goel et al., 2010; McPherson et al., 2001), or both. In any case, however, our results are directionally consistent with worries that social media increase segregation.

We further find that search engines are associated with the highest levels of segregation among the four channels we investigate: 0.12 for descriptive news and 0.20 for opinion. Some authors have argued that web search personalization is a key driver of such effects (Pariser, 2011). There are two alternative explanations. The first is that users implicitly influence the ideological leanings of search results through their query formulation, by, for example, issuing a query such as "obamacare" instead of "health care reform" (Borra and Weber, 2012). The second is that even when presented with the same search results, users are more likely to select outlets that share their own political ideology, especially for opinion content (Garrett, 2009; Iyengar and Hahn, 2009; Munson and Resnick, 2010). Though we cannot identify the underlying mechanism driving our results, our findings do suggest that the relatively recent ability to instantly query large corpora of news articles at least contributes to increased ideological segregation. Insofar as we view web search as expanding choices, the evidence thus indicates that increased choice does amplify ideological segregation, at least marginally for descriptive news, and substantially for opinion stories

At the other end of the spectrum, aggregators, perhaps surprisingly, exhibit the lowest segregation. In particular, even though aggregators return personalized news results from a broad set of publications with disparate ideological leanings (Das et al., 2007), the overall effect is relatively low segregation. Though even for aggregators, segregation for opinion (0.13) is far higher than for descriptive news (0.07),

possibly due to personalization or simply actively select which sources to read (Garrett, 2009; Iyengar and Hahn, 2009; Munson and Resnick, 2010).

Given that our results are directionally consistent with filter bubble concerns, how is it that in Section 3.1 we found largely moderate overall levels of segregation? The answer is simply that only a relatively small fraction of consumption is of opinion pieces or from polarizing channels (social and search). Indeed even after having removed apolitical categories like sports and entertainment, opinion still only constitutes 6% of news consumption. Further, both among descriptive news and opinion, direct browsing is the dominant consumption channel (79% and 67%, respectively), dwarfing social media and search engines. Finally, we note that even the most extreme segregation that we observe (0.20 for opinion articles returned by search engines) is not, in our view, astronomically high. In particular, that level of segregation corresponds to the ideological distance between *Fox News* and *Daily Kos*, which represents meaningful differences in coverage (Baum and Groeling, 2008), but is within the mainstream political spectrum. Consequently, though the predicted filter bubble and echo chamber mechanisms do appear to increase online segregation, their overall effects at this time are somewhat limited.

We conclude this section with a sensitivity analysis in which we examine segregation with a finer-grained measure of article subjectivity. As described in Section 2.1, we assign each article a score between 0 and 1 that indicates the fraction of words in the article that convey a subjective stance. For simplicity and consistency with our previous analysis, we bin articles into four discrete quartiles, and then fit a model analogous to the one described above. Specifically, letting X_{ijk} indicate the j-th article read by user i in the k-th subjectivity quartile ($1 \leq k \leq 4$), we model

$$X_{ijk} \sim N(\mu_i^k, \sigma_d^2) \qquad \vec{\mu_i} \sim N(\mu_p, \Sigma_p). \tag{8}$$

Estimates of segregation by article subjectivity are presented in Figure 3(b). Consistent with our finding that opinion articles are associated with higher segregation than descriptive news, Figure 3(b) shows that segregation increases with this measure of subjectivity as well. However, we note that even though the Internet has likely made it easier to publish, promote and discover subjective news content, the most subjective quartile of news stories is still not alarmingly more segregating than the least subjective (0.10 versus 0.13).

3.3 Ideological Isolation

We have thus far examined segregation in terms of the distance between individuals' *mean* ideological positions. Consequently, our preceding results say nothing of within-individual variation. It could be the case, for example, that individuals typically consume content from a variety of ideological viewpoints, though ultimately skewing toward the left or right, leading to moderate overall segregation. Alternatively, individuals might be tightly concentrated around their ideological centers, only rarely reading content from across the political spectrum. These two potential patterns have markedly different implications for the broader issues of political discussion and consensus formation.

