Do Journalists Drive Media Slant?

Karthik Srinivasan
University of Chicago
March 8, 2021

Abstract

I study the scope of a principle-agent problem in the field. I analyze news firms and journalists with possibly misaligned preferences over the partisan slant of content, and find that the firm's ability to exert control is limited. I construct a dataset that links 2,700 journalists to firms, news articles, and Twitter profiles. I measure article slant with a machine learning algorithm I train to identify partisan phrases. Using a movers design, I find firm ideology does not change the slant of a journalist's writing. In contrast, journalist ideology, estimated using the following decisions of Twitter users, is strongly correlated with article slant.

The production of news creates important social and political externalities: media coverage affects how people make high-stakes decisions related to elections, public health, and policing.[1] Given these consequences, it's important to understand why news outlets choose to produce slanted content. This study provides the first empirical evidence accessing the significance of a distinct channel for how these biases in coverage might arise — namely, journalist ideology.

Using web scraping, I produce a new dataset that links journalists to news outlets, Twitter accounts, and the full text of five hundred thousand news articles. Following Gentzkow et al. (2019), I train a machine learning algorithm to predict the party affiliation of members of Congress from the text of their speeches in the Congressional Record. This task allows me to identify phrases that are particularly predictive of partisanship, and I apply the trained algorithm to news articles to generate measures of slant.

My primary causal inference strategy compares the average slant of a journalist's writing across firms. This within-journalist strategy allows me to isolate the effect of the managerial components of the firm (e.g. editors and owners) holding fixed journalist-level characteristics (e.g. writing style). The central finding of this paper is a precise null result: when a journalist works for a more right (left) leaning publication as measured by it's mean article slant, the slant of the journalist's writing is unchanged on average. This result provides strong evidence against firms successfully exerting direct control over journalist's writing. For example, if a firm's culture or story assignment decisions pushed journalists to write in particular ways, we would expect a positive relationship between the firm's mean slant and the journalist's slant.

Beyond directly influencing a journalist's writing, firms may control their distribution of article slant indirectly by selectively hiring journalists who produce articles with the desired degree of slant. While my within-journalist strategy cannot rule out this kind of selection, I can provide some suggestive evidence by analyzing heterogeneity within my

[1] See DellaVigna and Kaplan (2007), Bursztyn et al. (2020) and Mastrorocco et al. (2020).

main result. I divide journalists into three groups based on the beats that they cover: soft news[2], policy, and politics. We would expect that soft news and policy journalists would be comparatively harder to select on ideology, and I find that the null result still holds within these groups. In contrast, the writing of politics reporters appears to be strongly responsive to the ideology of their outlet, suggesting that more obviously ideological writing is easier to directly control. Additionally, I find sharp heterogeneity by experience as measured by number of articles written: journalists in the bottom quartile of the experience distribution are strongly responsive to firm ideology while journalists in the top three quartiles are not. This provides further suggestive evidence that firms do try to exert direct control, but that their success depends on journalist characteristics.

Next, I turn to Twitter as a context where the firm exerts less control over content produced by journalists. I estimate a ridge regression that predicts DW Nominate scores for Twitter accounts associated with members of congress based on whether or not they are followed by a set of politically engaged Twitter users. I then apply the model to journalist Twitter accounts to generate a measure of journalist ideology. I find that this measure is strongly correlated with the slant of articles that journalists write, even after controlling for firm and beat-level fixed effects.[3] A 1 unit change in ideology (the ideological distance between Cory Booker and Mitt Romney) predicts a change 0.1 unit change in mean article slant (the ideological distance between the average New York Times and Fox News journalist). This exercise provides suggestive evidence that a journalist's personal ideology influences their writing.

This paper contributes to the literature on the causes of media slant. Early theoretical work posits that firms may want to produce slanted content to cater to consumer preferences (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006). Baron (2006) proposes a model where firms save money in labor costs by allowing journalists, who are motivated by career concerns, to produce more extreme content. A line of empirical work tries to access whether slant is more driven by profit concerns or the ideology of owners. Gentzkow and Shapiro (2010) finds that newspaper slant is more correlated with the Republican vote share than the political donations of owners, suggesting that owners are more driven by profit maximization. Supporting the consumer hypothesis, Martin and Yurukoglu (2017) estimate a structural model to identify to what degree consumers drive media polarization. In contrast, Grossman et al. (2020) analyzes the not-for-profit Isreali newspaper *Isreal Hayom* as an example of a political investment made by an owner.[4]

One common thread that runs throughout most of the literature is an emphasis on management-driven explanations of media slant. These explanations rest on an implicit assumption that owners and managers have the capacity to exert control over the slant of content that their firms produce. This paper breaks from the literature by quantitatively

[2]Soft news is defined as Arts and Entertainment, Beauty, Fashion, Food and Dining, Sports, Travel, and Weather.

[3]A journalist's beat is the issue, sector, organization or institution that they cover. In this paper, I derive information on a journalist's beat based on their categorization on muckrack.com.

[4]For a more complete discussion of the empirical literature on the causes and consequences of media slant, see Chapter 15 of the *Handbook of Media Economics* (Puglisi and Snyder Jr, 2015a).

accessing to what extent owners are able to exert control. One notable exception is Bursztyn et al. (2020), who try to understand how differential coverage of the risks of COVID-19 by *Fox News's* Sean Hannity and Tucker Carlson influenced the health-related behaviors of their viewers.

This study also relates to the literature concerned with measuring the partisanship of newspapers. One strand of this literature focuses on publishing decisions with clear ideological valence: think tank citations (Groseclose and Milyo, 2005), politician endorsements (Ansolabehere et al., 2006), editorials (Ho et al., 2008), letters to the editor (Butler and Schofield, 2010) and ballot proposition endorsements (Puglisi and Snyder Jr, 2015b) have all been used to compute publication-level measures of slant. Gentzkow and Shapiro (2010) propose the idea of comparing news text to the Congressional Record to produce measures of slant. I borrow methodology directly from Gentzkow et al. (2019) who propose the congressional party ID prediction task that I deploy in my machine learning measurement approach. Relative to the literature, this study is the first to attempt to measure ideology at the level of journalists instead of newspapers.

The rest of the paper is organized as follows. Section 1 defines a simple framework for thinking about firm control. Section 2 discusses the main measurement and causal inference strategy for analyzing the slant of online news articles. Section 3 presents an additional suggestive analysis comparing article slant to journalist ideology measured using data from Twitter.

1 Framework for Understanding Firm Control

1.1 Types of Firm Control

In the context of media slant, I will define two broad categories of firm control. First, firms can attempt to exert *direct control*. Here, I mean any behavior of the firm that causally influences the journalist's slant. This could include incentivized contracts, story assignments, editors or even the pressures of firm culture.

Second, firms can influence the distribution of slant through *indirect control*. In this category, I have in mind a selection story: for example, the *New York Times* might hire a conservative columnist away from the *Wall Street Journal* expecting the columnist to continue to write conservative columns. In this story, the management of the *New York Times* controls the publication's distribution of slant without casually influencing the way that journalists write.

Both direct and indirect control could be incomplete. Firms might have a partial causal influence on the writing of journalist, a possibility modeled in Subsection 1.2. Additionally, firms might select journalists on the basis of slant as well as other qualities like writing or investigative ability. To the extent that firm control over slant is incomplete, there is room for journalist ideology to causally influence the firm's distribution of slant.

1.2 Toy Model of Journalist-Firm Bargaining

I will formalize the idea of direct control with a toy model of journalist-firm bargaining. The model will serve as a micro-foundation for the main within-journalist across-firms causal inference strategy implemented in Section 2.

For simplicity, I represent partisan ideology as a one-dimensional spectrum. Suppose that the firm would prefer to have content with a slant of ϕ, but the journalist would like to write articles with a slant of μ. Let the observed slant of articles written by the journalist s be a weighted average of ϕ and μ with the bargaining weight of the firm denoted by β. That is:

$$s = \beta\phi + (1 - \beta)\mu$$

Now, consider the case of two firms, fixing the bargaining weight to be the same for both. If we observe the difference in slant between the articles written by the same journalist at each firm $(s_1 - s_2)$ as well as the difference between the firms' ideal slants $(\phi_1 - \phi_2)$, we can recover the bargaining weight β.

$$\begin{aligned} s_1 &= \beta\phi_1 + (1 - \beta)\mu \\ s_2 &= \beta\phi_2 + (1 - \beta)\mu \\ s_1 - s_2 &= \beta(\phi_1 - \phi_2) \end{aligned}$$

In practice, I'll estimate each journalist's ideal point from the text of their articles. Similarly, I'll estimate each firm's ideal point by from the text of all articles written for the firm. If the model is correctly specified and the mean slant of articles recovers ϕ, then slope estimated using the within-journalist design can be interpreted as the bargaining weight of firms.

