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

Ceren Budak
University of Michigan

Sharad Goel
Stanford University

Justin M. Rao
Microsoft Research

S1 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) with both linear and quadratic kernels (implemented in the open-source machine learning package Vowpal Wabbit (Langford, Li, and Strehl 2007)). The results show that logistic regression and SVM with linear kernel yield nearly identical performance while SVM with a quadratic kernel performs slightly worse than the other two approaches,[1] both on the subset of test articles where the two human judges agreed (Table S1), 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 S2). Given the inherent ambiguity in the task, the accuracy reported in Table S2, which captures the agreement between the classifier and a randomly selected human, provides a lower bound on the performance of our classifiers.

Table S1: 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 (linear kernel)	0.88	0.84	0.88	0.87	0.92	0.80
SVM (quadratic kernel)	0.84	0.79	0.86	0.85	0.89	0.80
Logistic Regression	0.87	0.82	0.90	0.87	0.91	0.81

S2 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

*Budak (cbudak@umich.edu) is the corresponding author.

1. SVM with a quadratic kernel is more prone to overfitting. Therefore, we experimented with various regularization settings and here report only the results from the best performing classifier on the test data set.

Table S2: 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 (linear kernel)	0.79	0.76	0.78	0.78	0.85	0.67
SVM (quadratic kernel)	0.77	0.72	0.78	0.76	0.82	0.67
Logistic Regression	0.79	0.74	0.81	0.78	0.84	0.69

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 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 (Kittur, Chi, and Suh 2008; Akkaya et al. 2010; Buhrmester, Kwang, and Gosling 2011; 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 S1. 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.

S3 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

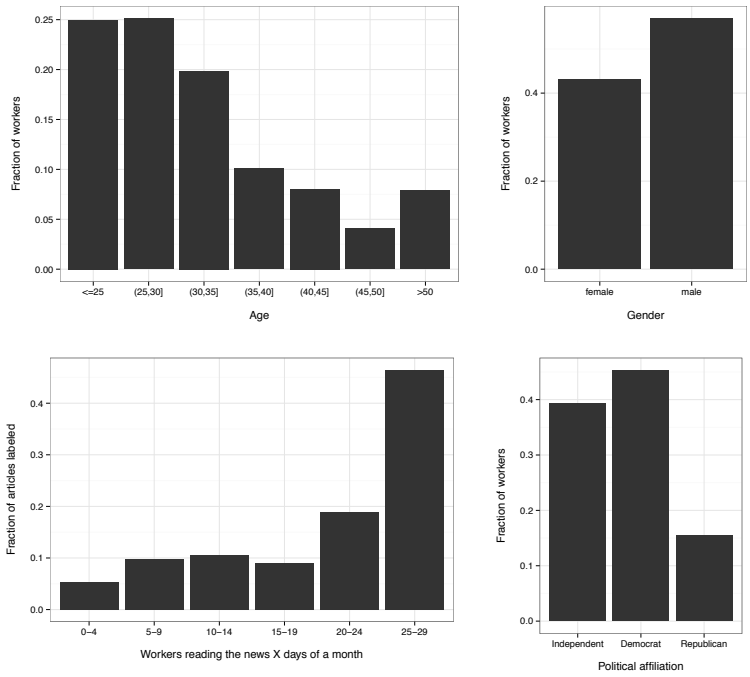

Figure S1: Demographic characteristics of Mechanical Turk workers participating in our experiments.

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 (one rater marking the article net-left while the other marks it net-right). 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. We find 65% topic agreement. Reported slant on average varied by 0.7 points on a five-point scale, and only in 5% of cases did the raters disagree in directionality. Both these figures are consistent with the results of the small, 20 article preliminary analysis, and also consistent with what others have achieved using traditional laboratory subjects (Baum and Groeling 2008).

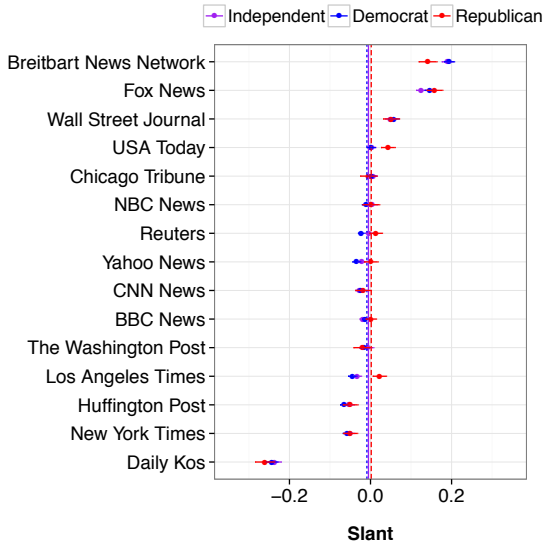

Figure S2: Outlet-level ratings by subjects' political affiliation. The points and lines are colored blue-to-red based on reported political affiliation, the vertical lines represent averages and lie virtually right on top of each other.