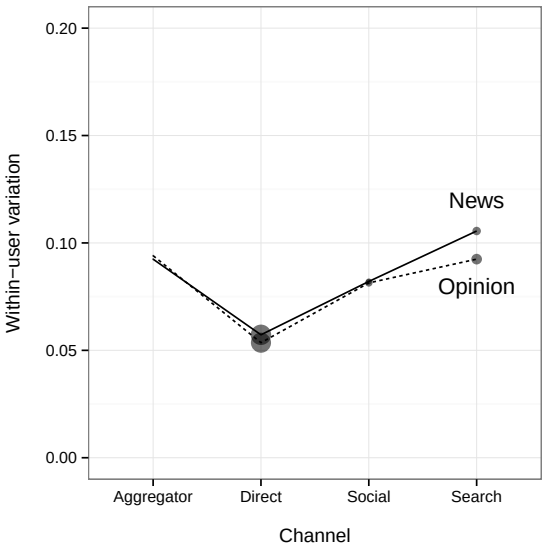
To investigate this question of within-user variation, we start by looking at the dispersion parameter σ_d in the overall consumption model described by Eqs. (4) and (5). We find that $\sigma_d = 0.06$, indicating that individuals typically read publications that are tightly concentrated ideologically.

This finding of within-user ideological concentration is driven in part by the fact that individuals often simply turn to a single news source for information: 78% of users get the majority of their news from a single publication, and 94% get a majority from at most two sources. As shown in the Appendix, however, this concentration effect holds even for those who visit multiple news outlets. In particular, Figure A.1 plots estimates of within-user variation as a function of the number of news outlets an individual visits. For example, among individuals who visited at least 10 news outlets, we find $\sigma = 0.09$, approximately the distance between Reuters and NBC News. Thus, even when individuals visit a variety of news outlets, they are, by and large, frequenting publications with similar ideological perspectives.

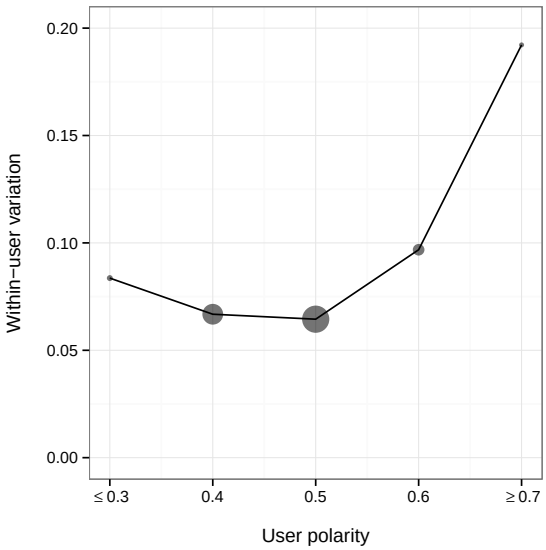
We now investigate ideological isolation across consumption channels and subjectivity categories. For each of the eight subjectivity-by-channel categories and for each user, we first estimate the variance of the polarities of articles read by that user in that category.[17] For each category, we then average these individual-level estimates of variance (and take the square root of the result) to attain category-level estimates. Figure 4(a) plots these estimates of within-user variation by channel and subjectivity.

Across all four consumption channels, Figure 4(a) shows that descriptive and

[17]For each category, we restrict to users who read at least two articles in that category.



(a) Within-user variation across various consumption channels, for both descriptive news (solid line) and opinion (dotted line). Point sizes indicate the relative fraction of traffic attributed to each source, normalized separately within the news and opinion lines.



(b) Within-user variation as a function of individual-level polarity, where an individual's polarity score is the average conservative share of news outlets he or she visits. Point sizes indicate the relative number of individuals in each polarity bin.

Figure 4: For a typical individual, variation (i.e., standard deviation) of the conservative share of news outlets he or she visits, as a function of channel and individual-level polarity.

opinion articles are associated with similar levels of within-user variation. Social media, however, is associated with higher variation than direct browsing. Though this may at first seem surprising given that social media also has relatively high segregation, the explanation is clear in retrospect: when browsing directly, individuals typically visit only a handful of news sources, whereas social media sites expose users to more variety. Likewise, web search engines, while associated with high segregation, also have relatively high diversity. Finally, the highest levels of within-user spread is observed for aggregators, as one might have expected.

