Fair and Balanced?
Quantifying Media Bias through Crowdsourced Content Analysis*

Ceren Budak
Microsoft Research

Sharad Goel
Stanford University

Justin M. Rao
Microsoft Research

Abstract

It is widely thought that news organizations exhibit ideological bias, but rigorously quantifying such slant has proven methodologically challenging. Through a combination of machine learning and crowdsourcing techniques, we investigate the selection and framing of political issues in 15 major U.S. news outlets. Starting with 803,146 news stories published over 12 months, we first used supervised learning algorithms to identify the 14% of articles pertaining to political events. We then recruited 749 online human judges to classify a random subset of 10,950 of these political articles according to topic and ideological position. Our analysis yields an ideological ordering of outlets consistent with prior work. We find, however, that news outlets are considerably more similar than generally believed. Specifically, with the exception of political scandals, we find that major news organizations present topics in a largely non-partisan manner, casting neither Democrats nor Republicans in a particularly favorable or unfavorable light. Moreover, again with the exception of political scandals, there is little evidence of systematic differences in story selection, with all major news outlets covering a wide variety of topics with frequency largely unrelated to the outlet's ideological position. Finally, we find that news organizations express their ideological bias not by directly advocating for a preferred political party, but rather by disproportionately criticizing one side, a convention that further moderates overall differences.

Word Count: 6,464

*Budak (cbudak@microsoft.com) is the corresponding author. We thank Seth Flaxman, Matthew Salganik, and Sid Suri for comments.

1 Introduction

To what extent are the U.S. news media ideologically biased? Scholars and pundits have long worried that partisan media may distort one's political knowledge and in turn exacerbate polarization. Such bias is believed to operate via two mechanisms: issue *filtering* (i.e., selective coverage of issues) (Iyengar and Kinder, 2010; Krosnick and Kinder, 1990; McCombs and Shaw, 1972), and issue *framing* (i.e., how issues are presented) (Gamson, 1992; Gamson and Lasch, 1981; Gamson and Modigliani, 1989; Iyengar, 1994; Nelson et al., 1997a; Nelson and Kinder, 1996; Nelson et al., 1997b). Prior work has indeed shown that U.S. news outlets differ ideologically (Patterson, 1993; Sutter, 2001), and can be reliably ordered on a liberal-to-conservative spectrum (Gentzkow and Shapiro, 2010; Groseclose and Milyo, 2005). There is, however, considerable disagreement over the magnitude of these differences (D'Alessio and Allen, 2000), in large part due to the difficulty of quantitatively analyzing the hundreds of thousands of articles that major news outlets publish each year. In this paper, we measure issue filtering and framing at scale by applying a combination of machine learning and crowdsourcing techniques. We find that on both dimensions, filtering and framing, U.S. news outlets are substantially more similar—and less partisan—than generally believed.

We focus our attention on 803,146 articles published during 2013 on 15 popular online news sites, including well-known traditional newspapers like the *New York Times*, mainstream online-only outlets such as *Fox News*, and two political blogs. First, we used methods from large-scale machine learning to separate political news from apolitical topics, such as sports, weather, and celebrity gossip, which are far less susceptible to partisan media bias. Ultimately, 114,814 (14%) of the articles were classified as political news. From this set of political articles, we constructed a sample of approximately 11,000 articles by randomly selecting two articles per outlet per day, where the sampling weights are proportional to the fraction of traffic an article accounted for at a given outlet.[1] These articles were then read and analyzed by a labor force of 749 online crowd workers from Amazon Mechanical Turk. Workers categorized articles into one or more of 14 high-level topics (e.g., economy, gay rights, and Democratic scandals)[2], and also judged whether an article was descriptive reporting or an opinion piece. Workers further evaluated whether each article was generally positive, negative, or neutral toward members of the Democratic party, and separately, toward members of the Republican party (i.e., each worker scored the article on two ideological dimensions). Importantly, workers were not asked to

[1] Readership statistics were obtained via the Bing Toolbar.

[2] These categories were determined through Latent Dirichlet Allocation (LDA), a commonly used topic modeling technique (Blei et al., 2003).

determine whether an article was "biased," which would have been a difficult task even for experts since it necessarily involves determining an objective "truth." We note that these workers do not constitute a representative sample of the U.S. population (Berinsky et al., 2012); however, as discussed in the Appendix, that does not appear to qualitatively affect our results.

We obtain an outlet-level partisanship rating by averaging the scores of all articles appearing in a given news outlet. This rating yields an ideological ordering of news organizations that is largely consistent with those from previous audience-based (Bakshy et al., 2015; Gentzkow and Shapiro, 2011) and content-based methods (Groseclose and Milyo, 2005).[3] We find, however, that the magnitude of the differences across outlets is surprisingly small. For example, 21% of descriptive news articles in the *New York Times* are net left-leaning, while 69% are neutral. In comparison, 16% for *Fox News* articles are net left-leaning and 60% are neutral.[4] Even though these two news outlets are regularly held up as exemplars of a partisan media establishment, we find that they in fact exhibit relatively small ideological differences in a representative sample of their political reporting.

We show that these muted differences are the result of four empirical regularities. First, most political topics are generally non-partisan. For example, articles about national security and civil rights are typically neutral toward both Republicans and Democrats. Second, to the extent that outlets present topics in a manner that is net pro-left or net pro-right, news organizations are largely in line with one another. For example, articles about Democratic scandals generally lean to the right, regardless of the outlet in which they appear, and similarly, articles about gay rights generally lean to the left. Third, news outlets rarely cast either Democrats or Republicans in a positive light but simply differ in the extent of their criticism, dampening partisan differences. For example, the three most conservative outlets—the *Breitbart News Network*, *Fox News*, and the *Wall Street Journal*—are all relatively critical of Democrats but give Republicans virtually the same treatment as a centrist outlet like *USA Today*. Finally, with the exception of political scandals, outlets do not disproportionately run stories that favor one party; and even for scandals, the relationship is weak outside the overtly partisan blogs. For example, the correlation in topic coverage between *Fox News* and the *New York Times* is 0.83.

Past empirical work on quantifying media bias can be broadly divided into two approaches: audience-based and content-based methods. Audience-based approaches are premised on the idea that consumers patronize the news outlet that is closest to their

[3]While most past automated methods have classified the *Wall Street Journal* as liberal-leaning, our approach positions it on the conservative side of the spectrum, in line with conventional wisdom.

[4]The gap among opinion articles is, as expected, larger: 39% and 6% net left-leaning, respectively.

ideological ideal point (as in Mullainathan and Shleifer (2005)), implying the political attitudes of an outlet's audience are indicative of the outlet's ideology. This approach has produced sensible ideological orderings of outlets (Bakshy et al., 2015; Gentzkow and Shapiro, 2011; Zhou et al., 2011). These methods, however, provide only relative, not absolute, measures of slant because small absolute differences between outlets could lead to substantial audience fragmentation along party lines.

Addressing this limitation, content-based methods, as the name implies, quantify media bias directly in terms of published content. For example, Gentzkow and Shapiro (2010) algorithmically parse congressional speeches to select phrases that are indicative of the speaker's political party (e.g., "death tax"), and then measure the frequency of these partisan phrases in a news outlet's corpus. Similarly, Groseclose and Milyo (2005) compare the number of times a news outlet cites various policy groups with the corresponding frequency among congresspeople of known ideological leaning. Ho et al. (2008) use positions taken on Supreme Court cases in 1,500 editorials published by various news outlets to fit an ideal point model of outlet ideological position. Using automated keyword-based searches, Puglisi and Snyder (2011) find that an outlet's coverage of political scandals systematically varies with its endorsement of electoral candidates. Finally, Baum and Groeling (2008) investigate issue filtering by tracking the publication of stories from Reuters and the Associated Press in various news outlets, where the topic and slant of the wire stories were manually annotated by 40 undergraduate students.

Collectively, these content-based studies establish a quantitative difference between news outlets, but typically focus on a select subset of articles, which limits the scope of the findings. For example, highly partisan language from congressional speeches appears in only a small minority of news stories, editorials on Supreme Court decisions are not necessarily representative of reporting generally, and political scandals are but one of many potential topics to cover. In response to these limitations, our approach synthesizes various elements of past content-based methods, combining statistical techniques with direct human judgments. This hybrid methodology allows us to directly and systematically investigate media bias at a scale and fidelity that were previously infeasible.

2 Data and Methods

Our primary analysis is based on articles published in 2013 by the top thirteen U.S. news outlets and two popular political blogs. This list includes outlets that are commonly believed to span the ideological spectrum, with the two blogs constituing the likely endpoints (*Daily Kos* on the left, and *Breitbart* on the right), and national outlets such as *USA Today* and *Yahoo News* expected to occupy the center. See Table 1 for a full list. To compile this

set of articles, we first examined the complete web browsing records for U.S.-located users who installed the Bing Toolbar, an optional add-on application for the Internet Explorer web browser. For each of the 15 news sites, we recorded all unique URLs that were visited by at least ten toolbar users, and we then crawled the news sites to obtain the full article title and text.[5] Finally, we estimated the popularity of an article by tallying the number of views by toolbar users. This process resulted in a corpus of 803,146 articles published on the 15 news sites over the course of a year, with each article annotated with its relative popularity.

2.1 Identifying political news articles

With this corpus of 803,146 articles, our first step is to separate out politically relevant stories from those that ostensibly do not reflect ideological slant (e.g., articles on weather, sports, and celebrity gossip). To do so, we built two binary classifiers using large-scale logistic regression. The first classifier—which we refer to as the *news classifier*—identifies "news" articles (i.e., articles that would typically appear in the front-section of a traditional newspaper). The second classifier—the *politics classifier*—identifies political news from the set of articles identified as news by the first classifier. This hierarchical approach shares similarities with active learning (Settles, 2010), and is particularly useful when the target class (i.e., political news articles) comprises a small overall fraction of the articles. Given the scale of the classification tasks (described in detail below), we fit the logistic regression models with the stochastic gradient descent (SGD) algorithm (see, for example, Bottou (2010)) implemented in the open-source machine learning package Vowpal Wabbit (Langford et al., 2007).[6]

To train the classifiers, we require both article features and labels. For features, we use a subset of the words in the article, as is common in the machine learning literature. Given the standard inverted pyramid model of journalism, we start by considering each article's title and first 100 words, which are strongly indicative of its topic. We then compute the 1,000 most frequently occurring words in these snippets of article text (across all articles in our sample), excluding stop words (e.g., 'a', 'the', and 'of'). Finally, we represent each article as a 1,000-dimensional vector, where the i-th component indicates the number of times the i-th word in our list appears in the article's title and first 100 words.