2 Empirical Analysis of Slant

In this section, I'll walk through my procedure for measuring article slant. Then, I'll present the results of analyses which depend on this measure, including the main within-journalist across-firm specification and an analysis of heterogeneous effects.

2.1 Measuring Slant Using Machine Learning

In order to develop a measure of the partisan slant of text, I estimate a machine learning model that uses phrases from speeches found in the Congressional Record to predict the speaker's party affiliation (Republican or Democrat). This approach is based on Gentzkow et al. (2019), who propose this prediction task and make publicly available a dataset containing the full text of congressional speeches labeled with party affiliation. The premise of this exercise is that phrases which predict congressional party ID will continue to be markers of partisanship even when observed in the news context.

I clean all text data using standard natural language processing techniques.[5]. I start with 44,000 speeches from the 113th and 114th sessions of Congress which I divide into testing, training and validation datasets. I convert speeches to lower case and remove punctuation as well as numeric characters. I exclude a list of stop words, which are words that contain little meaning (e.g. and, or, the). I lemmatize words (for example, converting swimming, swims and swam to swim). I then construct a set of right-hand side variables with a bag-of-words model: I compute indicator variables for the presence of each two or three-word phrase found in the training data at least 100 times. These indicator variables are then reweighed by inverse document frequency, which is often found to improve prediction accuracy.[6] In practice, the out-of-sample accuracy of my model is robust to many data cleaning decisions including the exclusion of three-word phrases, the inclusion of one-word phrases, and different frequency thresholds for phrase inclusion.

I predict party affiliation using a calibrated support vector machine (SVM) model, an off-the-shelf machine learning classification method (Boser et al., 1992). I select a penalty parameter value of $C = 1$ using five-fold cross-validation.

I show that the SVM model does a good job recovering partisan ideology in three ways. First, my model has high out of sample accuracy. When applied to the 20% testing sample, it correctly assigns party affiliation 80.1% of the time. Bayram et al. (2019) attempts this same task using a variety of different machine learning approaches, and finds that correctly classifying around 80% of speeches represents the upper bound on performance for this task. Intuitively, some speeches contain very little ideological content, and so may be difficult to classify, especially when restricting to a simple feature set of words and phrases.

Second, I look at the phrases with the largest coefficients predicting Republican or Democratic affiliation, and qualitatively verify that the model is selecting sensible phrases ex-post (see Figure 1). Overall, the phrases the model chooses are sensible. The model learns from the topics politicians talk about (climate change, gun violence, immigration, abortion) as well as how they talk about them (carbon pollution, background check, unborn child, illegal immigrant). The model also makes some mistakes. Phrases like "rise opposition" and "encourage colleague" are apparently predictive of Democratic and Republican speech in congress, but are unlikely to be markers of partisan slant in the news.

Finally, I try to allay potential concerns about the coarseness of party affiliation as a measure. I rely on the fact that the calibrated SVM returns not only a classification (Republican or Democrat), but also an underlying probability that the observations is a member of the class. I predict the probability Republican for each out of sample speech, and compare the average probability Republican for each member of congress to their DW Nominate scores. DW Nominate scores are a frequently used measure of the political ideology of members of Congress estimated using data on congressional voting decisions. I compare my estimates to dimension 1 of the DW Nominate score, which is widely interpreted to capture left-right partisanship (see Poole and Rosenthal (1985)). These scores are highly

[5]A complete description of my data cleaning steps procedure can be found in the Appendix

[6]Document frequency refers to the number of occurrences of a phrase across all speeches in the training data. This reweighing amounts to dividing the indicator variable column by that frequency.

correlated (87%), which means that this text-based measure of ideology is able to recover fine-grained information about voting decisions in congress (see Figure 2). For my analysis of article slant, I'll always use the underlying probability Republican measure rather than the discrete Republican/Democrat prediction.

2.2 Within-Journalist Across-Firm Design

I build a new journalist-level news article dataset using web scraping. I link journalists to article URLs and Twitter accounts using data from Muck Rack, a website which helps PR professionals identify friendly journalists by cataloging information about their previous work. Muck Rack contains firm and journalist level information. Each Muck Rack firm page contains a list of journalist pages associated with the firm which I use to create a sample of journalists. Each Muck Rack journalist page contains information about the beats the journalist covers, their previous places of employment, a set of recent Tweets, as well as a history of all URLs written by the journalist in the Muck Rack database. I write scrapers that visit each URL and return the full article text for eight online news websites: the *New York Times*, *USA Today*, *Fox News*, as well as five regional newspapers.[7] In total, I am able to get the full text of 500,000 articles written by 2,700 journalists. I apply the SVM model to predict the probability Republican for each article in my dataset, and then aggregate up to produce measures of mean slant at the journalist and journalist-outlet levels.

One unique feature of this dataset, relative to previous work on media slant, is that I can learn something about the distribution of article slant within each firm as opposed to just learning the mean. Figure 3 shows the 25th, 50th and 75th percentiles of journalist ideology in the 8 firms that comprise my sample. The triangular indentation around the median is a bootstrapped 95% confidence interval, so we can read off from this graph that the median *Fox News* journalist is significantly more right leaning that the median *USA Today* journalist who in turn is significantly more right leaning than the median *New York Times* journalist. Despite these differences in median slant, the other key feature that's apparent from the graph is that there is substantial overlap in journalist ideology across outlets: there are plenty of journalists who write for the *New York Times* whose average slant is not distinguishable from journalists who write for *Fox News* based on this measure.

Now, I turn to estimating the within-journalist across-firm design micro-founded in Section 1. I restrict attention to the 1,091 journalists who have written articles for multiple outlets. I plot the difference in mean outlet slant on the x axis and the difference in journalist slant on the y axis. Figure 4 describes the causal inference design. The dashed 45°-line represents a bargaining weight coefficient of 1, and corresponds to complete direct firm control. Intuitively, this line means that a journalist who moves to a 1 unit more right leaning firm writes content that is 1 unit more right leaning. In contrast, the x axis represents a bargaining weight of 0 and corresponds to no firm direct control. The x axis means that as a journalist moves to a 1 unit more right leaning firm, the slant of their writing does not change. The x axis is consistent with any degree of firm selection on ideology, from no

[7]The regional newspapers are the *Chicago Tribute*, the *Miami Herald*, the *Minnesota Star Tribune*, the *Boston Globe*, and the *Seattle Times*.

control to complete indirect control.

In the full sample, moving to a more right or left leaning outlet does not change the average slant of a journalist's writing on average. This central result of the paper is represented graphically in Figure 5. Each point represents the average of all within-journalist moves for a given outlet pair. A more intense red hue represents a larger number of moves. The most well-estimated points are labeled to give a sense of which moves are driving identification. The red line represents a linear regression, and the red shading represents a 95% confidence interval for the relationship between outlet slant and journalist slant.

This result should be interpreted with a few cautions in mind. First, the estimate is primarily identified off of small moves: most moves are between 0.01 and 0.03 units (for context, the distance between the mean *New York Times* and *Fox News* articles is about 0.1 units). Second, note that this is an average effect: as the many off-x-axis gray points suggest, there is a lot of underlying variance in these slant estimates at the journalist-outlet pair level.

2.3 Heterogeneity by Beat and Experience

The within-journalist design cannot rule out firm control through selection on ideology. To get at this question, I split the sample into three broad areas of coverage using Muck Rack's beat classification scheme. I define policy journalists as journalists who cover at least one beat in business and finance, crime and justice, education, energy, environment, health, media, military, politics, real estate, religion, science, technology, or transportation. I define soft news journalists as journalists with at least one beat in arts and entertainment, beauty, fashion, food and dining, sports, travel, or weather. Finally, I define political journalists as journalists who write op-eds or cover the politics beat. Figure 6 gives the distribution of journalists in my sample by beat.