We similarly examine within-user ideological variation as a function of user polarity (i.e., mean ideological preference). In this case, we first bin individuals by their polarity—as estimated in Section 3.1—and then compute the individual-level variation of article polarity, averaged over users in each group. As shown in Figure 4(b), within-user variation is small and quite similar for users with polarity

ranging from 0.3 to 0.6. Interestingly, however, the 2% of individuals with polarity of approximately 0.7 or more (significantly to the right of Fox News) exhibit a strikingly high within-user variation of 0.17.

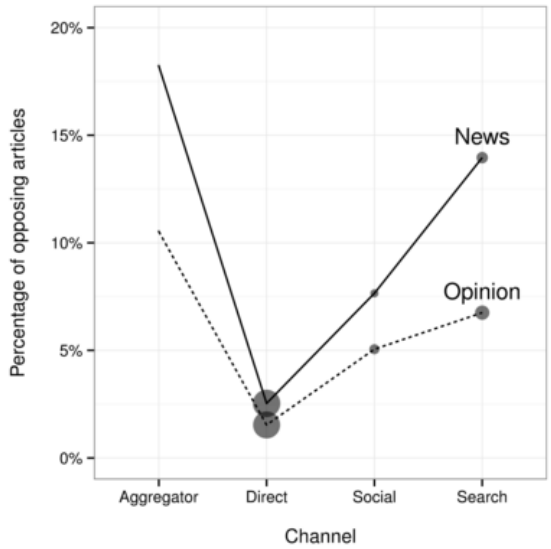
This preceding result prompts a question: Does the high within-user variation we see among extreme right-leaning readers result from them reading articles from across the political divide, or are they simply reading a variety of right-leaning publications? More generally, across channels and subjectivity types, what is the relationship between within-user variation and exposure to ideologically divergent news stories? We conclude our analysis of ideological isolation by examining these questions.

We start by defining a news outlet as left-leaning (resp., right-leaning) if it is in the bottom (resp., top) third of the 100 outlets we consider; the full ranked list of publications is given in the Appendix. The left-leaning publications include newspapers from liberal areas, such as the *San Francisco Chronicle* and the *New York Times*, as well as blogs such as the *Huffington Post* and *Daily Kos*; the right-leaning set includes newspapers from historically conservative areas, such as the *Fort Worth Star-Telegram* and the *Salt Lake Tribune*, and online outlets such as *Newsmax* and *Breitbart*; and centrist publications (i.e., the middle third) include, for example, *Yahoo! News* and *USA Today*. We refer to the combined collection of left- and right-leaning outlets as *partisan*.

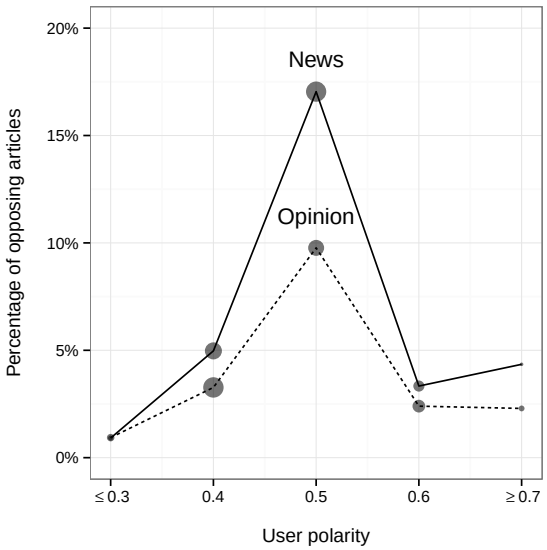
For each user who reads at least two partisan articles, define his or her liberal exposure ℓ_i to be the fraction of partisan articles read that are left-leaning. We define an individual's *opposing partisan exposure* $o_i = \min(\ell_i, 1 - \ell_i)$. Thus, for individuals who predominantly read left-leaning articles, o_i is the proportion of partisan articles they read that are right-leaning, and vice-versa. We note o_i is only defined for the 82% of individuals in our sample that have read at least two partisan articles.

Figure 5 shows average opposing partisan exposure, partitioned by article channel and subjectivity (Figure 5(a)), and by user polarity (Figure 5(b)).[18] For every subset we consider, only a small minority of articles—less than 20% in all cases, and is less than 5% for all non-centrist users—comes from the opposite side of an individual's preferred partisan perspective. Addressing the question posed above, even

[18]To compute the estimates of average opposing partisan exposure shown in 5(a), o_i is computed separately for each of the eight subjectivity-by-channel categories by restricting to the relevant articles, and limiting to users who read at least two partisan articles in that category.