The article labels for both the news and politics classifiers were collected through

[5] We estimate each article's publication date by the first time it was viewed by a user. To mitigate edge effects, we examined the set of articles viewed between December 15, 2012 and December 31, 2013, and and limit to those first viewed in 2013.

[6] In the Appendix we compare this approach to the use of support vector machines (SVM), and find nearly identical performance.

Amazon Mechanical Turk (`http://mturk.amazon.com`), a popular crowdsourcing platform. We required that workers reside in the U.S., have good Mechanical Turk standing (i.e., have completed at least 1,000 tasks on the platform and have 98% approval rate), and pass a test of political knowledge (described in the Appendix). Although the answers to the test could be found using web search, these types of screening mechanisms have nonetheless proven useful to ensure worker quality (Kittur et al., 2008).

For the news classification task, workers were presented with an article's title and first 100 words, and asked to categorize it into one of the following nine topics, roughly corresponding to the sections of a newspaper: (1) world or national news; (2) finance / business; (3) science / technology / health; (4) entertainment / lifestyle; (5) sports; (6) travel; (7) auto; (8) incoherent text / foreign language; and (9) other. We then collapsed topics (2)–(9) into a single "non-news" category, producing a binary division of the articles into "news" and "non-news". For the training set, workers categorized 10,005 randomly selected articles stratified across the 15 outlets (667 articles per outlet), with each article categorized by a single worker. Applying the trained news classifier to the full corpus of 803,146 articles, 340,191 (42%) were classified as news.

To evaluate the news classifier, we constructed a test set by first collecting labels for an additional random set of 1,005 articles (67 per outlet), where each article was now rated by two workers to ensure accurate ground-truth categories.[7] Of these 1,005 articles, 794 (79%) were identically labeled by the two workers. On this subset of articles, we find the classifier had 82% precision, 90% recall, and 87% overall accuracy. We also evaluated the classifier on the full set of 1,005 articles by randomly selecting one of the two labels as the "ground truth," again finding that it performed well, with 74% precision, 81% recall, and 79% overall accuracy.

Starting with the 340,191 articles classified as news, we next trained the politics classifier by again asking workers to label a random subset of 10,005 articles (667 per outlet), with each article classified by a single worker. In this case, we asked workers to "identify whether the following news article is about a U.S. political issue," and we provided three options: (1) political; (2) not political; and (3) incoherent / corrupted text. We also provided a list of examples to help the workers in their decisions. On the set of 340,191 news articles, 114,814 (34%) were classified as political. Thus, 14% of the original set of 803,146 articles were identified as political news stories.

To evaluate performance of the politics classifier, 1,005 randomly selected news articles (67 per outlet) were classified as political or not by two workers, and of these, 777

[7]For cost-effectiveness, only one label per article was collected for the training set, since the supervised learning techniques we used are robust to noise. However, to accurately *evaluate* the classifiers, it is important for the test set to be free of errors, and we thus collect two labels per article.

Outlet	Average number of "news" articles per day	Average number of "political news" articles per day
BBC News	72.8	4.3 (6%)
Chicago Tribune	16.0	3.8 (24%)
CNN News	100.1	29.1 (29%)
Fox News	95.9	44.2 (46%)
Huffington Post	118.7	44.8 (38%)
Los Angeles Times	32.5	9.1 (28%)
NBC News	52.6	14.6 (28%)
New York Times	68.7	24.7 (36%)
Reuters	30.3	10.8 (36%)
The Washington Post	65.9	37.9 (58%)
USA Today	33.7	11.8 (35%)
Wall Street Journal	11.7	4.6 (39%)
Yahoo News	173.0	53.9 (31%)
Breitbart News Network	15.1	11.2 (74%)
Daily Kos	14.0	9.8 (70%)

Table 1: Average number of daily "news" and "political news" stories identified in our sample for each outlet, with the percent of news stories that are political in parentheses.

(77%) had concordant labels. On this test set of 777 articles, the politics classifier had 91% precision, 81% recall, and 87% overall accuracy. As before we further evaluated the classifier on the full set of 1,005 news articles by randomly selecting one of the two labels as the official ground truth; on this full set, the classifier had 84% precision, 69% recall, and 78% accuracy.

Overall, both the news and politics classifiers performed well, yielding results inline with recent work (Flaxman et al., 2015). Moreover, as shown in the Appendix, we find almost identical results when we use support vector machines (SVMs) (Cortes and Vapnik, 1995) to classify the articles instead of logistic regression. Finally, we note that there may be genuine disagreement about an article's classification. For example, some may consider a story about President Obama's vacation plans political news, while others may classify it as a travel or lifestyle piece. Thus, at least part of the differences we observe between the algorithmic and human labels can be attributed to the ambiguity inherent in the classification task. A similar pattern has been observed in related work on linguistic annotation (Plank et al., 2014) where the authors show that disagreement between annotators reveal debatable linguistic cases and therefore should be embraced as opposed to eliminated.

Table 1 lists the average daily number of "news" and "political news" articles identified in our sample. Notably, there is substantial variation in the number of articles across

outlets. In part, this variation is due to real differences in the number of published articles—*Yahoo News* and *CNN*, for example, do indeed publish more news stories than the niche blogs *Daily Kos* and the *Breitbart News Network*. Some of the variation, however, is due to the fact that we examine only articles that were visited by at least ten toolbar users. We thus obtain lower coverage for smaller outlets (e.g., the *Chicago Tribune*), and those with a paywall (e.g., the *Wall Street Journal*). As described below, we conduct our analysis on a popularity-weighted sample of articles, and since these popular articles are likely represented in our sample, we do not expect the differences in coverage to qualitatively affect our results.

2.2 Identifying article topics and measuring ideological slant

Having identified approximately 115,000 political news articles, we next seek to categorize the articles by topic (e.g., gay rights, healthcare, etc.), and to quantify the political slant of the article. To do so, we again turn to human judges recruited via Mechanical Turk to analyze the articles. Even with crowdsourcing, however, classifying over 100,000 articles is a daunting task. We thus limit ourselves to a readership-weighted sample of approximately 11,000 political news articles. Specifically, for every day in 2013, we randomly selected two political articles from each of the 15 outlets we study, with sampling weights equal to the number of times the article was visited by our panel of toolbar users. We note that while we consider only a fraction of the political news articles in our corpus, our crowdsourcing approach allows us to analyze many more articles than would be feasibly in a traditional laboratory setting.

To detect and control for possible preconceptions of an outlet's ideological slant, workers, upon first entering the experiment, were randomly assigned to either a *blinded* or *unblinded* condition. In the blinded condition, workers were only presented with the article's title and text, whereas in the unblinded condition, they were additionally shown the name of the outlet in which the article was published. Each article was then analyzed by two workers, one each from the sets of workers in the two conditions.

For each article, each worker completed the following three tasks. First, they provided primary and secondary article classifications from a list of 15 topics: (1) civil rights; (2) Democrat scandals; (3) drugs; (4) economy; (5) education; (6) elections; (7) environment; (8) gay rights; (9) gun related crimes; (10) gun rights/regulation; (11) healthcare; (12) international news; (13) national security; (14) Republican scandals; and (15) other. We manually generated this list of topics with the aid of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a popular technique for exploring large text corpuses.[8] Only

[8]Though LDA was helpful for exploring the corpus and generating the list of topics, it did not produce

12% of all political articles received a primary classification of "other", suggesting our list of topics was sufficiently comprehensive. Second, workers determined whether the article was descriptive news or opinion. Third, to measure ideological slant, workers were asked, "Is the article generally positive, neutral, or negative towards members of the Democratic party?", and separately, "Is the article generally positive, neutral, or negative towards members of the Republican party?". Choices for these last two questions were provided on a 5-point scale: very positive, somewhat positive, neutral, somewhat negative, and very negative. To mitigate question ordering effects (Schuman and Presser, 1996), workers were initially randomly assigned to being asked either the Democratic or Republican party question first; the question order remained the same for any subsequent articles the worker rated.

Finally, we assigned each article a partisanship score between -1 and 1, where a negative rating indicates the article is net left-leaning and a positive rating indicates it is net right-leaning. Specifically, for an article's depiction of the Democratic party, the 5-point scale from very positive to very negative is encoded as: -1, -0.5, 0, 0.5, 1. Analogously, for an article's depiction of the Republican party, the scale is encoded as 1, 0.5, 0, -.0.5, -1. The score for each article is defined as the average over these two ratings. Thus, an average score of -1, for example, indicates that the article is very positive toward Democrats *and* very negative toward Republicans. The result of this procedure is a large, representative sample of political news articles, with direct human judgments on partisanship and article topic.

Whereas past work has relied on undergraduate student judges to evaluate media bias (Baum and Groeling, 2008; Ho et al., 2008), ours is the first to use crowdsourcing. This approach facilitates far greater scale and diversity of workers, but also raises concerns regarding data quality (Berinsky et al., 2012). For instance, the small partisan differences we observe across outlets (discussed below) could simply reflect limited political awareness of workers. With these concerns in mind, we took several steps (described in more detail in the Appendix), consistent with established best practices (Mason and Suri, 2012), to ensure high quality ratings. First, we restricted participation to U.S.-located workers with an exceptional track record on the crowdsourcing platform. Second, we required workers to pass a screening test. Third, in a preliminary analysis multiple workers were assigned to the same article; we found inter-rater reliability was on par with previous studies, even if we consider only those articles rated to have a political leaning (i.e., excluding "neutral" articles). Fourth, we limited the number of articles a single worker could rate to 100, ensuring a large pool of independent evaluations. Finally, as noted above, we only

article classifications that were sufficiently accurate for our purposes; we thus relied on human labels for the final article-level topic classification.

presented the name of the publication venue to a randomly selected subset of the workers so as to check whether their perceptions of an outlet's ideological leaning affected their ratings. We found that article ratings were similar regardless of whether the outlet name was listed. Nevertheless, to be cautious, we limit our primary analysis to ratings generated by workers who did not see the outlet source.