I rerun the main within-journalist specification for each of these three groups of journalists.[8] The main result essentially holds for policy journalists (see Figure 9). The coefficient on soft news journalists is negative, which is hard to rationalize in the bargaining framework (see Figure 10). However, this estimate has a very wide confidence interval due to the fact that there are relatively few soft news moves. To the extent that it's harder to observe and therefore select on the ideology of these journalists since their content is less political, this is evidence that firms are note exerting complete indirect control.

In contrast, when I run the movers specification on political journalists, I find a bargaining coefficient of almost exactly 1, suggesting that political journalists are very responsive to firms (see Figure 11). One explanation that is consistent with this finding is that the partisan valence of political writing is less costly to observe, so firm direct control is more successful.

Finally, I conduct a heterogeneity analysis by experience, using number of articles written by a journalist that are in my sample as a proxy for their experience. Figure 7 describes the distribution of article counts in my sample, and demonstrates that I observe a wide

[8]Note that I recompute the mean article at each firm including only articles within the category, which changes the size of differences between outlet slants.

range of journalist experience. Figure 8 shows the within-journalist specification computed for journalists who are and are not in the bottom quartile of the distribution. I find that inexperienced journalists are responsive to firm slant, while experienced journalists are not. This suggests that inexperienced journalists may have an easier time changing writing styles or may be more susceptible to firm pressure.

3 Predicting Journalist Ideology Using Twitter Followers

One downside of the analysis of slant is that every article is fundamentally a product of an interaction between a journalist and a firm. In this section, I leverage the fact that Twitter allows journalists to produce content outside of the firm. I use the "following" decisions of a set of politically engaged Twitter users to infer journalist ideology, and then compare that ideology to article slant. The purpose of this exercise is to provide suggestive evidence that journalist ideology is informing the way that journalists write.

3.1 Training a Ridge Model on Congressional Twitter Accounts

My goal is to produce a model that takes a Twitter account as an input and produces an estimate of partisan ideology. To do this, I need a set of accounts whose ideology is already known. I rely on 900 Twitter accounts belonging to members of congress, and I use DW Nominate scores as a measure of partisan ideology. I estimate a ridge regression to predict DW Nominate scores where my right hand side variables are indicators for whether or not these congressional accounts are followed by a set of 15,000 Twitter user accounts randomly sampled from the followers of journalists and members of congress. I divide the set of congressional accounts into training, testing and validation samples, and I calibrate the ridge penalty parameter using five-fold cross validation.

This model preforms well. Predicated DW Nominate scores have a 89% positive correlation with actual scores when the model is applied out of sample (see Figure 12). This demonstrates that the following decisions of these users do contain information about the partisan ideology of members of congress.

3.2 Comparing Journalist Ideology to Article Slant

I apply this ridge model to generate predicted DW Nominate scores for each journalist in my sample. A nice feature of this setting is that journalists have a high take up of Twitter: I am able to link over 80% of journalists in my article slant sample to Twitter accounts.

I compute the correlation between the average slant of a journalist's writing and their predicted DW Nominate score (see Figure 13). The raw correlation is large. A 1 unit change in predicted DW Nominate score is associated with a 0.22 unit change in article slant. To benchmark these results, a 1 unit difference in DW Nominate scores is about the difference in ideology between an average Democratic and Republican congressmember, and a 0.1 unit

change in slant is about the difference between the average article at Fox News and the New York Times.

I then recompute this correlation at the journalist and journalist-outlet levels including fine grained outlet and beat-level fixed effects. These controls reduce the size of the correlation, but it remains quantitatively large, with a 1 unit change in DW Nominate scores being associated with at least a 0.12 unit change in article slant (see Table 1).

There are two potential concerns with this method that I cannot address. The first is that users may follow journalists and members of Congress for different reasons. If, for example, users like to follow members of Congress to display their ideological preferences, but follow journalists for non-partisan reasons like keeping up with certain issues, then the score may not reflect the journalist's own ideology, but instead could be interpreted as a measure of the ideology of their Twitter followers. A second, related concern is that user following decisions are likely endogenous to a journalist's choice of firm. For example, working at Fox News may garner a journalist more conservative followers, thus making them appear to hold a more conservative ideology. Because of these problems, I cannot tease apart two stories: (1) journalists who have more conservative followers on Twitter are more conservative, and therefore write more conservative articles, and (2) journalists who write more conservative articles garner more conservative followers on Twitter.

Nevertheless, this exercise is useful in establishing that variation in journalist slant is meaningful. If you believe the first story that the predicted the DW Nominate score measures a journalist's ideology, then this exercise is evidence that ideology shapes writing. If instead you believe the second story that Twitter followers are due to their writing, this analysis still suggests that journalists' article slant decisions are affecting career-relevant outcomes, as Twitter following is widely believed to be important for journalist's career prospects (evidenced in part by the unusually high take up of Twitter among journalists). Either way, the fact that the correlation holds within-firm is evidence that journalists are making consequential decisions that cannot be accounted for by firm-level factors.

4 Conclusion

I provide the first systematic empirical assessment of the role that journalists play in the production of slant in online news media. To do this, I produce new article-level estimates of slant using a machine learning approach and a journalist-article linked dataset created through web scraping. I find that as journalists write for more ideologically extreme firms, the slant of their writing does not change. Moreover, I find that this effect is primarily driven by experienced journalists and journalists who cover policy and soft news topics. In contrast, inexperienced journalists and journalists who cover political topics are responsive to the ideology of their firms.

Additionally, I compute estimates of journalist ideology based on the following decisions of Twitter users. I find that this measure is strongly correlated with article slant, even after controlling for firm and beat level fixed effects.

Together, these analysis provide evidence that journalists play an important role in the

production of slant. While I cannot rule out indirect firm control through selection, my causal inference design does provide evidence that firms are not using direct channels like editors or cultural pressure to change the slant of journalists writing on average. Moreover, I find that there is important within-firm variation in article slant that correlates with journalist's predicted DW Nominate scores and by definition cannot be accounted for by firm-level characteristics.

Much of the theoretical and empirical media slant literature implicitly assumes that owners have a high degree of control over the slant of content produced at their firm. This project undermines that assumption, and suggests that owners are at the minimum subject to labor market constraints based on the distribution of journalists' ideology.

References

Stephen Ansolabehere, Rebecca Lessem, and James M Snyder Jr. The orientation of newspaper endorsements in us elections, 1940–2002. *Quarterly Journal of political science*, 1 (4):393, 2006.

David P Baron. Persistent media bias. *Journal of Public Economics*, 90(1-2):1–36, 2006.

Ulya Bayram, John Pestian, Daniel Santel, and Ali A Minai. What's in a word? detecting partisan affiliation from word use in congressional speeches. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Leonardo Bursztyn, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. Misinformation during a pandemic. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-44), 2020.

Daniel M Butler and Emily Schofield. Were newspapers more interested in pro-obama letters to the editor in 2008? evidence from a field experiment. *American Politics Research*, 38 (2):356–371, 2010.

Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

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

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

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87 (4):1307–1340, 2019.

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

Guy Grossman, Yotam Margalit, and Tamar Mitts. Media ownership as political investment: The case of israel hayom. 2020.

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

Gregory J Martin and Ali Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–99, 2017.

Nicola Mastrorocco, Arianna Ornaghi, et al. Who watches the watchmen? local news and police behavior in the united states. Technical report, University of Warwick, Department of Economics, 2020.

Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95(4):1031–1053, 2005.

Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384, 1985.

Riccardo Puglisi and James M Snyder Jr. Empirical studies of media bias. In *Handbook of media economics*, volume 1, pages 647–667. Elsevier, 2015a.

Riccardo Puglisi and James M Snyder Jr. The balanced us press. *Journal of the European Economic Association*, 13(2):240–264, 2015b.