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 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 S1 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 (Vallone, Ross, and Lepper 1985; Perloff 1989; Eveland and Shah 2003). For example, in analyzing the 1982 Beirut massacre, Vallone, Ross, and Lepper (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 S2 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.[2] 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 S3 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.

S4 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. Moreover, as expected, the proportion of left- and right-leaning opinion stories in each

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

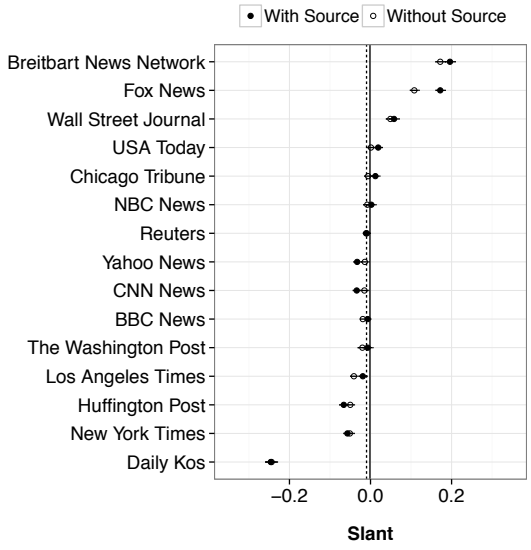

Figure S3: The solid points show reported slant when the article source was revealed, and the hollow points show the reported slant when the outlet name is not disclosed.

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 S4 shows the distribution of highly left- and right-leaning descriptive and opinion articles across outlets; the overall pattern is quite similar to the findings presented for overall net-left (score < 0) and net-right (score > 0) articles, indicating that our results are robust to the precise score threshold one applies.

S5 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, we additionally 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.[3] 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-

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

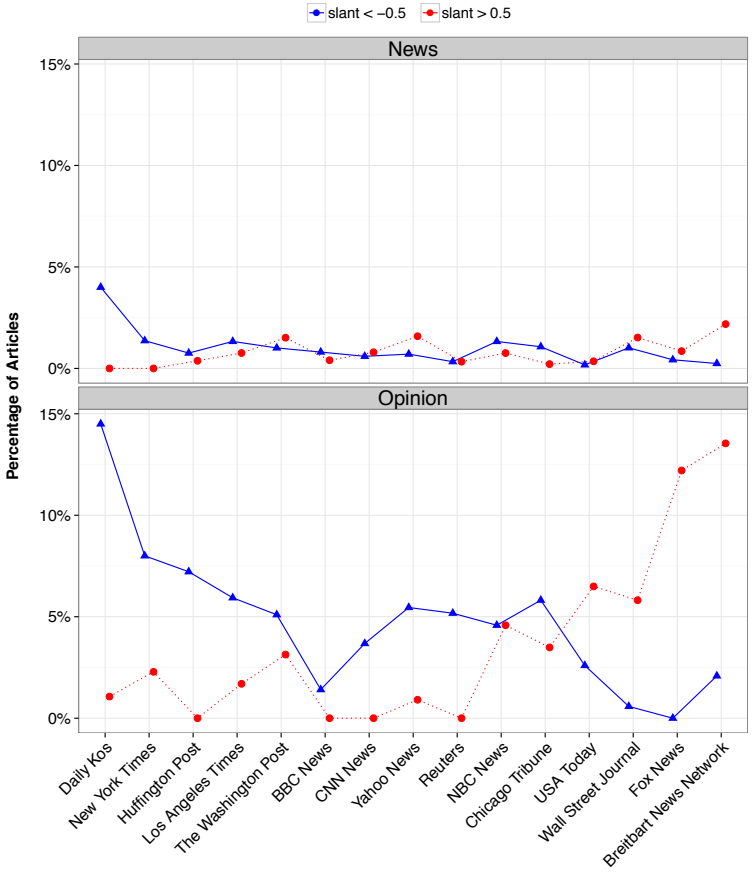

Figure S4: Fraction of articles that are highly partisan, defined as articles with a slant score greater than 0.5 in absolute value.

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,

$$\hat{s}_{\text{full}} = \frac{1}{n} \sum_{i=1}^{m} \frac{r_i}{p_i}. \tag{S1}$$

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 S3 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.

S6 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:

$$Y_i = \beta_0 + \beta_1 I_i + \epsilon_i \tag{S2}$$

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). 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 S4.

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:

$$Y_i = \beta_0 + \beta_1 |I_i| + \epsilon_i \tag{S3}$$

Table S3: 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

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 S4).

S7 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 S5, on this subset of outlets the crowd labels yield results in line with those from the publishers' own classifications.

S8 Negative coverage by issue

In our primary analysis, we show that news outlets almost universally portray both Democrats and Republicans negatively. Figure S6 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.

