Analyzing Text Data Indexed by a Hidden Characteristic: Method and Application to the Study of Media Bias in the U.S.

Yimeng Li*

August 2021

Click here for the latest version.

This is a preliminary draft. Please do not cite.

Abstract

Partisan bias enters political news coverage through selective coverage and biased presentation, and understanding the contribution through each channel necessitates estimating the difference in ideological leanings conditional on covering the same events. Political events, however, are an underlying characteristic that is unavailable to researchers absent large-scale hand-coding. In this paper, I propose a novel method to estimate quantities of substantive interest conditional on a hidden characteristic of the text data. The method overcomes difficulties of existing methods that are sensitive to critical tuning parameters and produce biased estimates. Applying the proposed method to the study of media bias, I find the ideological difference between media organizations is not always driven by how events are covered. Instead, the contribution of presentation bias to the total difference in ideological leanings between news articles published by media organizations varies in ways that are consistent with the structural difference between different media types. My results shed light on the exposure of different types of voters to the partisan bias of political news coverage.

*Ph.D. candidate in social sciences, Division of the Humanities and Social Sciences, California Institute of Technology. Email: yimeng.li@caltech.edu. I thank R. Michael Alvarez, Federico Echenique, Alexander V. Hirsch, Jonathan N. Katz, Gabriel Lopez-Moctezuma, Robert P. Sherman, and Matthew Shum for discussions and comments. I thank Kosuke Imai, Burt Monroe, Jonathan Nagler, and Molly Roberts, and the audience at APSA 2020 for comments. All errors are my own.

1. Introduction

Understanding partisan bias in the news is crucial as it has critical implications for political discourse, which may potentially affect political behavior or policies through the consumption of news by both ordinary voters and political elites. Partisan bias may enter political news coverage through two types of choices by newsmakers: which events or information to cover and how to cover them (Groeling 2013; see also D'Alessio and Allen 2000), referred to as "selection bias" and "presentation bias", respectively. The impact of media bias depends on the interaction of how it manifests and the way voters and elites acquire knowledge about ongoing political events. For example, if a voter only follows online breaking news of the day, which all major media organizations cover, perhaps through links shared on social media, then they would not be subject to selection bias. Similarly, aggregator sites such as Google may undo the selection of news coverage by the media organizations and leave voters with presentation bias, depending on whether they feature predominantly liberal or conservative media sources. On the other hand, if a voter relies exclusively on an ideologically-aligned media organization for news consumption, then they would likely be affected by the partisan framing and selective coverage. Estimating the magnitude of these two components, therefore, helps us understand the exposure of different types of consumers of news to a media organization's partisan bias.

Many text data in political science, including but not limited to corpora of news articles, have the feature that an observation is indexed by its author (individuals or organizations) and a characteristic (e.g., political events) that is often hidden in the text and unorganized without large-scale hand-coding. For instance, researchers often observe a stream of news articles from each media

organization but do not know which pairs or groups of articles are covering the same event without close inspection of each article. Ideally, one would want to read and compare all pairs or groups of articles and recover the underlying data structure, but this may be infeasible in large corpora. On the other hand, many quantities of substantive interest depend crucially on the hidden index. As discussed above, American politics scholars may be interested in measuring differences in ideological leanings between media organizations in the coverage of the same underlying political events. Comparative politics scholars may be interested in whether state-run media restrict discussion on sensitive news coverage more heavily than non-state counterparts in authoritarian countries. In both cases, estimating quantities that necessitate controlling for event-level heterogeneity is challenging.

In this paper, I propose a novel method to estimate quantities of interest which are conditional on a characteristic hidden in the text. Instead of extensive hand-coding, the method proceeds in the following four steps. It starts by forming candidate pairs of observations that potentially (but not necessarily) share a common index. After this initial pairing procedure, I require a sample of these candidate pairs to be labeled by researchers or human subjects as training data. Utilizing the training data, I estimate predictive models of the probability of a pair of observations having the same hidden index (1) as a function of a summary measure of text similarity and (2) as a function of the similarity measure and the difference in the outcome variable. In the context of media bias, the simplest case would be estimating the probability that a pair of articles cover the same event (1) given text similarity and (2) given both text similarity and the difference in ideological leanings. Finally, I construct an estimator of the

quantities of interest using the fitted predictive model, the similarity of all pairs, and metadata (if available).

I then apply my method to decompose media bias of U.S. news outlets into presentation bias and selection bias. A close examination of the corpus of political news articles yields two observations about text similarity between pairs of articles. The distributions of text similarity are generally smooth with no clear cutoff, and text similarity is correlated with the difference in ideological leaning conditional on covering the same event. These features make existing methods undesirable and call for an alternative method better fitted to estimate these quantities of interest. My approach overcomes this problem by assigning probabilities to pairs of articles being matches, and incorporating these probabilities in the estimation yields substantively different conclusions. In particular, I find the ideological difference between media organizations is not always driven by how events are covered. Instead, the contribution of presentation bias to the total difference in ideological leanings between news articles published by media organizations varies in ways that are consistent with the structural difference between different media types.