A Machine Learning Data Cleaning Steps

The dataset of congressional speeches is split into training, testing, and validation samples in the following way:

- I start with the set of 44,000 congressional speeches given during the 114th congress (January 3, 2015, to January 3, 2017)
- I withhold 10% validation random sample
- I keep only speeches with valid firstnames and a party ID that is either R or D
- I keep speeches with at least 200 words. This is meant to be a lower bound on the length of news articles, and to focus on speeches that are long enough to contain ideological content
- The subsample of speeches is then split into 80% training data and 20% testing data

The training speeches are then processed according to the following steps:

- All text is converted to lowercase.
- All text is lemmatized using the SpaCy English lemmatizer. This is a process, similar to stemming, that converts words to a root form (e.g. "swim", "swimming", and "swims" are converted to "swim"). Unlike stemming, lemmatizing guarantees that the words are "valid" words in the English language.
- Stop words are removed. Stop words are words that are common and thought to convey little meaning. I remove the set of English stopwords defined in NLTK's English stopword corpus, as well as a set of congressional stopwords identified in Gentzkow, Shapiro, Taddy (2019).
- Punctuation and numbers are removed

The function maps from words and phrases to ideology scores. To determine the set of words and phrases that the function takes as inputs, I compute the set of all 1 and 2 word phrases that occur at least 100 times within the training data. Rather than using the count of times each of these phrases occurs as the features (variables) in the function, words are weighted by the frequency that they appear in the original set of congressional speeches. This

practice, called Term-Frequency, Inverse-Document Frequency (TFIDF) weighting, helps improve the accuracy of the model.

When computing a score on a new set of text (e.g. test congressional speeches or news articles), the text is processed using the same initial steps. Then, phrases that appeared in the congressional vocabulary are counted and weighted by Inverse Document Frequency before the function is applied.

B Appendix Figures

Figure 1: Top Democratic and Republican Features

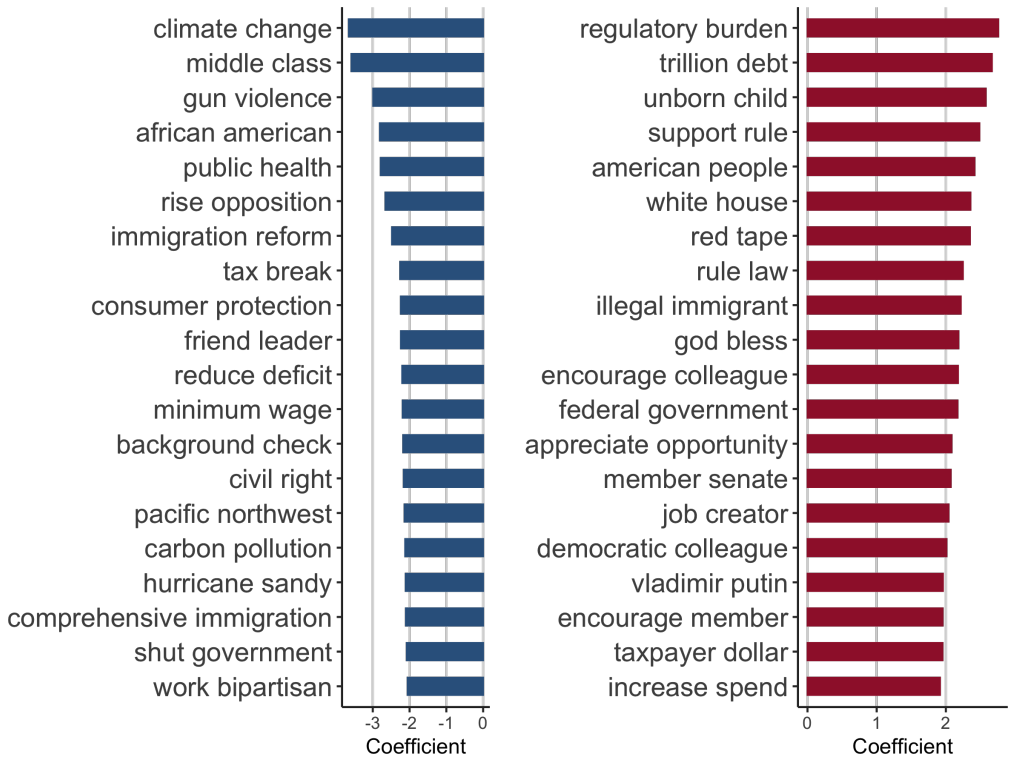

Table 1: Twitter Ideology vs. Article Slant

	Journalist Level	Beat FE	Journalist-Outlet Level	Outlet FE	Outlet, Beat FE
Twitter Ideology	0.217***	0.196***	0.170***	0.134***	0.127***
	[0.181, 0.254]	[0.159, 0.233]	[0.136, 0.204]	[0.098, 0.170]	[0.089, 0.165]
Num.Obs.	2326	2326	4601	4601	4601
R2	0.055	0.123	0.021	0.034	0.074