(a) Percentage of opposing articles read by a typical individual across various consumption channels, for both descriptive news (solid line) and opinion (dotted line). Point sizes indicate the relative fraction of traffic attributed to each source, normalized separately within the news and opinion lines.



(b) Percentage of opposing articles read by a typical individual, for both descriptive news (solid line) and opinion (dotted line), as a function of individual-level polarity, where an individual's polarity score is the average conservative share of news outlets he or she visits. Point sizes indicate the relative number of individuals in each polarity bin.

Figure 5: 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 channel and individual-level polarity.

extreme right-leaning readers have strikingly low opposing partisan exposure (3%); thus, their relatively high within-user variation is a product of reading a variety of centrist and right-leaning outlets, and not exposure to truly ideologically diverse content. In contrast, the relatively higher levels of within-user variation associated with social media and web search (Figure 4(a)) do translate to increased exposure to opposing viewpoints, though this effect is still small. Lastly, we note that these findings are only partially a consequence of individuals typically visiting just a small number of news outlets. As shown in Figure A.2, even among those who visit 5–9 news outlets, average opposing partisan exposure is only 14%; and it is still just 20% for those users visiting 10–14 outlets.

Summarizing our results on ideological isolation, we find that individuals gener-

ally read publications that are ideologically quite similar, and moreover, users that regularly read partisan articles are almost exclusively exposed to only one side of the political spectrum. In this sense, many, if not all, users exist in a so-called echo chamber. We note, however, two key differences between our findings and some previous discussions of this topic (Pariser, 2011; Sunstein, 2009). First, the lack of within-user variation appears largely driven by direct browsing, not social media or personalization. Second, consistent with Gentzkow and Shapiro (2011), the outlets that dominate partisan news coverage are still relatively mainstream, ranging from *The New York Times* on the left to *Fox News* on the right; the more extreme ideological sites (e.g., *Breitbart*), which presumably benefited from the rise of online publishing, do not appear to qualitatively impact the dynamics of news consumption.

4 Ideological Segregation on Twitter

While our preceding analysis investigated a variety of channels through which individuals read the news, it was limited to a particular opt-in sample of individuals, namely those who use the Internet Explorer toolbar application and who further consent to making their (anonymized) browsing data available. In this 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.[19]

Aside from the selection effects mentioned above, 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

[19]Twitter positions itself as a fully-customizable information portal, exemplified by the first paragraph on `www.twitter.com/about`: "Twitter is a real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting. Simply find the accounts you find most compelling and follow the conversations."

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.[20] 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 (i.e., "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) at least 10 front-section news articles and at least 2 opinion pieces.[21] In total, 1.5 million users meet all of these restrictions.

We begin our analysis by estimating the distribution of user polarity, the average conservative share of the articles to which a user is exposed (i.e., articles that are posted by an account the user follows). 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, where we recall that the polarity of an article is the conservative share of the news outlet in which it appeared.

Figure 6 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

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

[21]As with the toolbar analysis, articles were classified as front-section news and opinion according to the methods described in Section 2.

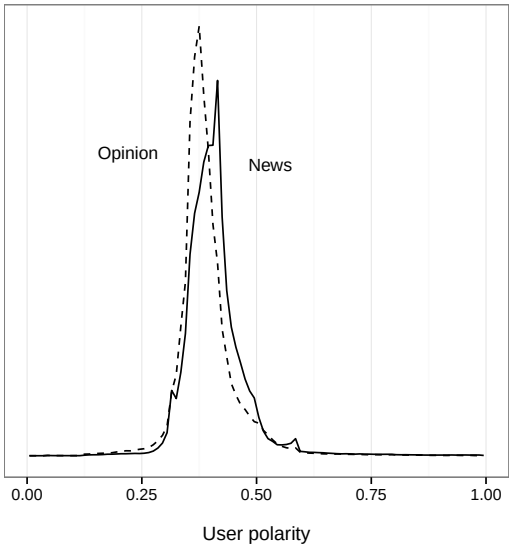
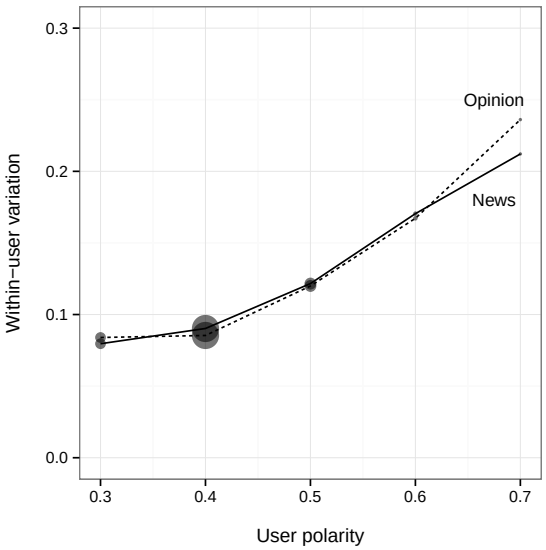