We additionally conducted ex-post checks to validate the quality of the article slant ratings. The median subject spent over two minutes reading and rating each article, in line with expectations. The ratings were uncorrelated with a worker's stated political affiliation and only weakly related to a worker's intensity of news consumption. Inter-rater reliability—computed by comparing labels when the source was revealed versus blinded—is 81%, consistent with our preliminary analysis and with previous studies (Baum and Groeling, 2008). We further note that this number should be considered a lower bound on agreement, since some differences could be due to the impact of revealing the source. Detailed statistics on inter-rater reliability can be found in the Appendix. The totality of evidence thus suggests that our workers produced high-quality article ratings. This finding is consistent with the growing literature demonstrating that crowd workers reliably replicate the behavior of undergraduate students across a wide variety of behavioral experiments (Berinsky et al., 2012; Buhrmester et al., 2011; Goodman et al., 2013; Mason and Suri, 2012; Paolacci et al., 2010), and produce verifiably high-quality work in labeling tasks (Callison-Burch, 2009; Sorokin and Forsyth, 2008).

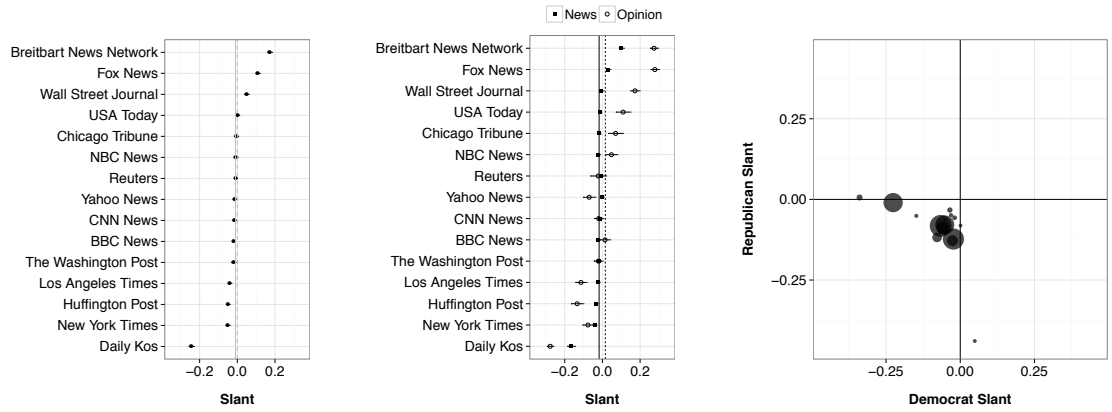
3 Results

3.1 Outlet-level slant

We start by providing outlet-level estimates of ideological position. As described above, each article is first assigned a partisanship score between -1 and 1, with negative values indicating a net left-leaning article and positive values indicating a net right-leaning article. For each outlet, we then average the scores of the corresponding articles in our sample. Thus, since articles were randomly sampled in proportion to their popularity, an outlet's score is the average, popularity-weighted slant of articles in that outlet.

As can be seen in Figure 1(a), these ideological scores result in an ordering of outlets that is largely in line with past research.[9] For example, the *Breitbart News Network* is the most right-leaning outlet, and the *Daily Kos* is the most left-leaning outlet in our set. However, though the rank ordering mirrors past work, the magnitude of the observed

[9]One difference is that we identify the *Wall Street Journal* as a relatively conservative outlet—in accordance with convention wisdom—while past automated methods based on audience and co-citation measures have characterized it as left-leaning (Flaxman et al., 2015; Groseclose and Milyo, 2005).



(a) (b) (c)

Figure 1: **1(a):** Overall outlet-level slant; **1(b):** Outlet-level slant estimated separately for opinion and descriptive news. **1(c):** Democrat and Republican slant for each news outlet, where point sizes indicate the relative popularity of the outlet.

differences between the mainstream news outlets is remarkably small. For example, the *New York Times* and *Fox News*—which are the most ideologically distant mainstream outlets in our sample—have slant coefficients that differ by only 0.16 points (-0.05 vs. 0.11). To put these numbers in perspective, we recall that the distance between each category in our five-point ideology scale (e.g., the distance between neutral and somewhat positive for Democrats) is 0.5. The two political blogs exhibit much larger differences (-0.24 vs. 0.17), both from each other and from the mainstream media.

The outlet-level partisanship score is based both on descriptive news and opinion articles. Concerns over partisan media, however, largely stem from worries that *descriptive news* is ideologically biased, since such coverage is not necessarily interpreted by readers as representing only a single individual's perspective. To investigate this issue, we next examine outlet-level partisanship scores separately for opinion pieces and descriptive reporting. As expected, Figure 1(b) shows that partisanship is much more extreme for opinion than for descriptive news. For example, opinion stories on Fox News have a slant of 0.28, compared to 0.03 for descriptive news stories. Notably, even though descriptive news stories are largely neutral across the outlets, the differences still produce an ideological ranking of the outlets that is approximately the same as when we include opinion stories.[10] This finding indicates that ideological slant, while small in an absolute sense, is indeed present in descriptive reporting, and is directionally consistent with conventional

[10]There are exceptions: for instance, the*Huffington Post* is more left-leaning on descriptive news than the *New York Times*.

wisdom.

Why is it that the partisan differences we find are so small? Figure 1(c) in part answers this question by splitting outlet-level slant into its two constituent pieces: Democratic and Republican slant. That is, we look at how, on average, outlets portray Democrats, and separately, how they portray Republicans. Strikingly, nearly all the outlets (with the exception of *Daily Kos* and *Breitbart News Network*), lie in the lower-left quadrant, meaning that on average, they portray both Democrats and Republicans negatively. While one might have expected that net left-leaning or net right-leaning outlets would favorably portray one party while unfavorably characterizing the other, what we find is quite different. An outlet's net ideological leaning is identified by the extent of its criticism, rather than its support, of each party. In particular, net conservative outlets treat Republicans about the same way as centrist outlets, but are much more critical of Democrats. Analogously, net liberal outlets are more critical of Republicans but treat Democrats quite similarly compared to centrist outlets. This apparently widespread reporting practice of critical rather than supportive coverage in turn limits the ideological differences between outlets.[11]

Two additional contributing factors for the similar outlet-level slants that we observe are: (a) the vast majority of political articles in most outlets are neutral; and (b) among the partisan pieces, ideologically opposed articles are present in a single outlet. In Figure 2, we plot the fraction of descriptive articles in each outlet that are net-left (score < 0) and net-right (score > 0), with the remaining fraction consisting of articles rated as neutral. We find that in most mainstream outlets, about one-third of descriptive news articles show a partisan leaning, and among these, slightly more than half are net-left, with the exact ratio varying with the outlet's overall ideology. For example, in the *New York Times*, about 20% of articles are net-left, 10% are net-right, and 70% are neutral. This similarity among outlets persists even when we restrict to more clearly partisan articles. In particular, Figure A4 (in the Appendix) shows that about 1% of descriptive articles in most mainstream outlets have a slant score above 0.5 (indicating the article is quite right-leaning), and about 1% have a score below -0.5 (indicating the article is quite left-leaning). Among opinion articles, there are, as expected, typically far more partisan pieces; and among these partisan articles, the balance of net-left and net-right coverage generally reflects an outlet's overall ideology. For example, opinion pieces in *Fox News* are 63% net-right, 6% net-left, and 31% neutral. Since the outlets in Figure 2 are ordered left-to-right by ideological position, this relationship is revealed by the downwards sloping net-left line and upward sloping net-right line.

[11] In the Appendix, we look at treatment of Democrats and Republican by issue, and similarly find that on most issues, both parties are portrayed slightly negatively on average.

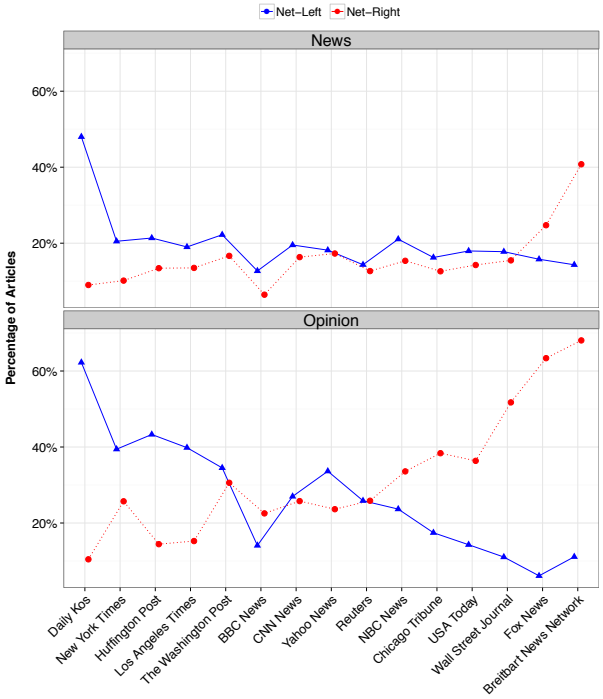


Figure 2: Article-level news and opinion ideological position, by outlet.