* p < 0.1, ** p < 0.05, *** p < 0.01

Figure 2: Probability Republican vs. DW Nominate Scores

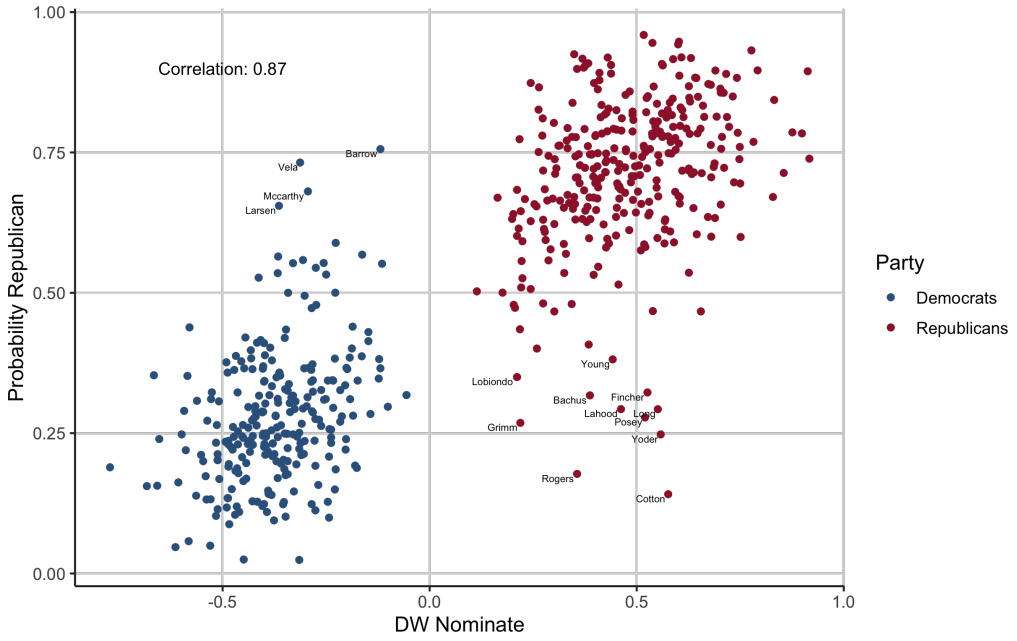

Figure 3: Distribution of Journalist Slant by Outlet

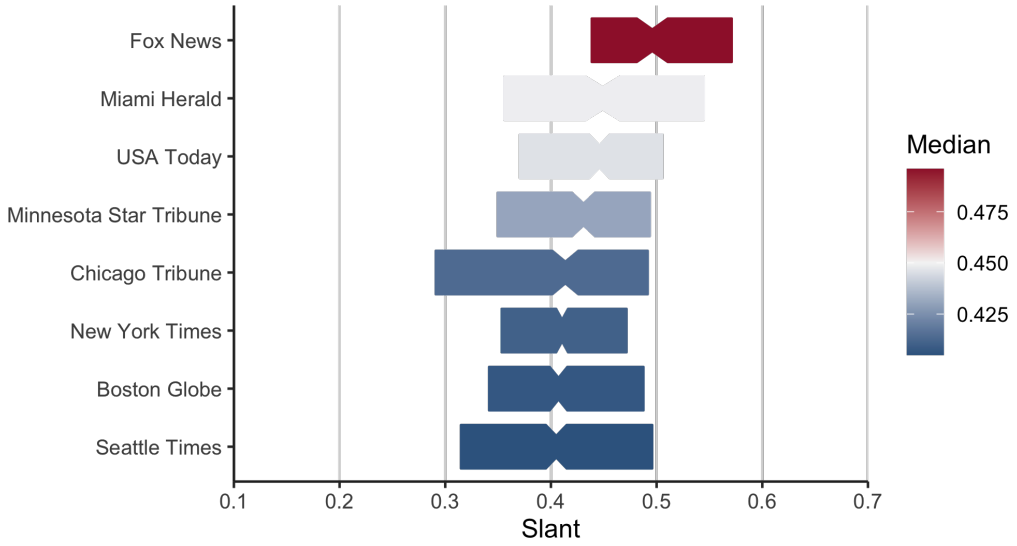

Figure 4: Within-Journalist Across-Firm Design

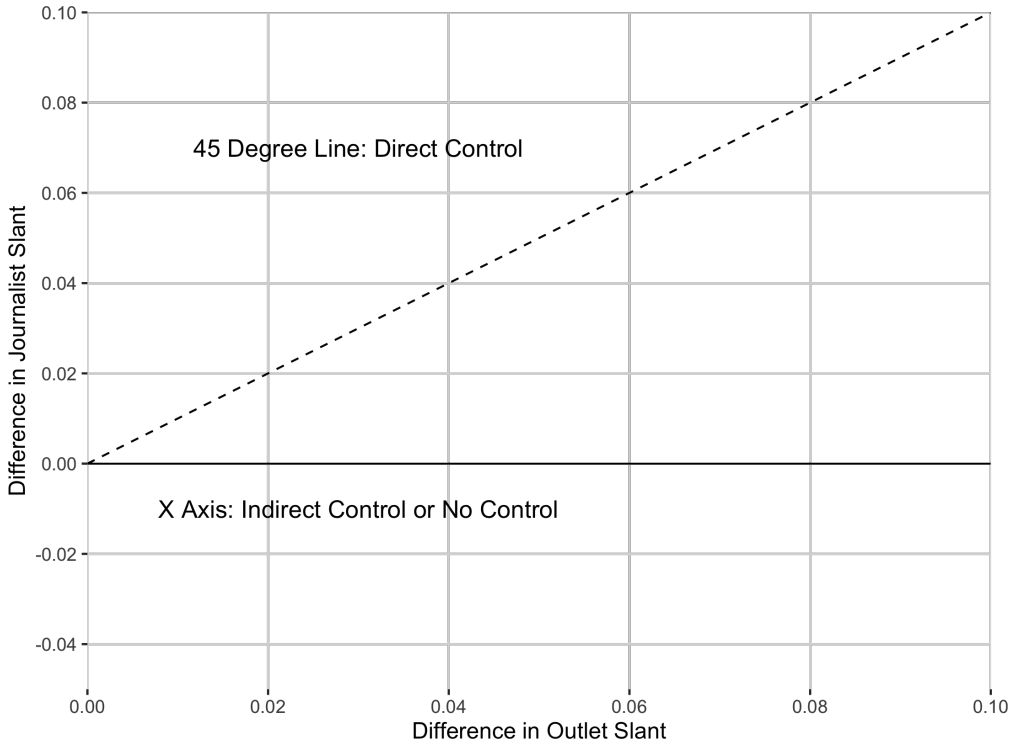