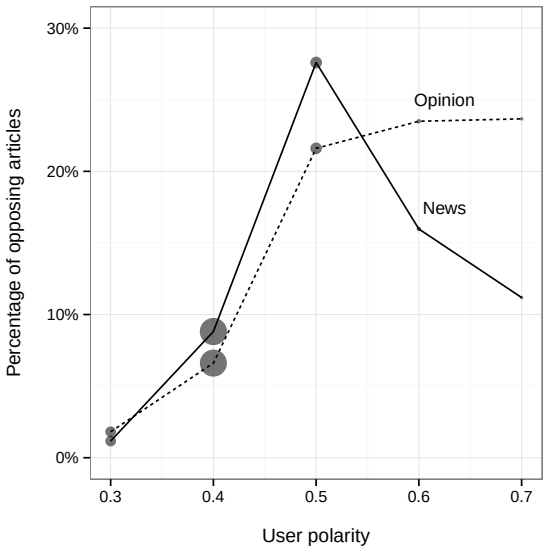
Figure 6: Distribution of individual-level polarity for Twitter users, where an individual's polarity score is the average conservative share of news outlets to which he or she is exposed, computed separately for descriptive news articles (solid line) and opinion pieces (dashed line).

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, we do not find evidence of significant ideological fragmentation on Twitter.

As discussed in our analysis of the toolbar data, user polarity only reflects the *average* ideological position of articles to which a user is exposed, and nothing more about the article distribution. In particular, it could be the case that individuals are largely concentrated around their ideological centers, or alternatively, that they have diverse exposure, while still exhibiting left- or right-leaning tendency. We investigate this question, as before, via 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



(a) Within-user variation as a function of individual-level polarity.



(b) Opposing partisan views as a function of individual-level polarity.

Figure 7: Within-user variation and opposing partisan views 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).

results are plotted in Figure 7, as a function of user polarity.

As indicated by Figure 7(a), 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 4(a)), 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.

To further understand the individual-level consumption distribution, we plot opposing partisan exposure in Figure 7(b), 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.

5 Discussion and Conclusion

We began our investigation with a puzzle: laboratory experiments and theoretical arguments suggest that the rise of online publishing, social media, and personalized recommendations should create a so-called filter bubble or echo chamber in which individuals are ideologically isolated (Pariser, 2011; Sunstein, 2001, 2009); yet mainstream news outlets still dominate the market, and by most metrics political polarization in the general U.S. population has not spiked in recent years (Baldassarri and Gelman, 2008; Fiorina and Abrams, 2008). We reach a simple and intuitive resolution to this apparent paradox by conducting one of the largest studies of online news consumption to date.

We find that stories shared on social media or found via web search engines are indeed more segregating than those an individual reads by directly visiting news sites, an effect that is almost entirely driven by opinion articles. However, a relatively small amount of online news consumption is driven by the polarizing social and search channels, and opinion pieces—which are typically the focus of laboratory studies—constitute just 6% of articles relating to world or national news. Indeed we may have missed the effect entirely if we had not carefully separated out opinion content using natural language processing. Rather, we find that individuals typically consume descriptive reporting, and do so by directly visiting a handful of their preferred news outlets. Even within opinion, we do not see the extreme choice segmentation observed in the lab. This is likely because the hot-button issues used in those studies, such as the death penalty and abortion rights, are a poor analog for your average opinion piece.