3.2 Issue framing

A key strength of our approach is that we not only can assess an outlet's overall slant, but we can also evaluate bias on an issue-by-issue basis. Figure 3 compares the ideological slant of the *New York Times* to *Fox News* for each of the 14 topics we consider. The issues are ordered top-to-bottom from largest to smallest differences in slant between the two outlets—thus issues at the top of the list can be thought of as the most polarizing. The points sizes reflect the coverage intensity in the corresponding outlet. The plot illustrates three high-level points. First, *Fox News* is consistently to the right of the *New York Times* on every issue we examined. Second, for many issues, the differences are remarkably small. For civil rights, for example, the net slant for the *New York Times* is -0.01, compared to 0.07 for *Fox News*. Finally, even for topics where there are relatively large differences between the two outlets, their slants are related. For example, in both outlets, Republican scandals have the most left-leaning coverage, and analogously, Democratic scandals have the most right-leaning coverage. This last observation further explains the relatively small overall differences between the outlets: many issues (e.g., scandals) are inherently left- or right-leaning, and thus mitigate the potential for bias; it would be difficult, for example, for the *New York Times* to write a net-left article about a scandal perpetrated by Democrats.

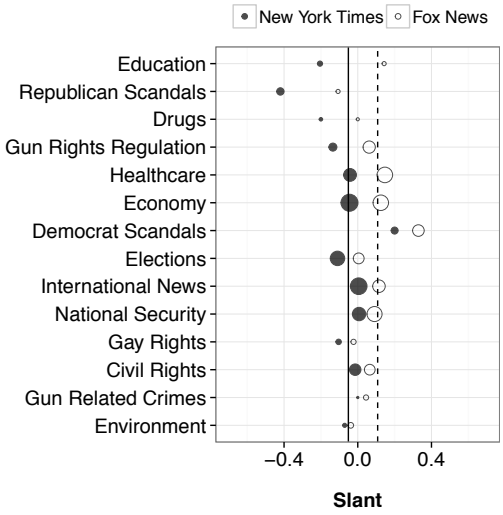


Figure 3: Comparison of issue-level slant of the *New York Times* to *Fox News*. Point sizes indicate coverage intensity, and vertical lines give outlet averages.

Figure 4 generalizes these findings to the 15 outlets we study. Outlets are ordered on the x-axis from left to right based on overall outlet-level slant; each line corresponds to an issue, colored according to its mean slant across the outlets, and the y-axis indicates the average slant of articles on that topic in each outlet. As noted above, across outlets, Democrat and Republican scandals are among the few issues exhibiting large partisan slant. Moreover, on all the issues, *Fox News* and the two political blogs—*Daily Kos* and the *Breitbart News Network*—and consistently more partisan than the other outlets. For the remaining issues and outlets, the ideological differences are quite small and do not appear to vary systematically.

3.3 Issue filtering

We next examine the extent to which news outlets selectively report on topics (i.e., issue filtering). Such potential issue filtering is consequential for at least two reasons. First, by selectively reporting on partisan topics (e.g., scandals), issue filtering can amplify an outlet's overall ideological slant. Second, even for issues that are reported in a largely non-partisan manner, selective coverage may leave readers of different outlets with materially different exposure to political issues.

To gauge filtering effects, for each outlet we first estimate the proportion of articles that were categorized (by the human judges) under each topic.[12] Figure 5(a) compares

[12]We performed this analysis using both the primary and the secondary topics; as a robustness check,

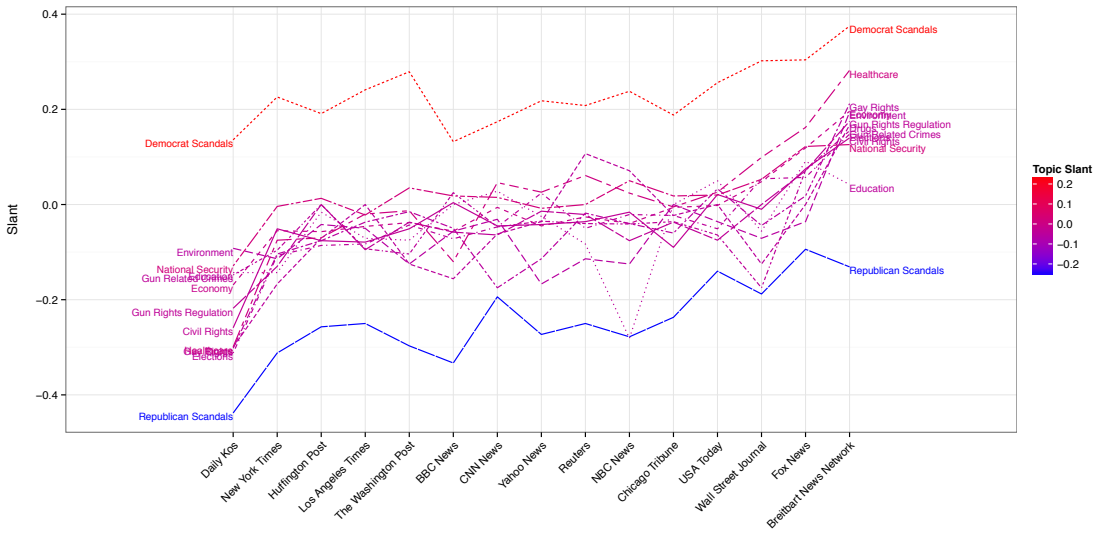


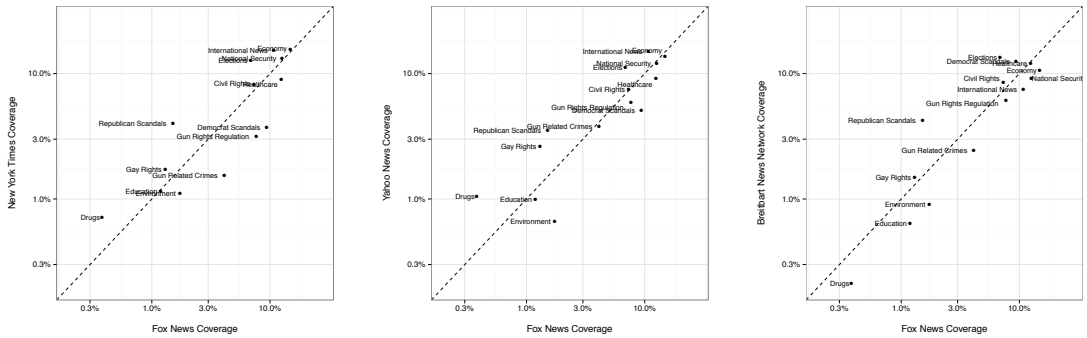
Figure 4: Issue specific slant. Outlets are ordered left-to-right by their overall ideological position and issues are colored blue-to-red according to their slant. The y-axis gives the average relative Republican slant for a particular domain on a specific issue.

the distribution of topics in *Fox News* to that in the *New York Times*, where points lying on the dashed diagonal line indicate equal coverage. Perhaps surprisingly, the plot shows that most topics receive similar coverage in the two outlets. Moreover, this similarity is not solely restricted to the popular topics—such as the economy, international news, and elections—but also carries over to more niche areas, including civil rights, gay rights, and education. Overall, the correlation in topic coverage between the *New York Times* and *Fox News* is 0.83. As another point of comparison, Figures 5(b) and 5(c) contrast *Fox News* to the centrist *Yahoo! News* and to the right-wing blog *Breitbart News Network*. We again find, strikingly, that the distribution of topics is remarkably similar, despite their ideological differences. One exception is coverage of scandals. For example, Democrat scandals make up only 4% of political articles in the *New York Times*, while they account for almost 10% of those on *Fox News*. Similarly, Republican scandals make up 4% of all political articles in the *New York Times* and account for just 1.5% in *Fox News*.

Figure 6 extends these results to the full set of outlets.[13] Outlets are ordered left to right based on their overall ideological slant. Each line corresponds to a particular topic, and is colored according to the average ideological slant of outlets that cover that topic:

we used only the primary topic and did not find any qualitative differences. We used the labels gathered from both the blinded and unblinded groups, since seeing where the article was published should have little to no effect on the topic identified by participants. Moreover, we exclude articles labeled as "other".

[13]For ease of viewing, we remove *BBC News* and the international news topic. That *BBC News* emphasizes international news is orthogonal to the question we address here, but its inclusion in the plot would have substantially changed the scale of the y-axis.



(a) *Fox News* vs. *New York Times* (b) *Fox News* vs. *Yahoo News* (c) *Fox News* vs. *Breitbart*

Figure 5: Comparison of *Fox News* topic coverage with *New York Times*, *Yahoo News*, and *Breitbart News Network*

the more blue the line is, the more it is covered by left-leaning outlets, and the more red it is, the more it is covered by right-leaning outlets. Since the lines in Figure 6 are largely flat across the outlets, there appears to be little systematic issue filtering in the U.S. news media.

As a more quantitative test of this visually suggestive result, for each topic we separately fit the following regression model:

$$C_i = \beta_0 + \beta_1 I_i + \epsilon_i \tag{1}$$

where C_i is the coverage of the topic in outlet i, and I_i is the outlet's overall ideological slant. The estimated coefficient β_1 thus captures how coverage relates to the position of the outlet on the political spectrum. We find β_1 is uniformly small across the topics we study, and is in fact statistically significant for only two issues: Republican scandals ($p = 0.018$) and Democrat scandals ($p = 0.0006$). We note that previous work identified similar coverage differences for scandals (Baum and Groeling, 2008; Gentzkow and Shapiro, 2011); our results show that selective coverage of scandals, while consequential, are not representative of issue filtering more broadly.