Figure 5: Full Sample

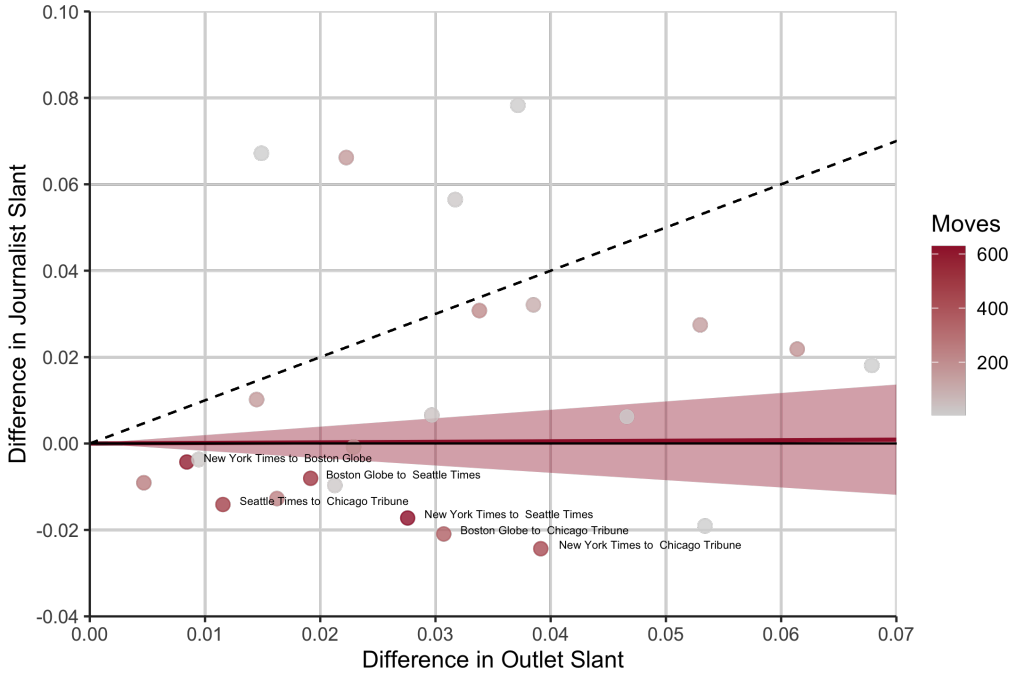

Figure 6: Journalists by Beat

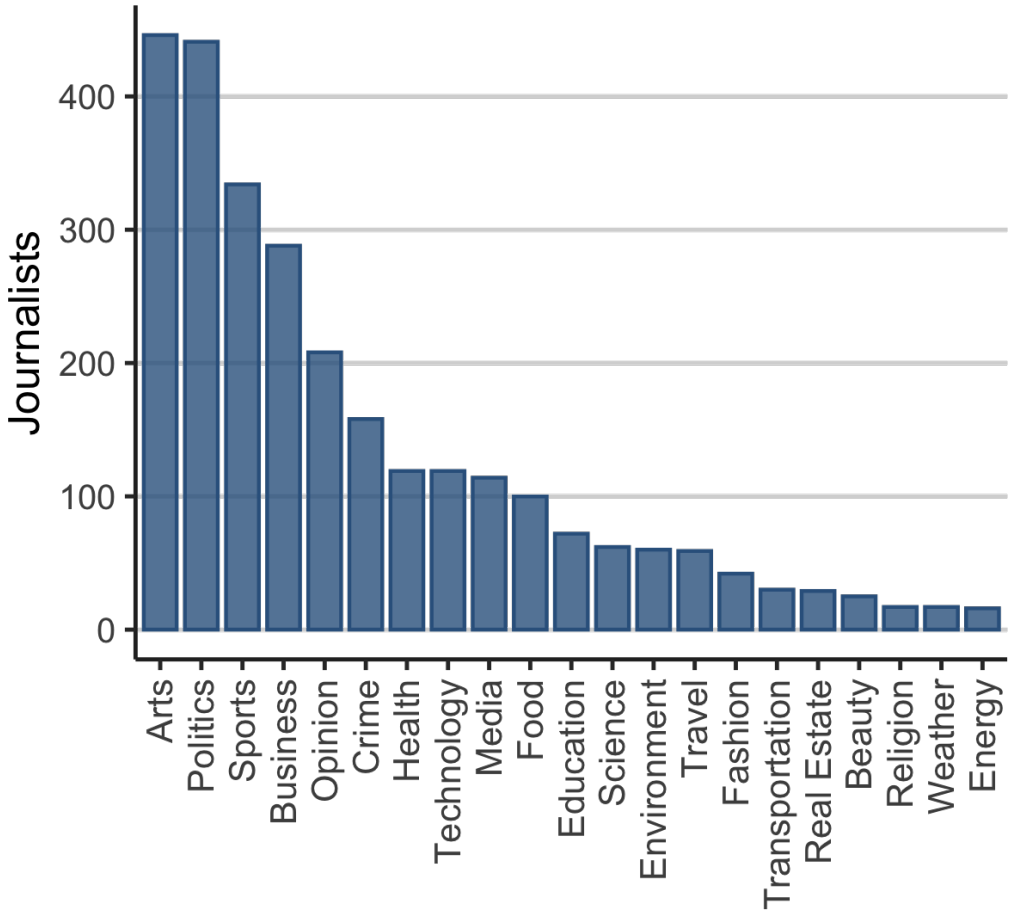

Figure 7: Article Count Distribution

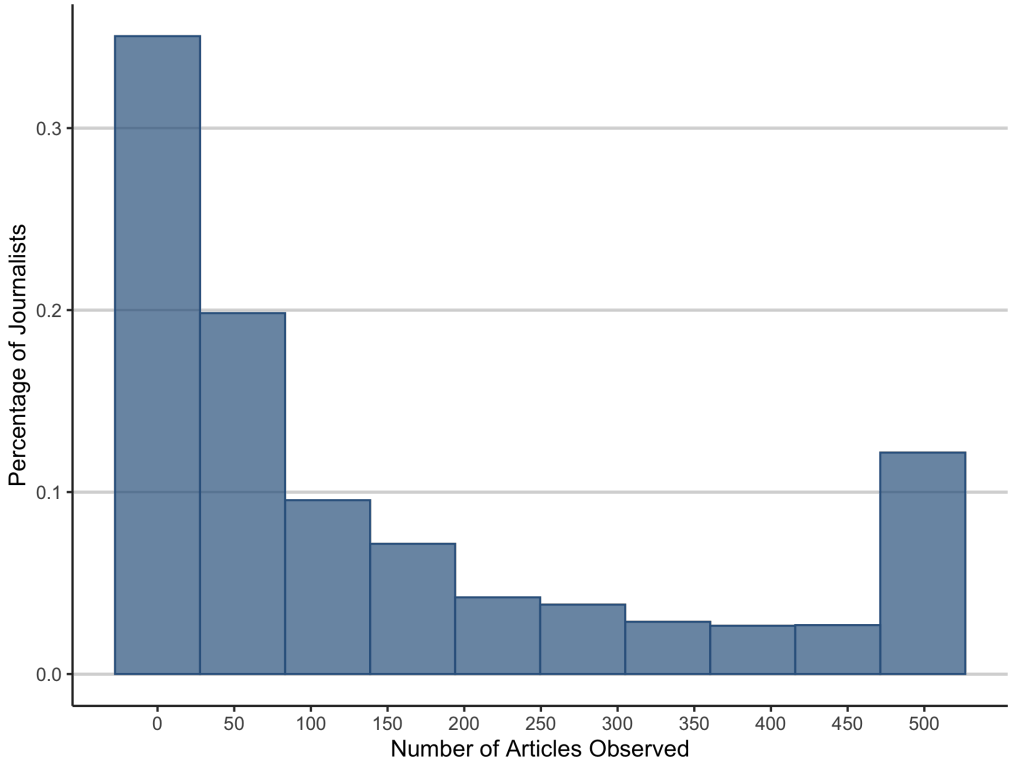

Figure 8: By Experience Quartile

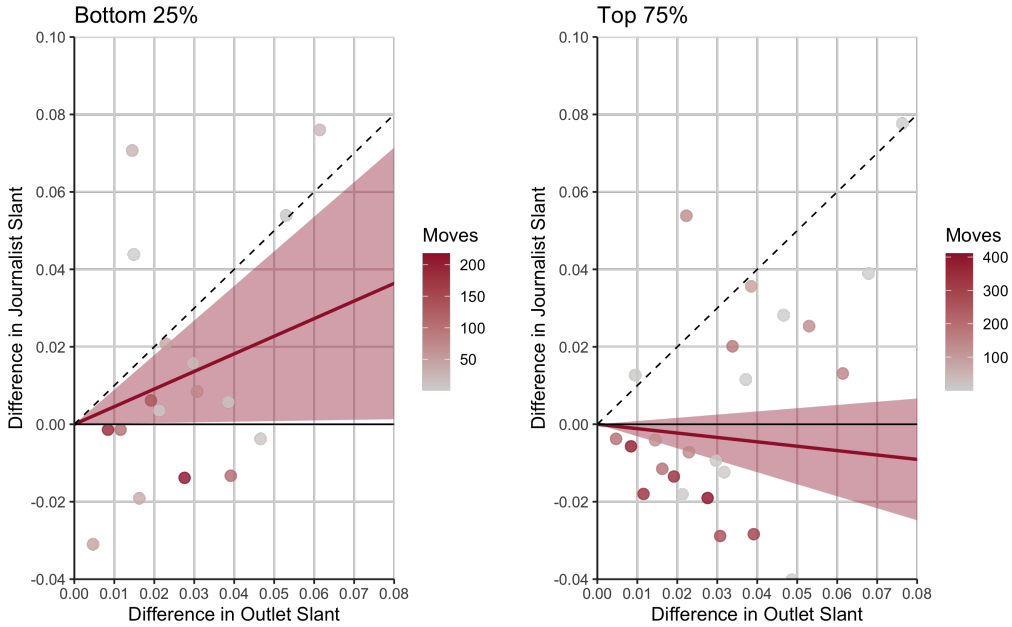