Putting it all together, despite the plethora of available news sources, there seems to be limited consumer demand for content more extreme than that available

on broadcast television. Within this range, though, we do find that individuals have little exposure to ideologically diverse perspectives, especially for opinion, a phenomenon that likely occurs in both online and offline news consumption (Gentzkow and Shapiro, 2011). Thus, though many elements of ideological fragmentation are operating as predicted (directionally) by filter bubble theories, the overall impact of these factors appears to be somewhat limited at this time.

Although we validate our core findings on two different datasets, our study is subject to some limitations. First, as with past work (Gentzkow and Shapiro, 2010, 2011; Groseclose and Milyo, 2005), for methodological tractability we focus on the ideological slant of news outlets, as opposed to that of specific articles. As such, we would misinterpret, for example, the news preferences of an individual who primarily reads liberal articles from generally conservative sites. We suspect, however, that this type of behavior is relatively limited, in part because individuals typically visit ideologically similar news outlets, suggesting their own preferences are in line with those of the sites that they frequent. Second, we focus exclusively on news consumption itself, and not on the consequences such choices have on, for example, voting behavior or policy preferences.[22] Given that we find social media and web search have a limited impact on news exposure—although these channels are more important for opinion—it is likely that their effects on attitudes and behaviors are corresponding small. It is, however, possible—and even plausible—that of the hundreds of news articles one reads, a single, persuasive opinion story shared via social media could have the greatest impact. Finally, and related to the previous point, as we have focused our study on the (natural) subpopulation of active news consumers, it is unclear what impact recent technological changes have on the majority of individuals who have little exposure to the news, but who may get that limited amount largely from social media.

So while it seems we have thus far largely avoided the detrimental, segregating effects of social media and personalization, what the next generation of Internet users will experience is less certain, which is especially relevant since social networking services tend to skew young (Pew, 2013). In particular, our results suggest that if social media become the predominant channel for disseminating news that transformation could significantly increase ideological segregation. Looking forward, our

[22]Establishing and measuring the causal effects of partisan news exposure is difficult, though not impossible (Prior, 2013).

substantive and methodological contributions provide a framework for understanding and monitoring the effects of future systems for producing, distributing, and consuming online news.

References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM.

Athey, S. and Mobius, M. (2012). The impact of news aggregators on internet news consumption: The case of localization. Technical report, Harvard University.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM.

Baldassarri, D. and Gelman, A. (2008). Partisans without constraint: Political polarization and trends in american public opinion1. *American Journal of Sociology*, 114(2):408–446.

Bates, D., Maechler, M., and Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

Baum, M. A. and Groeling, T. (2008). New media and the polarization of american political discourse. *Political Communication*, 25(4):345–365.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bernhardt, D., Krasa, S., and Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics*, 92(5):1092–1104.

Borra, E. and Weber, I. (2012). Political insights: exploring partisanship in web search queries. *First Monday*, 17(7).

Broder, A. (2002). A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM.

De los Santos, B., Hortacsu, A., and Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review*, 102(6):2955–2980.

DellaVigna, S. and Kaplan, E. (2007). The Fox News effect: media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Downs, A. (1957). *An economic theory of democracy*. New York.

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the American public. *Annu. Rev. Polit. Sci.*, 11:563–588.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from US daily newspapers. *Econometrica*, 78(1):35–71.

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Gentzkow, M. and Shapiro, J. M. (2013). Ideology and online news. Working paper.

George, L. M. and Waldfogel, J. (2006). The "New York Times" and the market for local newspapers. *The American Economic Review*, 96(1):435–447.

Goel, S., Hofman, J. M., and Sirer, M. I. (2012a). Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*.

Goel, S., Mason, W., and Watts, D. J. (2010). Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*, 99(4):611.

Goel, S., Watts, D. J., and Goldstein, D. G. (2012b). The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 623–638. ACM.

Groseclose, T. and Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*, pages 527–538. International World Wide Web Conferences Steering Committee.

Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. (2005). Conversations in the blogosphere: An analysis. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 107b–107b. IEEE.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA. ACM.

Lawrence, E., Sides, J., and Farrell, H. (2010). Self-segregation or deliberation? blog readership, participation, and polarization in American politics. *Perspectives on Politics*, 8(01):141–157.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:568.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of personality and social psychology*, 47(6):1231.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 1. MIT Press.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Moscovici, S. and Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of personality and social psychology*, 12(2):125.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *American Economic Review*, pages 1031–1053.