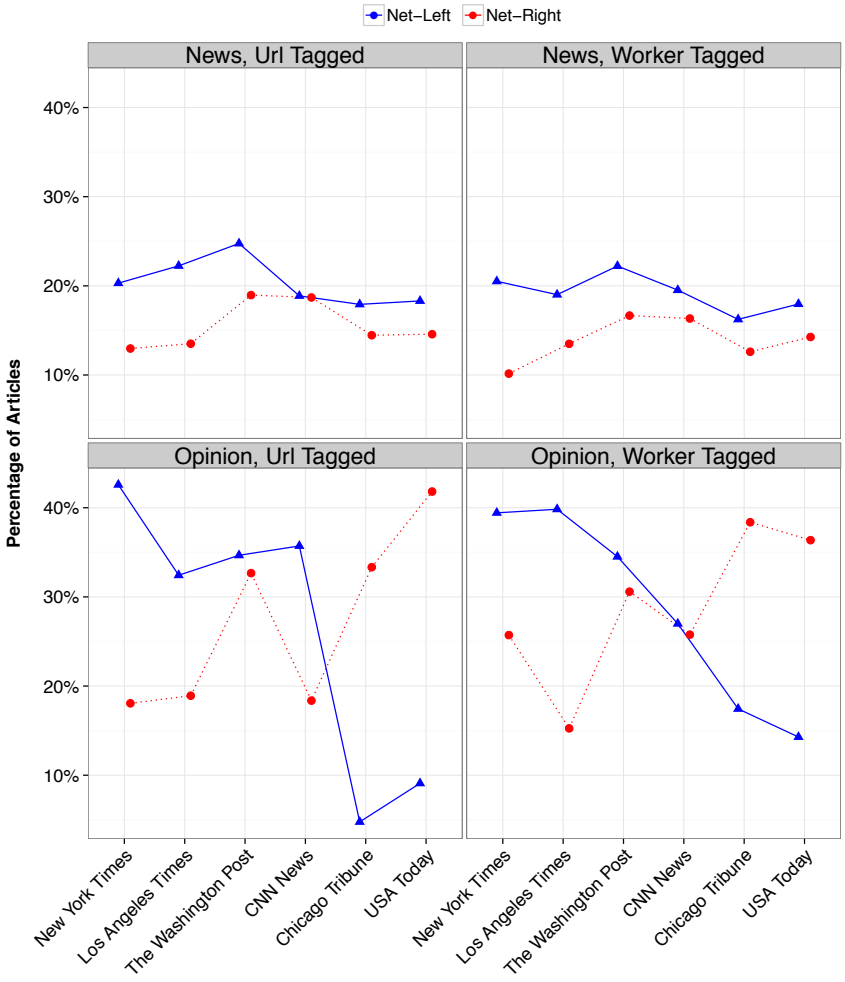

Figure S5: Comparison of worker-labeled and url-labeled categorizations of news and opinion.

Table S4: 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

S9 Coverage similarity

Table S5 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 S5: Correlation of coverage for all pairs of news outlets

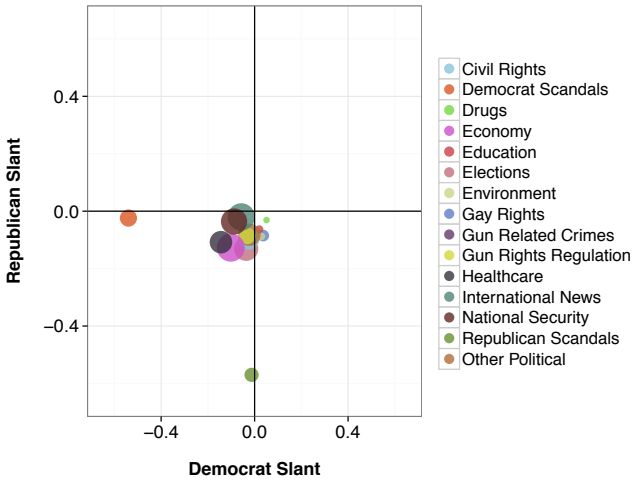

Figure S6: Democratic and Republican slant of articles per issue, aggregated across all outlets. The size of the points signify the overall coverage of the issue. The two extreme points are Democratic (orange) and Republican (green) scandals.

References

Akkaya, Cem, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 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,* 195–203. Stroudsburg, PA, USA: Association for Computational Linguistics.

Baum, Matthew A, and Tim Groeling. 2008. "New media and the polarization of American political discourse." *Political Communication* 25 (4): 345–365.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2011. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science* 6 (1): 3–5.

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

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

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

Kittur, Aniket, Ed H Chi, and Bongwon Suh. 2008. "Crowdsourcing user studies with Mechanical Turk." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 453–456. ACM.

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

Mason, Winter, and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior research methods* 44 (1): 1–23.

Perloff, Richard 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.* Accessed September 27, 2012. `http://www.people-press.org/2012/09/27/section-3-news-attitudes-and-habits-2/`.

Vallone, Robert P, Lee Ross, and Mark R Lepper. 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.