3.4 Consumption vs. production choices

We have throughout considered a popularity-weighted sample of articles in order to study what a typical reader of each news outlet might consume. For example, our primary analysis aims to measure the differences—in terms of coverage and slant—between articles read by those who frequent *The New York Times* and those who frequent *Fox News*. These consumption choices are ostensibly driven by the availability and promotion of content

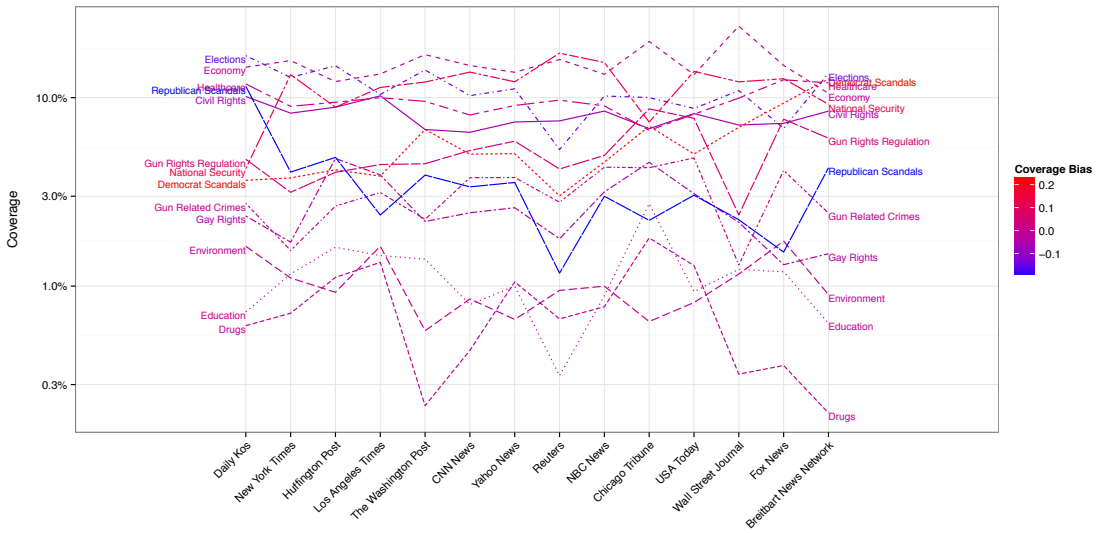


Figure 6: Issue coverage by outlet. Issue lines are colored blue-to-red to reflect the degree to which the issue-level average is is relatively pro-Republican.

by the news publishers: the headline story is likely to attract considerable attention, regardless of its topic or ideological slant. One might, however, reasonably seek to isolate the inherent *production* bias of an outlet, absent the consumption choices of its readers.

Given that online outlets frequently update which content to promote based on continuous popularity metrics, the distinction between production and consumption is not clear-cut. Nevertheless, here we report the results of two natural variations on our analysis. First, we estimate outlet-level slants for the full corpus of articles appearing in each outlet; second, we compute these metrics for articles appearing on an outlet's home page, the online equivalent of the traditional front page of a newspaper.[14] For each of the 15 news sites, the three outlet-level estimates (popularity weighted, full corpus, and homepage articles) are nearly identical (Figure 7). This pattern is indicative of the close relationship between production and consumption decisions, and illustrates the robustness of our results to the specifics of the measurement procedure. At the extremes of the ideological spectrum (e.g., *Daily Kos* and *Breitbart News Network*), popular articles do appear to be more partisan than those across an outlet's entire corpus or articles promoted on the home page, though the effect is small. In the Appendix, we further investigate this relationship between article slant and popularity with article-level regression models, and confirm there is a statistically significant but substantively small effect.

[14]Our analysis is specifically based on articles that garnered at least 10 views by our sample of users. Due to data retention policies, we could determine if an article appeared on an outlet's home page only for the last 6 months of our sample period, which accounts for the slight shuffling in the ordering of outlets. Details are provided in the Appendix.

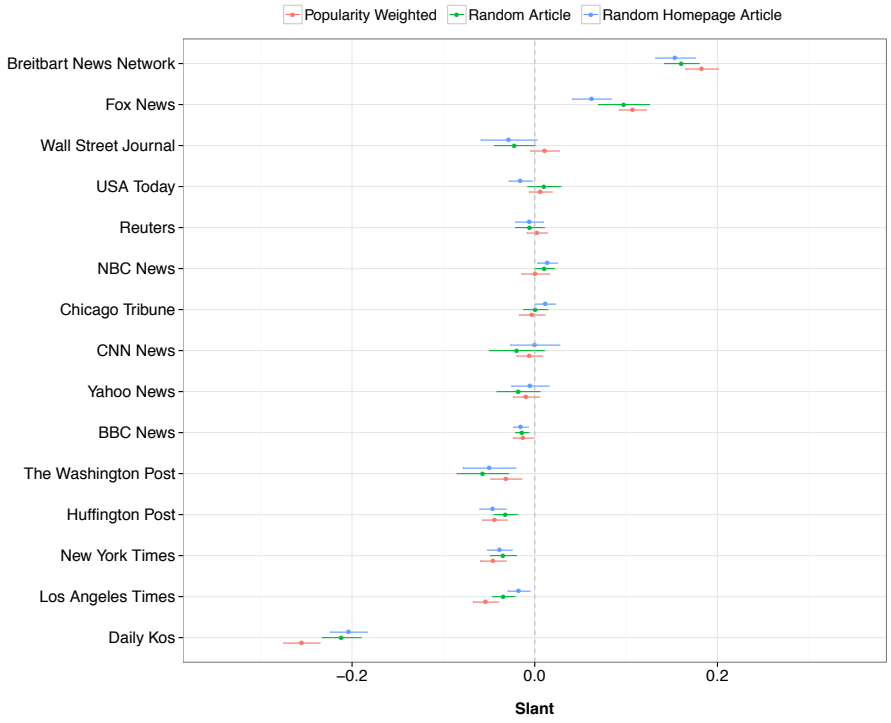


Figure 7: Overall outlet-level slant using different sample weighting procedures: readership-weighted, uniform weights and uniform weights for those articles appearing on the outlet's homepage. Whiskers give ±1 standard errors.

4 Discussion and Conclusion

Both in terms of coverage and slant, we find that the major online news outlets—ranging from *The New York Times* on the left to *Fox News* on the right—have surprisingly similar, and largely neutral, descriptive reporting of U.S. politics. This result stands in contrast to what one might reasonably conclude from past academic studies, and from the regular laments of popular commentators on the rise of partisan media. For example, by analyzing the think tanks news outlets cite, Groseclose and Milyo (2005) conclude there is a "strong liberal bias" in the U.S. media. Further, given the ideological fragmentation of media audiences (Iyengar and Hahn, 2009), theoretical and empirical work in economics predicts news outlets would present biased perspectives in an effort to cater to their readers (Gentzkow and Shapiro, 2006, 2010; Mullainathan and Shleifer, 2005). Indeed, Gentzkow and Shapiro (2010) find that an outlet's use of highly partisan language (e.g., "death tax" instead of "estate tax") is strongly correlated with popular perceptions of its political leanings.

How, then, can we reconcile our results with the prevailing conventional wisdom? In part, the differences stem from the difficulty of directly measuring the ideological slant of articles at scale. For example, most articles do not cite policy groups nor do they use highly partisan language, and those that do are not necessarily representative of political reporting more generally. In contrast, our combination of machine learning and crowdsourcing techniques does appear to yield accurate, article-level assessments. Moreover, despite an increasingly polarized American public (Pew Research Center, 2014), a substantial fraction of news consumers (49%) still prefer sources that do not have a systematic ideological leaning—and the proportion is particularly high among those who get their news online (Purcell et al., 2010)—tempering partisan pressures. Finally, the traditional news-desk / editorial divide (Machin and Niblock, 2006) may yet encourage publishers to maintain a degree of non-partisanship in their descriptive reporting, reserving their opinion pages to show their ideological stripes and appeal to their audience base.

It may be tempting to cheer the relative lack of overt partisanship in the descriptive political reporting of most major American news outlets. Our analysis, however, also reveals some hurdles for robust political discourse. First, given the ideological distance between Democrats and Republicans—by some measures the largest of the modern political era—balanced coverage in the point / counter-point paradigm (D'Alessio and Allen, 2000) may not optimally inform voters about the issues. One might reasonably expect the facts to favor one party over the other—at least on some of the issues—and thus largely non-partisan reporting may not accurately reflect the substantive differences between the political parties. (Coverage of political scandals is a notable exception to this

convention of non-partisanship; for example, scandals involving Democrats unsurprisingly portray Democrats more harshly than Republicans.) Second, both for Democrats and Republicans, we find news outlets are almost universally critical rather than supportive, a practice some have called "gotcha journalism." For example, as many political commentators have observed, the failures of the Affordable Care Act received far more media attention than its successes. This tendency to print predominantly critical news may stem from publishers' desires to appear non-partisan by avoiding apparent advocacy, or from readers' appetites for negative coverage, or from a combination of the two. Regardless of the rationale, predominantly critical coverage likely masks relevant facts and may hinder readers from developing informed opinions. Finally, though the relative uniformity we observe across news outlets provides readers with a common base of knowledge, it also limits the diversity of available perspectives.

We conclude by noting three important limitations of our study. First, we have characterized ideological slant by assessing whether an article is generally positive, neutral, or negative towards members of the Democratic and Republican parties. This codification has the advantage of side-stepping the tricky—and perhaps impossible—task of assessing bias against an objective "truth." However, it is certainly possible for an article not to explicitly favor a political party while still advocating a generally liberal or conservative position. Second, we considered a relatively short, 12-month timespan that did not coincide with a presidential or midterm election. While several hot-button issues attracted substantial media attention during this stretch—such as healthcare reform and marriage equality—other periods may exhibit more partisan dissent, which could in turn amplify differences between outlets. Third, we do not consider the effects of media bias on readers' attitudes or actions, such as voting, volunteering, and donating. It could be the case, for example, that even though we find outlets have ideologically similar coverage overall, a single partisan article could have an out-sized effect on readers. Despite these limitations, we believe our study is a natural starting point for investigating media bias at scale, and we hope the approach we have taken will benefit future exploration of such issues.

References

Akkaya, C., Conrad, A., Wiebe, J., and Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 195–203, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*.

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

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

D'Alessio, D. and Allen, M. (2000). Media bias in presidential elections: a meta-analysis. *Journal of communication*, 50(4):133–156.

Eveland, W. P. and Shah, D. V. (2003). The impact of individual and interpersonal factors on perceived news media bias. *Political Psychology*, 24(1):101–117.

Flaxman, S., Goel, S., and Rao, J. M. (2015). Ideological segregation and the effects of social media on news consumption. *Available at SSRN*.

Gamson, W. A. (1992). *Talking politics*. Cambridge University Press New York.

Gamson, W. A. and Lasch, K. E. (1981). The political culture of social welfare policy. *University of Michigan CRSO Working Paper No. 242*.