Munson, S. A. and Resnick, P. (2010). Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466. ACM.

Myers, D. G. and Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science*.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, page 71.

Poole, K. T. and Rosenthal, H. (2001). D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting. *Legislative Studies Quarterly*, pages 5–29.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, 16:101–127.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics.

Schkade, D., Sunstein, C. R., and Hastie, R. (2007). What happened on deliberation day? *California Law Review*, pages 915–940.

Spears, R., Lea, M., and Lee, S. (1990). De-individuation and group polarization in computer-mediated communication. *British Journal of Social Psychology*, 29(2):121–134.

Speretta, M. and Gauch, S. (2005). Personalized search based on user search histories. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 622–628. IEEE.

Sunstein, C. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2009). *Republic.com 2.0*. Princeton University Press.

Tewksbury, D. (2005). The seeds of audience fragmentation: Specialization in the use of online news sites. *Journal of broadcasting & electronic media*, 49(3):332–348.

Webster, J. G. and Ksiazek, T. B. (2012). The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of communication*, 62(1):39–56.

White, M. J. (1986). Segregation and diversity measures in population distribution. *Population index*, 52(2):198–221.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.

Zhou, D. X., Resnick, P., and Mei, Q. (2011). Classifying the political leaning of news articles and users from user votes. In *ICWSM*.

Appendix

Table A.1: 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

Conservative shares for the top 100 news outlets
(Continued from previous page)

	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

† Includes Star-Ledger, South Jersey Times, Warren Reporter, Independent Press, Cranford Chronicle, The Times, Jersey Journal, Messenger Gazette and Suburban News

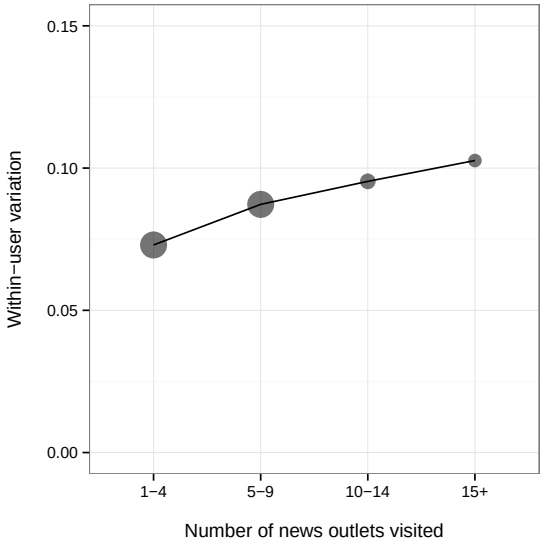


Figure A.1: For a typical individual, within-user variation (i.e., standard deviation) of the conservative share of news outlets he or she visits, as a function of the number of outlets visited.

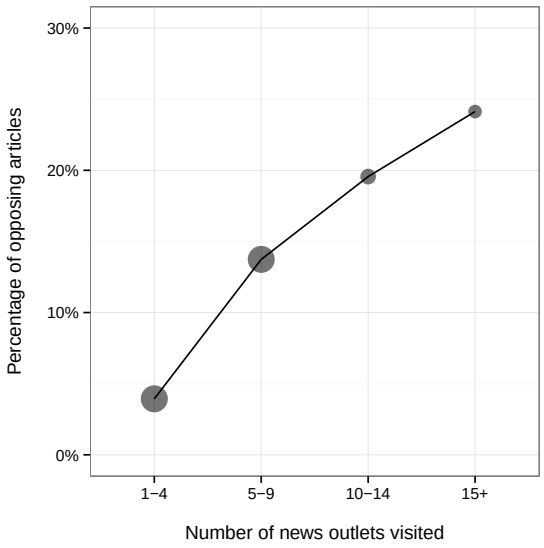


Figure A.2: For a typical individual, fraction of partisan articles that are on the opposite side of the ideological spectrum from those he or she generally reads, as a function of the number of news outlets visited.

		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 A.2: Variance-covariance matrix for the model used to estimate ideological consumption by channel and subjectivity type, as described in Eqs. (6) and (7).

		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 A.3: Correlation matrix for the model used to estimate ideological consumption by channel and subjectivity type, as described in Eqs. (6) and (7).