Figure 9: Policy Journalists

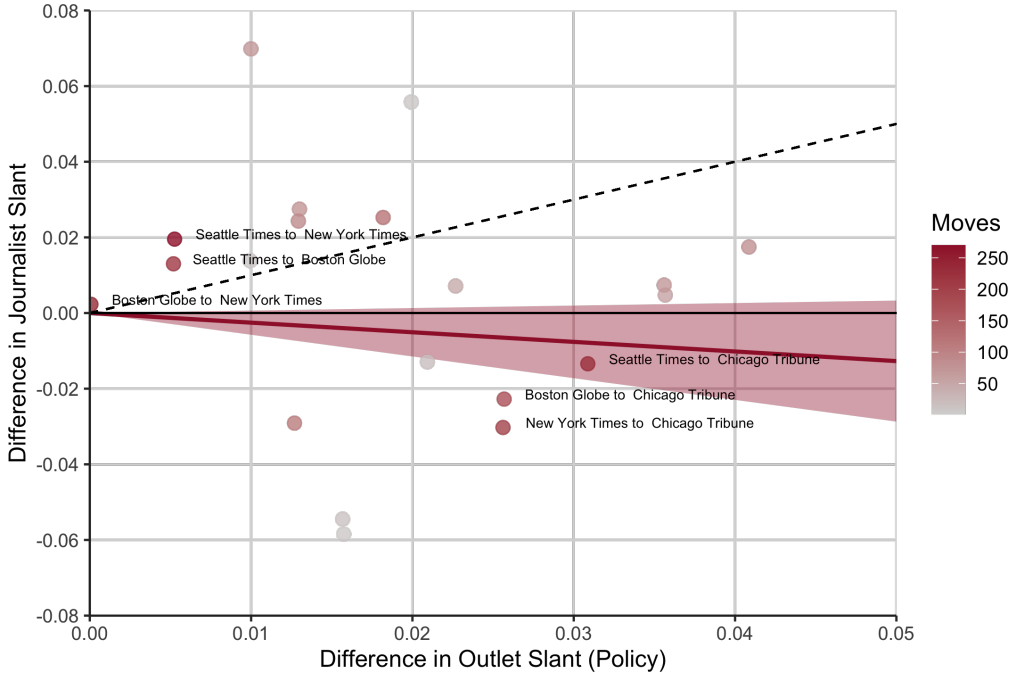

Figure 10: Soft News Journalists

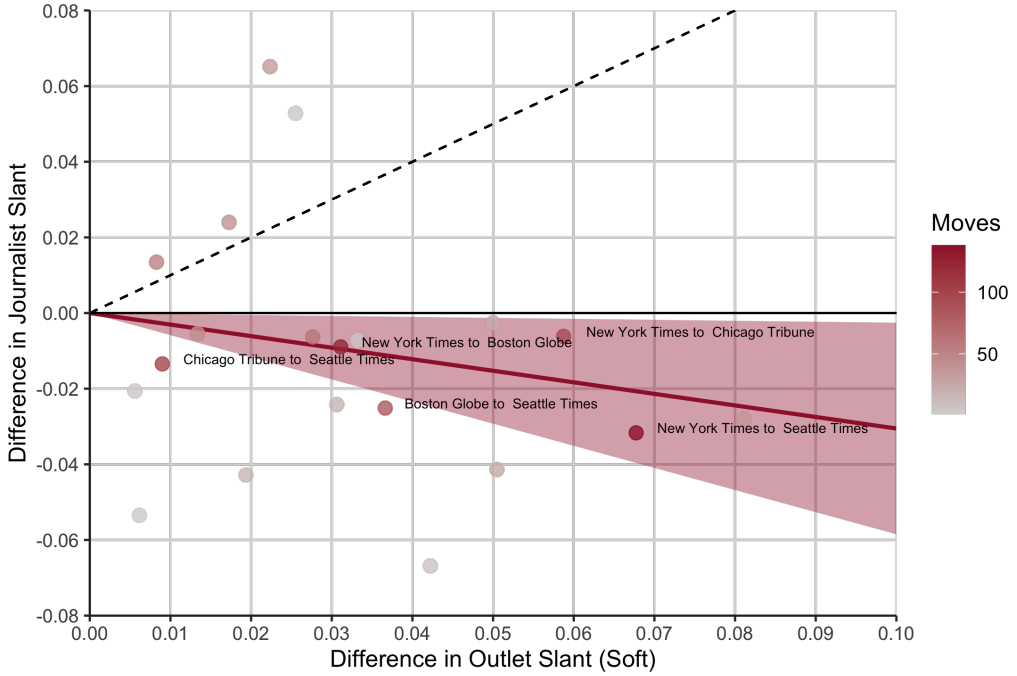

Figure 11: Politics Journalists

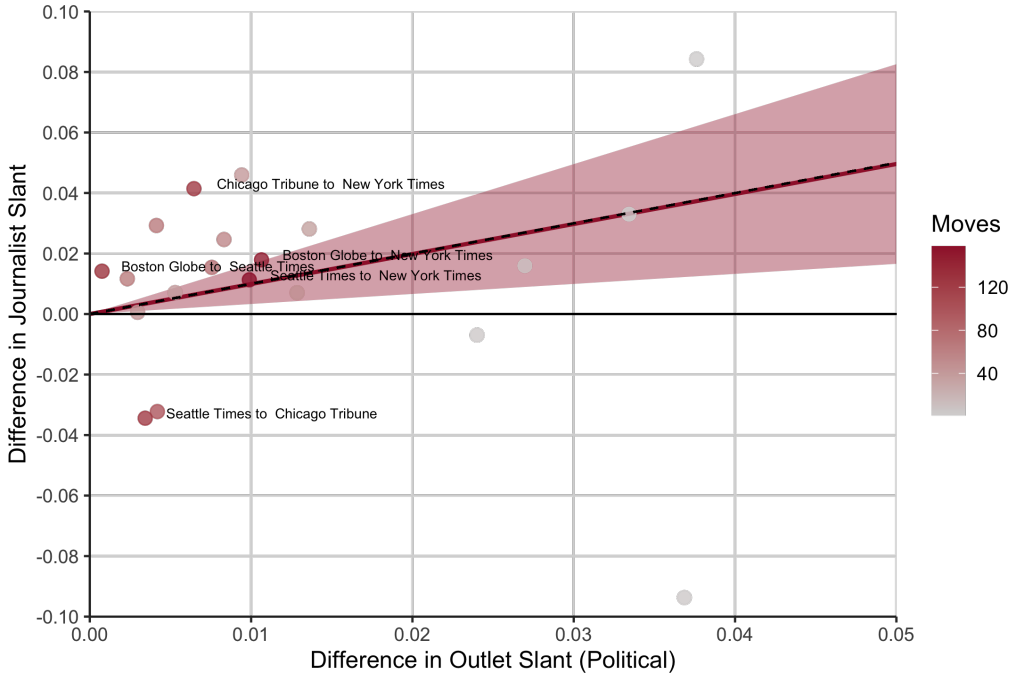

Figure 12: Predicted vs. Actual DW Nominate Scores

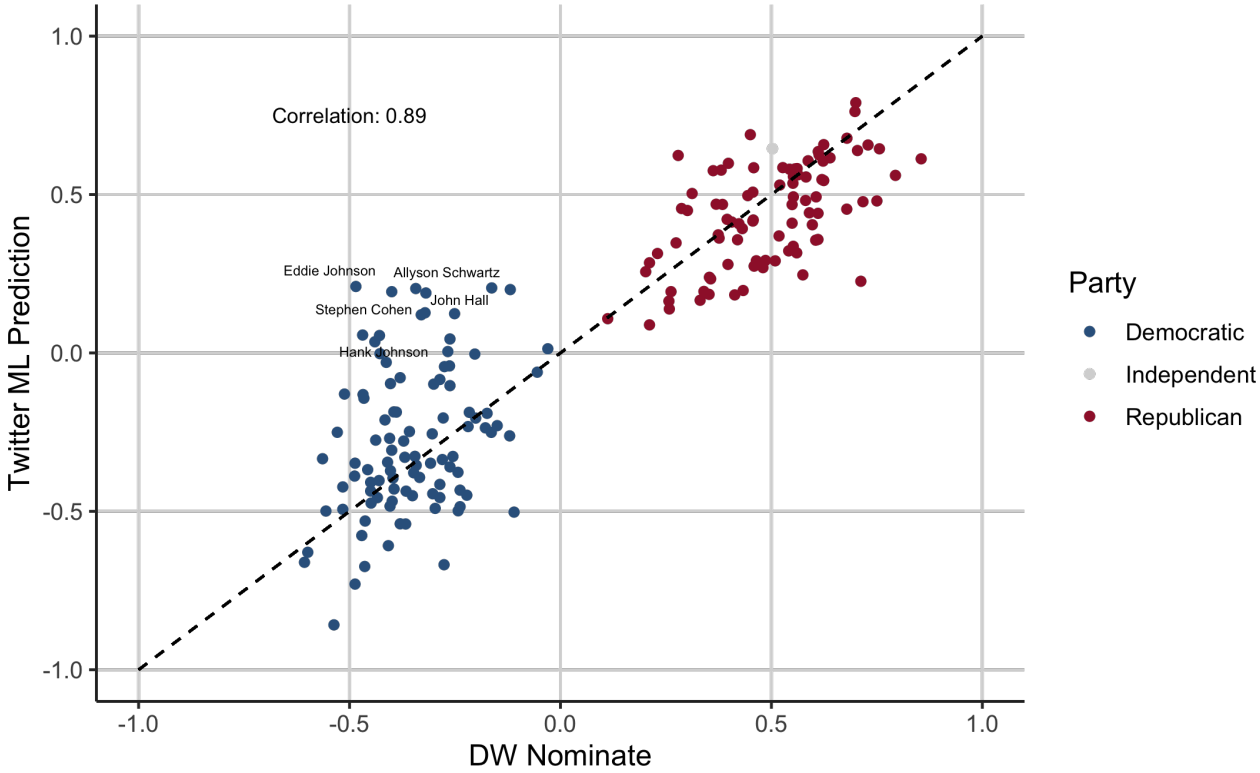