Gamson, W. A. and Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, pages 1–37.

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.

Goodman, J. K., Cryder, C. E., and Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224.

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

Ho, D. E., Quinn, K. M., et al. (2008). Measuring explicit political positions of media. *Quarterly Journal of Political Science*, 3(4):353–377.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Iyengar, S. (1994). *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

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.

Iyengar, S. and Kinder, D. R. (2010). *News that matters: Television and American opinion*. University of Chicago Press.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

Krosnick, J. A. and Kinder, D. R. (1990). Altering the foundations of support for the president through priming. *The American Political Science Review*, pages 497–512.

Langford, J., Li, L., and Strehl, A. (2007). Vowpal wabbit online learning project.

Machin, D. and Niblock, S. (2006). News production. *Theory and Practice. London. Routledge*.

Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

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

Nelson, T. E., Clawson, R. A., and Oxley, Z. M. (1997a). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, pages 567–583.

Nelson, T. E. and Kinder, D. R. (1996). Issue frames and group-centrism in american public opinion. *The Journal of Politics*, 58(04):1055–1078.

Nelson, T. E., Oxley, Z. M., and Clawson, R. A. (1997b). Toward a psychology of framing effects. *Political behavior*, 19(3):221–246.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Patterson, T. E. (1993). *Out of Order*. New York:Knopf.

Perloff, R. M. (1989). Ego-involvement and the third person effect of televised news coverage. *Communication Research*, 16(2):236–262.

Pew Research Center (2012). In changing news landscape, even television is vulnerable: Trends in news consumption: 1991-2012.

Pew Research Center (2014). Political polarization in the american public.

Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 507–511.

Puglisi, R. and Snyder, J. M. (2011). Newspaper coverage of political scandals. *The Journal of Politics*, 73(03):931–950.

Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., and Olmstead, K. (2010). Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21.

Schuman, H. and Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66.

Sorokin, A. and Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820.

Sutter, D. (2001). Can the media be so liberal? the economics of media bias. *Cato Journal*, 20:431–451.

Vallone, R. P., Ross, L., and Lepper, M. R. (1985). The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of personality and social psychology*, 49(3):577.

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

A Appendix

A.1 Article classification with support vector machines

In our primary analysis, we used logistic regression (fit via stochastic gradient descent) to classify news articles, and separately, to classify political articles. Here we compare this approach to an alternative, popular classification method, support vector machines (SVMs) (Cortes and Vapnik, 1995). The results below show that the two approaches yield nearly identical performance, both on the subset of test articles where the two human judges agreed (Table A1), and on the full set of test articles, where for each article one of the two raters was chosen at random as the providing the "ground truth" (Table A2).

Table A1: Performance of the news and politics classifiers on the subset of articles for which the two human raters provided identical labels.

	News			Politics		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVM	0.88	0.84	0.88	0.87	0.92	0.80
Logistic Regression	0.87	0.82	0.90	0.87	0.91	0.81

Table A2: Performance of the news and politics classifiers on the full set of articles, where the ground truth label for each article was chosen at random from the two provided by the human raters.

	News			Politics		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVM	0.79	0.76	0.78	0.78	0.85	0.67
Logistic Regression	0.79	0.74	0.81	0.78	0.84	0.69

A.2 Experiment protocol and worker demographics

In total, 749 online crowd workers were recruited via Amazon Mechanical Turk to identify the topic and slant for 10,950 political articles. Workers were paid ten cents for each article they reviewed. Upon entering the experiment, workers were randomly assigned to either a *blinded* or *unblinded* condition, determining whether or not they were shown the name of the outlet in which an article was published. Each article was rated by one worker from each condition, and each worker could only rate up to 100 articles so as to limit the impact of any one worker's idiosyncratic judgements on the overall results.

To ensure high-quality ratings, we required that workers: (1) reside in the U.S.; (2) had

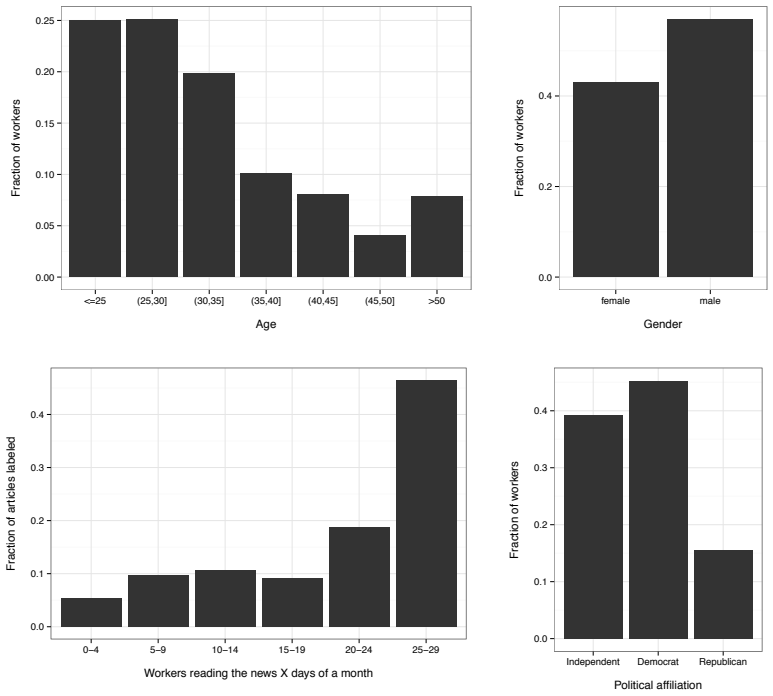


Figure A1: Demographic characteristics of Mechanical Turk workers participating in our experiments.

successfully completed at least 1,000 Mechanical Turk "human intelligence tasks" (HITs); (3) had an approval rate of at least 98%; and (4) correctly answered the following three multiple choice questions:

1. Who is the U.S. Senate Majority Leader [in Fall 2013]? (Harry Reid)
2. Which Amendment protects the right to keep and bear arms? (Second Amendment)
3. Who was the U.S. Secretary of State in 2012? (Hillary Clinton)

If workers failed to correctly answer all three questions on their first try, they could retake the test after one hour. Use of qualifications tests is a common and effective practice on Mechanical Turk to remove spammers and low quality workers, even when the test answers are easily looked up online (Akkaya et al., 2010; Buhrmester et al., 2011; Kittur et al., 2008; Mason and Suri, 2012).

After workers passed the qualification test, they were required to provide the following demographic information before evaluating articles: gender, age, the highest degree or level of school completed, political affiliation (Democrat, Republican or independent), and frequency of news consumption (i.e., number of days on which the individual reads

the news in an average month). The demographic distribution of the workers is given in Figure A1. Though the workers constitute a relatively diverse group—and are highly active news readers—they are clearly not representative of the general population. Below, we investigate the extent to which sample composition affects our results, and find that our conclusions are largely robust to selection issues.

A.3 Assessing the quality of article labels

The strength of our findings rests squarely on the quality of the judgements from the crowd workers. Although the use of crowdsourcing is relatively new in the media bias literature, it has been used extensively for other tasks with similarly high intellectual demands. Nevertheless, to ensure that our conclusions are robust to the specific nature of our study population, and to the particulars of our experimental design, we conducted a series supplementary analyses, which we now describe.

Inter-rater reliability. In a preliminary experiment, a random sample of 20 articles was labeled by four workers. The results indicate high agreement: on average the slant reported varied by less than one point (0.8 to be precise) on the five-point scale, and only in 3% of cases did the raters disagree in slant directionality. Topic agreement was also high: in 53% of cases the raters agreed on the article's primary topic, and in 65% of cases the raters agreed on at least one of the two topics they listed. Given the large number of possible choices (14 topics), this percentage indicates high inter-rater reliability.

An additional check of inter-rater reliability comes from the fact the each article was rated two times, once from a subject in the unblinded group, where the article source was revealed, and once from a subject in the blinded group, where the source was not shown. Of course, some differences in slant could be due to the impact of revealing the source, so we view this comparison as providing only a lower bound on agreement. For topic, agreement is 65%. For slant, we code each score as left, right or neutral and find an agreement of 81%. The majority of disagreement occurred when one rater found the article neutral and the other labeled it as either net-left or net-right. These figures are largely consistent with the results of the small, 20 article preliminary analysis. Further, given there is some genuine ambiguity in an article's classification, these numbers are quite high and consistent with what others have achieved using traditional laboratory subjects (Baum and Groeling, 2008).

Effect of raters' news consumption. According to a Pew Research Center (2012) study, 37% of the U.S. population reads the news regularly. In comparison, over 70% of our workers read the news regularly (at least every other day). This difference is

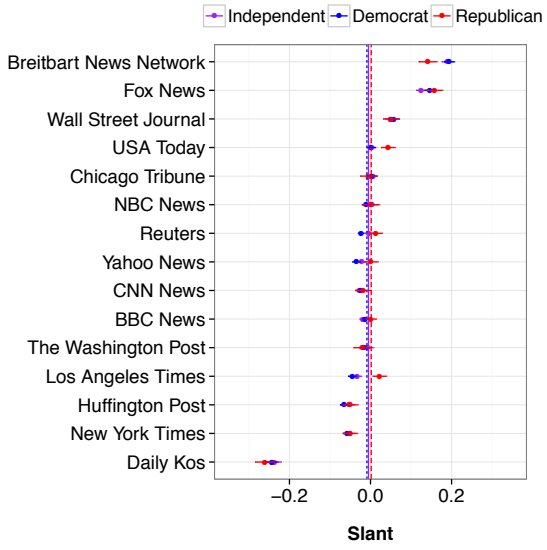


Figure A2: Outlet-level ratings by subjects' political affiliation. The points and lines are colored blue-to-red based on reported political affiliation, the vertical lines represent averages and lie virtually right on top of each other.

likely a consequence of our political screening mechanisms, and the fact that workers on Mechanical Turk regularly access the Internet. We investigated whether the labels gathered from workers who read the news less frequently are qualitatively different from those who read the news more. In particular, we checked whether workers who read the news more are able to pick up on slant that is hidden to the workers that read the news less. To that end, we computed the correlation between news consumption and the absolute value of the slant detected by the worker. This correlation is non-zero (0.1), and is statistically significant, but it is small enough not to materially influence our results.