The rest of the paper is organized as follows: I first discuss related methodological literature in Section 2. In Section 3, I formally lay out the setup, define the quantity of interest, introduce my method, and compare it to the state-of-the-art method. Section 4 is devoted to the application to the study of media bias. I then conclude.

2. Literature

There is a large body of text literature devoted to the analysis of text data with underlying structures. One big area is text classification, where researchers usually invoke supervised learning methods to assign texts to researcher specified categories. For example, Budak et al. (2016) train a classifier that identifies "political" news articles from a pool of news articles using logistic regression models and support vector machines. Grimmer and Stewart (2013) consider an example of classifying Russian language public statements by political and military elites as having a restrained, activist, or neutral position on the Russian use of force, with a random forest algorithm. Once documents are classified into different categories, researchers can estimate quantities of interest conditional on the categories by taking sample analogues. Text classification via the supervised learning method, while commonly used in content analysis, presents challenges when there are more than a handful of categories. When applied to the study of media bias, it can yield quantities such as differences in ideological leanings between media organizations covering certain issue areas. It, however, cannot disentangle presentation bias from selective coverage within an issue area. Correlation between misclassification and outcomes of interest may also introduce bias to the estimates.

An alternative is document clustering, where researchers partition texts to mutually exclusive and exhaustive groups according to text similarity. For example, the most widely used clustering algorithm, the K-Means algorithm (MacQueen, 1967), minimizes the within-cluster sum of squared distances. Grimmer and King (2011) develop a computer-assisted clustering method and apply it to classify 250 Reuters news articles into one of 22 categories, among other exam-

ples. Subsequent analysis can be performed based on the categories uncovered by the unsupervised learning methods. While clustering algorithms have the advantage of uncovering insightful organizations of text, it is often hard to know how much the resulting clustering is different from the underlying structure governing the quantities of substantive interest. Clustering models can handle a large number of categories, but the number of categories is a crucial tuning parameter that must be determined. When applied to the study of media bias, unsupervised methods can yield differences in ideological leanings between media organizations within each cluster. Correlation between text similarity and outcomes of interest here, however, implies the choice of the number of clusters will govern the direction and magnitude of bias in estimates.

Both text classification and document clustering place texts into categories. An alternative route to the estimation of quantities of interest dependent on the underlying structure is to find matching pairs of texts that are aligned on the hidden characteristic. Recently, Roberts et al. (2020) develop a matching approach to address the problem of text-based confounding in observational studies. In the settings they consider, conditioning on features of the texts by finding matched pairs allows researchers to conduct analysis as if the treatment were randomly assigned. Mozer et al. (2020) employ a text-matching method on a corpus of news articles collected by Budak et al. (2016) to estimate the magnitude of presentation bias, and conclude that "most differences in favorability appear to be driven by presentation bias." However, as I discuss in Section 3.6, estimating quantities of interest such as presentation bias is a different task from handling text-based confounding for which the text matching method is designed. The nature of their corpus being only a sample of political news articles over the

time window introduces additional difficulty to this text-matching solution to the problem. The method proposed in this paper, by contrast, is better suited to estimating such quantities of interest conditional on a hidden characteristic and does not suffer from the biases inherent in a deterministic matching approach.

3. Method

3.1 Setup

Consider the following empirical framework of news coverage by media organizations. T events $\{E_1, \ldots, E_T\}$ occur, and each of I media organizations $\{1, \ldots, I\}$ decides whether to cover each event and if so, how to cover it. Table 1 displays an example where three media organizations each decided to publish a news article about four out of eight events that occurred (e.g., media 1 covered events E_2 through E_5).

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
Media 1		$\times$	$\times$	$\times$	$\times$			
Media 2	$\times$		$\times$	$\times$		$\times$		
Media 3	$\times$	$\times$		$\times$			$\times$	

Table 1: Underlying Data Structure

Without comprehensive hand-coding, the mapping between news articles to events as illustrated by Table 1 is unobservable to researchers. Researchers, instead, observe a stream of news articles from each media i. Let A_i^k denote article k published by media organization i. Let E_i^k and Y_i^k denote the event and the outcome of interest associated with article A_i^k, respectively. For instance, Y_i^k

is a measure of ideological leaning of article k by media i in the case of media bias.

3.2 Quantity of Interest

This paper proposes a novel method to estimate quantities of interest taking the form $\mathbb{E}[f(Y_i^k, Y_j^l)|g(E_i^k, E_j^l)]$, i.e., quantities conditional on a hidden characteristic—an event—which is unobservable without hand-coding.[1] Examples of such estimands include $\mathbb{E}[Y_i^k|E_i^k \in C]$, where C is some prespecified category, e.g., mean ideological leaning for media i's articles covering events in category C, and $\mathbb{E}[\mathcal{I}((Y_i^k, Y_j^l) = (1,1))|E_i^k, E_j^l \in T]$, where $\mathcal{I}$ is the indicator function and T is some prespecified topic, e.g., the rate at which two articles, one from each media organization, on topic T are both prohibited from discussion. Hereafter, I focus on $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$, quantities of interest conditional on the same events, as the main application necessitates. But similar strategies in terms of estimation can be employed for quantities conditional on a hidden characteristic generally.