Figure 13: Twitter Ideology vs. Article Slant

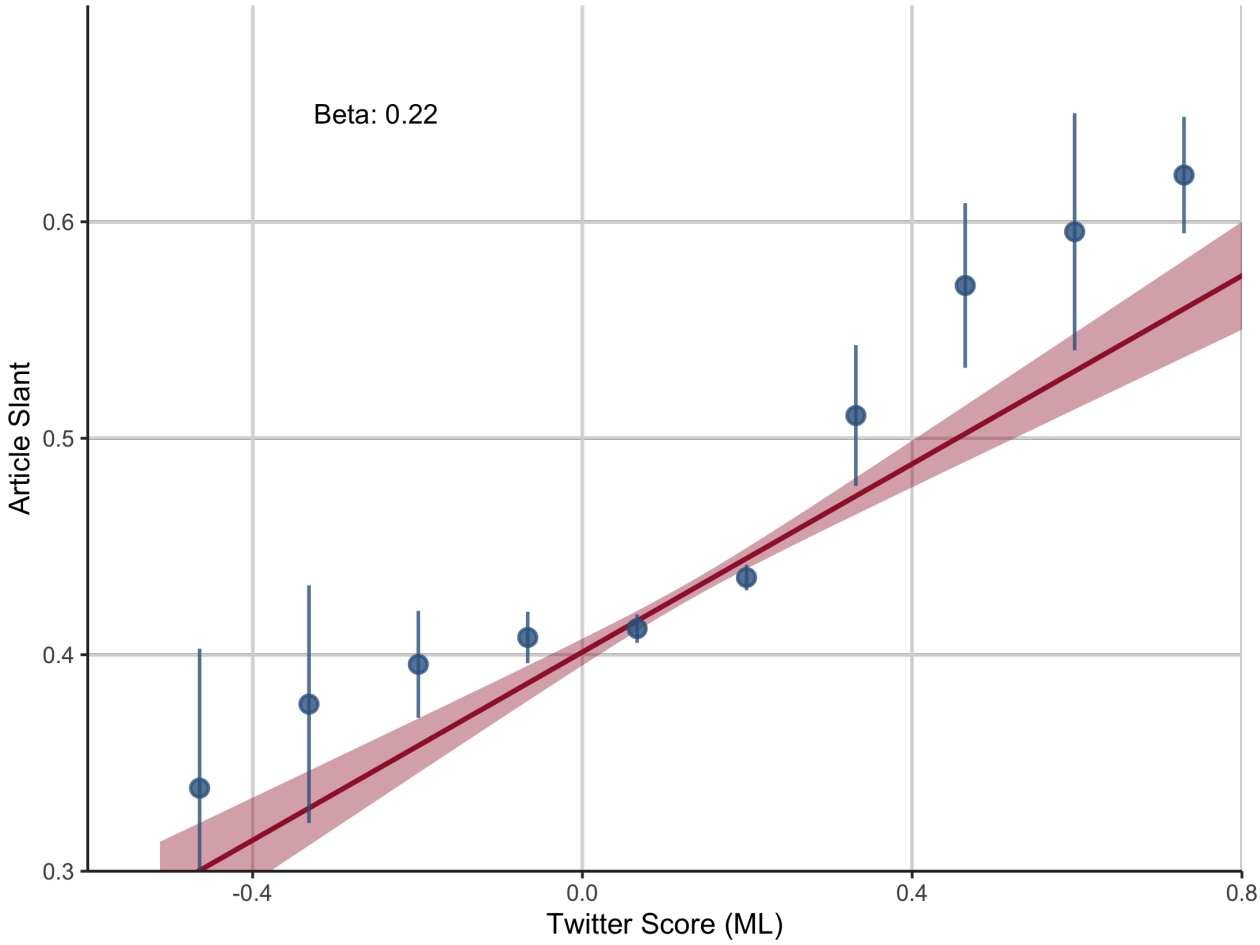

Figure 14: Policy Journalist Slant by Outlet

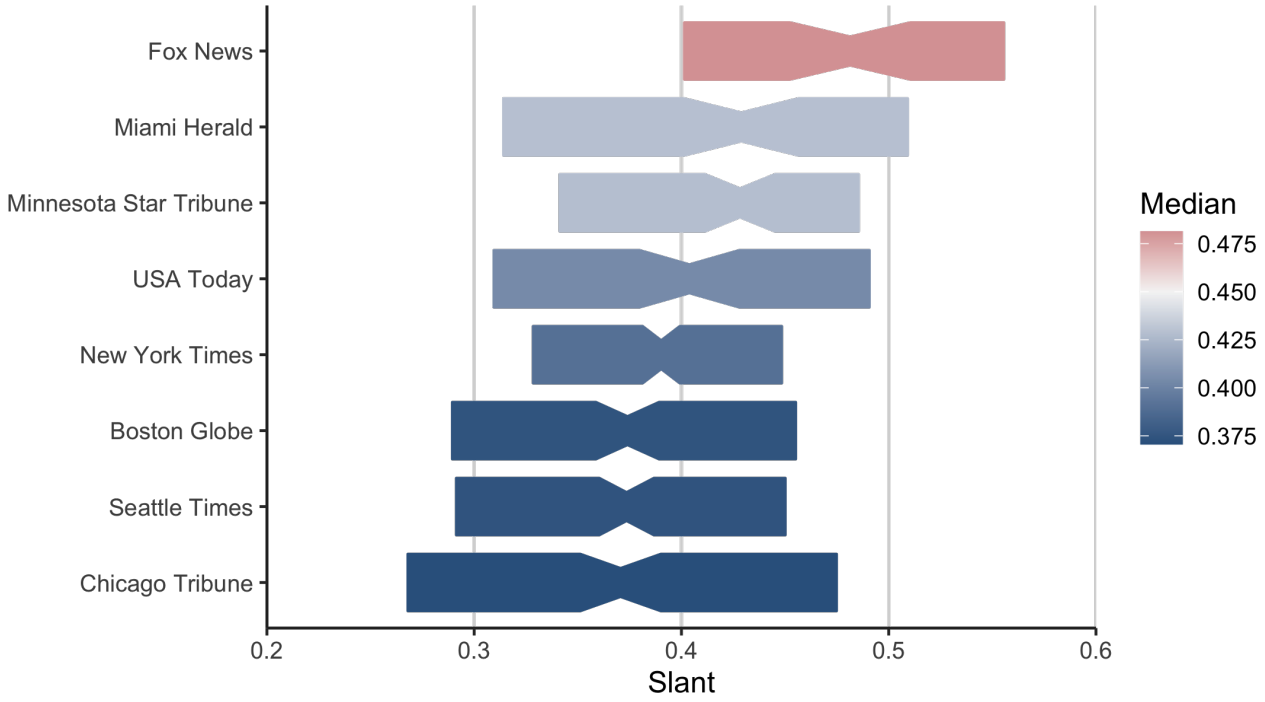

Figure 15: Soft News Journalist Slant by Outlet

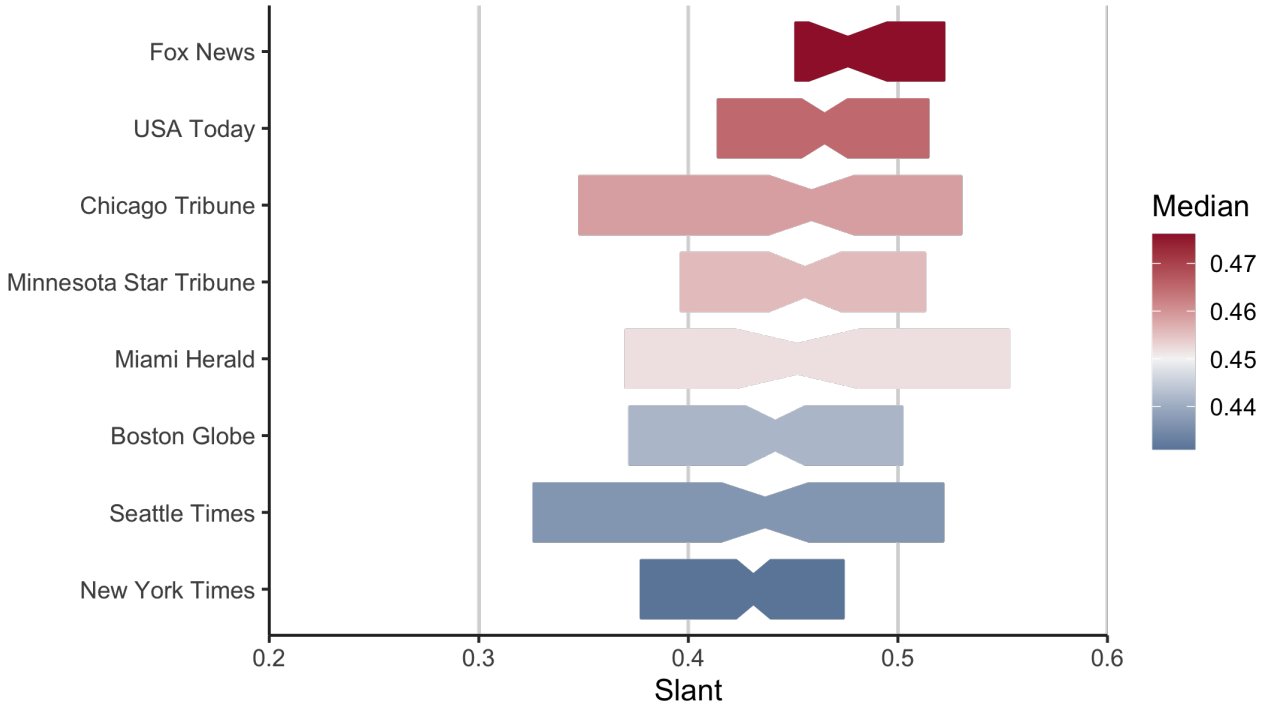

Figure 16: Political Journalist Slant by Outlet