Effect of raters' political affiliation. Figure A1 reveals that independents are over-represented in our sample and Republicans are under-represented. This is a potential concern since research suggests that individuals who see bias in political news reporting believe that the direction of the bias is counter to their own political beliefs (Eveland and Shah, 2003; Perloff, 1989; Vallone et al., 1985). For example, in analyzing the 1982 Beirut massacre, Vallone et al. (1985) find that both pro-Arab and pro-Israeli subjects interpret the same news stories on the event as hostile to their personal opinion. Perloff (1989) uncovers a similar set of differential responses to news coverage of the war in Lebanon among Arab and Jewish subjects.

Based on these findings, we were careful in how we asked subjects to rate the ideological

position of each article. Instead of asking whether the article was *biased* we instead asked how the article portrayed each party. We expect this to reduce differences across raters as it focuses subjects on a more objective assessment. For example, a liberal might find an article that is highly critical of Republicans as "unbiased," but can nonetheless report it as portraying Republicans in a negative light. Given our framing, if there is an association between political affiliation and ratings we would expect it to take the form of a person viewing their own position as neutral. For example, we might expect Republicans to rate articles as more left-leaning than Democrats, because Republicans may plausibly view neutral articles as left-leaning since they are to the left of their own opinions. Figure A2 suggests there is no such bias in our data. Indeed, the plot shows that for almost all outlets, the slant estimated by the Republicans, Democrats and independents in our sample are virtually indistinguishable from one another.

Rater attention. Another potential concern is that raters are not carefully reading the articles and thus missing key differences. To check, we first examined how much time workers spent examining articles. On average, a worker spent 160 seconds per article (with a standard deviation of 110 seconds), substantially longer than web users typically spend reading a news article.[15] Only 34 of our 749 workers spent on average less than 50 seconds (one standard deviation below than the mean) evaluating articles. Though they constitute a small subset of the total, one might still worry that these fast workers could skew our results. To check, we recomputed slant for each news outlet after removing the labels gathered from the fast workers, and find that the results are highly correlated (0.998) with our original estimates.

Effect of revealing an article's source. Finally, we examine whether workers were using the source of an article as a heuristic to make an informed guess, rather than analyzing the article content as instructed. For example, knowing that an article was published by the *New York Times*, one might rate the article as left-leaning regardless of the actual content. To check for this potential effect, a random subset of the raters in our experiment were shown the article source, while the others were not. Figure A3 shows that the slant estimated does not differ significantly across these two groups of raters, with the exception of *Fox News*. In the case of *Fox News*, the slant shifts to the right when the outlet name is exposed to the reader. Though the effect is relatively small, we err on the side of caution and restricted our primary analysis to raters who were not shown the article source.

[15] Quantitative estimates of article read times can be found here: http://time.com/12933/what-you-think-you-know-about-the-web-is-wrong.

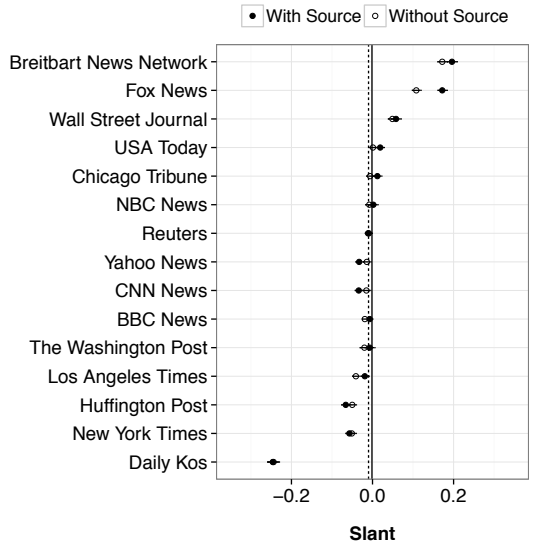


Figure A3: The solid points show reported slant when the article source was revealed, and the hollow points show the reported slant when the outlet name is not disclosed.

A.4 Coverage of highly partisan articles

In our primary analysis, for each outlet we estimated the fraction of descriptive articles that are net left-leaning (score < 0) and that are net right-leaning (score > 0), finding that most mainstream news sites covered such partisan stories in roughly equal proportion (Figure 2). Moreover, as expected, the proportion of left- and right-leaning opinion stories in each outlet was strongly associated with the outlet's overall ideological position. One might, however, reasonable wonder if these patterns persist when we consider only highly partisan articles (i.e., those with slant greater than 0.5 in absolute value). This definition requires, for example, that a net-left article be very negative towards Republicans *and* at least moderately positive towards Democrats, or very positive towards Democrats and at least moderately negative towards Republicans. Figure A4 shows the distribution of highly left- and right-leaning descriptive and opinion articles across outlets; the overall pattern is quite similar to that shown in Figure 2, indicating that our results are robust to the precise score threshold one applies.

A.5 Slant estimates for the full corpus and for homepage articles

Our main findings concern a popularity-weighted sample of articles for each news outlet. To partially disentangle production and consumption effects, in Figure 7 we additionally

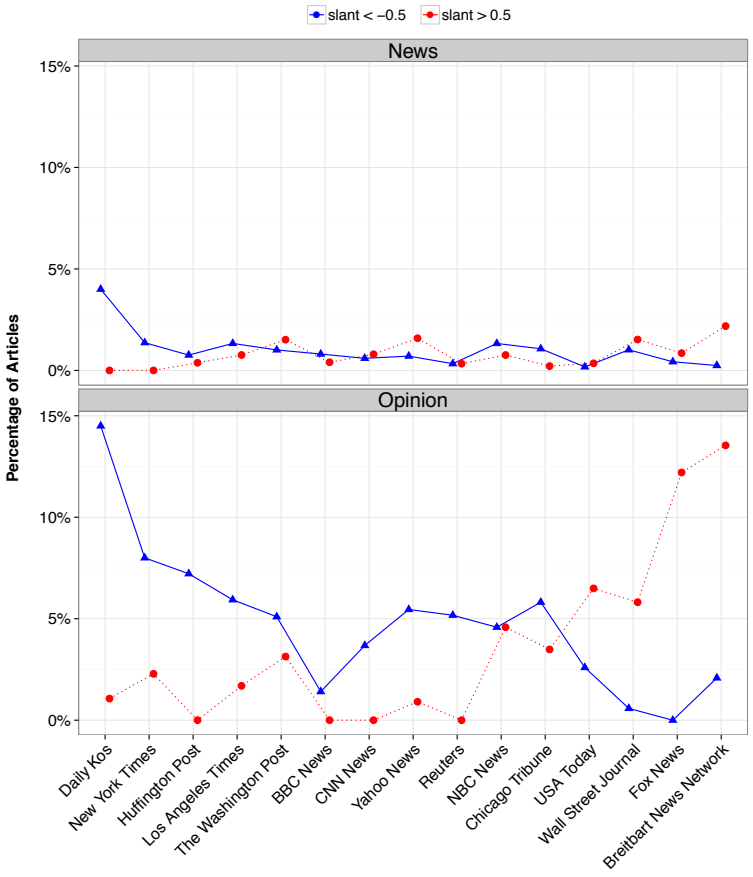


Figure A4: Fraction of articles that are highly partisan, defined as articles with a slant score greater than 0.5 in absolute value.

estimated for each news site the average (unweighted) slant of every article in our corpus from the outlet, and separately, the average (unweigted) slant of articles appearing on the outlet's homepage.[16] In both cases, we found the slant estimates were nearly identical to the original numbers.

To measure these outlet-level slants for the full corpus of articles, we apply the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) to the popularity-weighted sample of articles analyzed by the human judges. Specifically, consider a single outlet and let n denote the number of articles from the full political article corpus published in this outlet, let $A = \{a_1, a_2, ..., a_m\}$ be the set of articles included in the popularity weighted sample, let $R = \{r_1, r_2, ..., r_m\}$ be the slant ratings associated with these articles by the human judges, and let $P = \{p_1, p_2, ..., p_m\}$ denote the ex-ante probability of selecting each of these articles in the popularity weighted sample. Then an unbiased estimate $\hat{s}_{\text{full}}$ of the outlet's slant across the full corpus is given by,

$$\hat{s}_{\text{full}} = \frac{1}{n} \sum_{i=1}^{m} \frac{r_i}{p_i}. \tag{2}$$

We use an analogous estimation technique to infer, for each news site, the average slant of homepage articles. In this case, however, we first need to identify homepage articles, which we do by noting whether a reader accessed an article via a link from the site's homepage (as indicated by the URL's referrer). Outlets typically rotate the articles that appear on the homepage throughout the day, and so for simplicity we consider the set of articles that were accessed at least once from the homepage. It is difficult to definitely determine the precise location of a story on the homepage (e.g., whether or not it was the featured headline). However, in the case of *The New York Times* the position of an article is indicated in its URL, and we were thus able to confirm that the majority of homepage articles (for this particular news site) did appear in visually prominent places. Table A3 shows the fraction of our popularity-weighted sample appearing on the homepage, and as a point of comparison, the fraction of all political articles that make it to the homepage. Most articles in our sample appeared on the homepage, as did most political articles, but as expected, articles from our readership-weighted sample were more likely to appear on the homepage.

[16] Specifically, our "full corpus" consists of articles that were viewed by the toolbar users at least 10 times.

Table A3: Percentage of articles appearing on each outlet's homepage

Outlet	Popularity-weighted sample	All political articles
BBC News	0.93	0.88
Breitbart News Network	0.96	0.94
Chicago Tribune	0.79	0.79
CNN News	0.84	0.52
Daily Kos	0.95	0.92
Fox News	0.96	0.82
Huffington Post	0.93	0.77
Los Angeles Times	0.79	0.71
NBC News	0.97	0.84
New York Times	0.97	0.87
Reuters	0.71	0.66
The Washington Post	0.84	0.69
USA Today	0.89	0.80
Wall Street Journal	0.71	0.69
Yahoo News	0.92	0.75

A.6 Article-level slant vs. popularity

We next investigate the relationship between article slant and the readership it garners by using article-level regression models. In the first set of models (Model 1), for each outlet we separately regressed the popularity of an article against its slant:

$$Y_i = \beta_0 + \beta_1 I_i + \epsilon_i \tag{3}$$

where Y_i is the number of visits, as recorded by the Bing Toolbar, and I_i is the article's ideological slant (between -1 and 1). Consistent with Figure 7, we find β_1 is statistically significant for only three outlets (*Daily Kos*, the *Washington Post*, and *USA Today*). For *Daily Kos*, β_1 is negative, indicating that left-leaning articles attract more attention; and for *USA Today* and the *Washington Post*, the result is the opposite. However, even for these cases where β_1 is statistically significant, the magnitude of the effect is relatively small, as shown in Table A4.

We further examined this issue with a second set of models (Model 2), where for each outlet, we test whether more polarized articles attract more attention. In this case, we regress article popularity on the absolute value of the slant of the article:

$$Y_i = \beta_0 + \beta_1 |I_i| + \epsilon_i \tag{4}$$

Table A4: Article popularity vs. slant

Outlet	Model 1				Model 2			
	β_0	β_1	p-value	R^2	β_0	β_1	p-value	R^2
Daily Kos	92.55	-34.33	0.01*	<0.01	84.08	51.90	<0.01*	<0.01
N.Y. Times	490.29	-106.49	0.13	<0.01	438.23	381.30	<0.01*	<0.01
Huff. Post	1409.94	374.92	0.16	<0.01	1364.43	156.06	0.62	<0.01
L.A. Times	92.26	-2.20	0.96	<0.01	96.22	-29.12	0.53	<0.01
CNN News	576.18	95.67	0.21	<0.01	572.86	7.01	0.94	<0.01
Yahoo News	1129.81	54.16	0.80	<0.01	1022.39	752.90	<0.01*	<0.01
BBC News	244.67	-31.54	0.57	<0.01	243.16	26.29	0.67	<0.01
Reuters	112.03	-5.09	0.83	<0.01	112.88	-8.08	0.77	<0.01
Wash. Post	264.87	144.74	<0.01*	0.01	246.76	81.97	0.16	0.01
Chicago Tribune	60.85	-34.77	0.11	<0.01	56.64	33.69	0.19	<0.01
NBC News	2917.67	-671.16	0.09	<0.01	2887.98	209.74	0.66	<0.01
USA Today	155.88	74.19	0.01*	0.01	151.19	42.33	0.19	0.01
WSJ	174.70	210.70	<0.01*	0.04	135.50	303.93	<0.01*	0.04
Fox News	969.24	1.11	0.99	<0.01	956.10	62.06	0.58	<0.01
Breitbart	216.23	35.90	0.35	<0.01	220.54	8.53	0.86	<0.01

We find β_1 is statistically significant for *Daily Kos*, the *New York Times*, *Yahoo News* and the *Wall Street Journal*. In all four cases, the effect is positive, indicating that more partisan articles are more popular. However, similar to Model 1, the effects are rather small (see Table A4).

A.7 Publisher classifications of news vs. opinion

In our primary analysis, we used worker labels to determine if an article was an opinion piece or descriptive reporting, since many outlets do not have the well defined sections found in traditional newspapers. However, for six of the 15 outlets, the news sites clearly categorize the articles, and moreover, these categories are readily apparent from the article links. As can be seen in Figure A5, on this subset of outlets the crowd labels yield results in line with those from the publishers' own classifications.

A.8 Negative coverage by issue

In our primary analysis, we show that news outlets almost universally portray both Democrats and Republicans negatively. Figure A6 shows that this result holds at the level of issues as well, with articles on nearly all issues having, on average, neutral to negative slant for both parties. Notable exceptions are gay rights, environment, drugs, and education. Note however, that these four issues have rather small overall coverage.

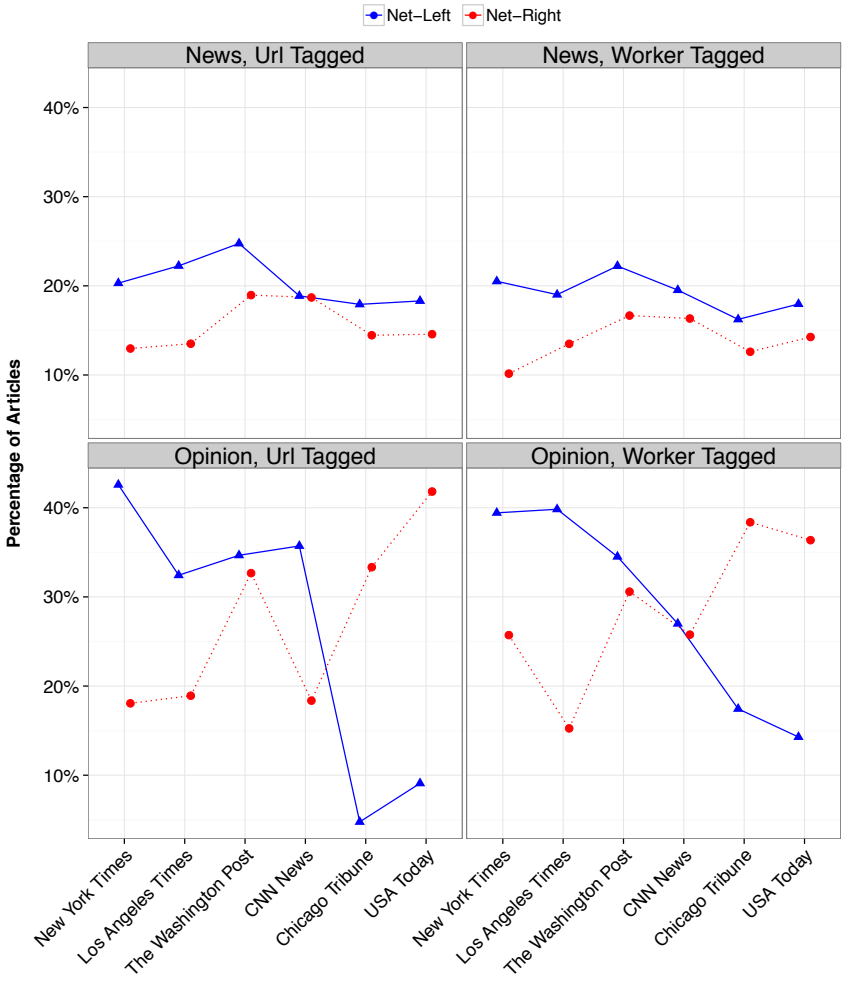


Figure A5: Comparison of worker-labeled and url-labeled categorizations of news and opinion.

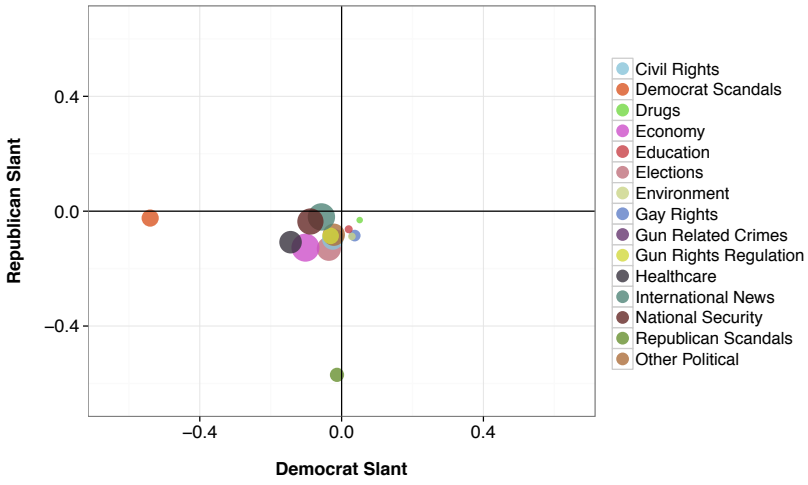


Figure A6: Democratic and Republican slant of articles per issue, aggregated across all outlets. The size of the points signify the overall coverage of the issue. The two extreme points are Democratic (orange) and Republican (green) scandals.

A.9 Coverage similarity

Table A5 shows the pairwise correlation in topic coverage across all pairs of news outlets we consider. By and large, we find high correlation in coverage between the outlets. One exception is *BBC News*, where the relatively low coverage similarity is attributable to its international focus.

	Daily Kos	New York Times	Huffington Post	Los Angeles Times	Washington Post	BBC News	CNN News	Yahoo News	Reuters	NBC News	Chicago Tribune	USA Today	WSJ	Fox News
New York Times	0.62													
Huffington Post	0.82	0.90												
Los Angeles Times	0.64	0.95	0.92											
The Washington Post	0.76	0.92	0.92	0.91										
BBC News	0.02	0.73	0.47	0.64	0.44									
CNN News	0.51	0.97	0.84	0.94	0.89	0.77								
Yahoo News	0.56	0.98	0.87	0.95	0.90	0.76	0.99							
Reuters	0.22	0.87	0.61	0.82	0.67	0.91	0.91	0.90						
NBC News	0.50	0.96	0.83	0.94	0.88	0.73	0.98	0.97	0.90					
Chicago Tribune	0.66	0.69	0.74	0.74	0.85	0.20	0.71	0.70	0.46	0.65				
USA Today	0.51	0.92	0.82	0.94	0.88	0.66	0.96	0.96	0.86	0.97	0.75			
Wall Street Journal	0.63	0.89	0.79	0.86	0.93	0.49	0.87	0.86	0.74	0.85	0.86	0.84		
Fox News	0.47	0.83	0.70	0.85	0.85	0.53	0.87	0.88	0.79	0.88	0.75	0.91	0.86	
Breitbart News Network	0.72	0.73	0.82	0.76	0.86	0.28	0.71	0.75	0.47	0.71	0.68	0.73	0.72	0.82

Table A5: Correlation of coverage for all pairs of news outlets