While estimating these quantities would be straightforward if one were to read and compare all pairs of articles and recover the underlying data structure, hand-coding at such scale is typically infeasible or undesirable in large text corpora.

[1] Quantities of other forms may also be of interest. Unconditional quantities $\mathbb{E}[f(Y_i^k, Y_j^l)]$, e.g. difference in ideological leanings of articles between media i and media j, can be estimated by taking the sample mean. Conditional quantities $\mathbb{E}[f(Y_i^k, Y_j^l)|g(X_i^k, X_j^l)]$, where X_i^k denotes observable characteristics associated with article A_i^k such as author gender, can be estimated by sample mean for given values of observable characteristics nonparametrically, or parametrically via regressions.

3.3 What Constitutes an Event?

News coverage is often more complicated than this simple setup. While some newsworthy stories are stand-alone, other news stories are developing with follow-ups. In other cases, significant or controversial speeches or activities often spark responses that may be worth coverage on their own. So what constitutes an event? While an event may be defined broadly, e.g., the Trump-Kim Vietnam Summit, or granularly, e.g., the Trump-Kim Vietnam Summit ended with no joint agreement, how researchers should specify events depends on the goal of the research. If researchers define events at a granular level, then ideological leanings conditional on the same event will differ only due to framing and assembling of details. On the other hand, when researchers specify events broadly to encompass circumstances like follow-ups and responses, then choices among developments or responses will also contribute to presentation bias.

The specification of events should be compatible with the substantive quantities of interest. The method, however, only requires researchers to consistently determine the event each news article covers, if he or she were to inspect the articles manually. Moreover, to estimate $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$, it suffices to determine whether two articles cover the same event according to researcher specified criteria.[2]

I shall make two additional remarks before proceeding. First, in the empirical framework, each news article is identified with one event. This is important as, eventually, quantities of interest need to be expressed in terms of news articles instead of underlying events that are unobservable without manual inspections. Theoretically, a news article covering multiple events should be

[2]Similarly, to estimate $\mathbb{E}[Y_i^k|E_i^k \in C]$, where C is some prespecified category, it suffices to determine whether each article falls into category C.

split into multiple shorter pieces, but automatic splitting is likely infeasible. Operationally, most news articles have main themes, and I identify an article with the main event under coverage in the application. Second, the setup permits a media organization to publish multiple stories on an event.

3.4 Initial Pairing Procedure

Quantities of interest, such as the difference in ideological leanings between news articles covering the same events and the difference in the rates of discussion prohibition or censorship between news posts on the same event, of form $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$, fundamentally concern pairs of articles rather than single articles. This feature of the estimands introduces a complication as all pairs of articles between two media organizations cannot ostensibly be considered to be statistically independent. For example, if a researcher defines an event at the most granular level so that each media organization publishes at most one article about each event, then article A_i^k and A_j^l covering the same event implies A_i^k and $A_j^{l'}$ covering different events for all $l' \neq l$. This observation motivates the following iterative procedure pairing articles between media i and media j before subsequent statistical analysis.

The procedure takes a particular form of nearest-neighbor matching without replacement. It starts by computing a similarity measure $s(A_i^k, A_j^l)$ between each pair of news articles (A_i^k, A_j^l), $k = 1, \ldots, K$, $l = 1, \ldots, L$. The method does not require a particular similarity measure to be chosen by the researcher. Examples of similarity measures include similarity based on structural topic models (Roberts et al., 2020) and cosine similarity over a Term-Document-Matrix-based (TDM-based) representation (Mozer et al., 2020). The procedure then finds the

pair with the highest similarity, i.e., (A_i^k, A_j^l) such that $s(A_i^k, A_j^l) > s(A_i^{k'}, A_j^{l'})$ for all $(k', l') \neq (k, l)$. I shall call (A_i^k, A_j^l) a candidate pair. After removing A_i^k and A_j^l from the pool of articles, the procedure repeats the previous step among pairs of remaining articles until exhausting all articles published by at least one media organization. This initial pairing procedure yields the candidate pairs of news articles $(A_i^{k_1}, A_j^{l_1}), \ldots, (A_i^{k_N}, A_j^{l_N})$. The method estimates $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$ by $\mathbb{E}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n})$.

While matching procedures such as nearest-neighbor matching have been used extensively to adjust for confounding or nonparametric preprocessing (Ho et al., 2007), it serves an entirely different purpose here. It is employed here to pick up pairs of news articles that (1) potentially (but not necessarily) cover the same event, and (2) do not involve the same article, and hence are plausibly independent conditional on observable characteristics. The following two propositions illustrate the theoretical properties of this procedure. The first proposition concerns a simple case where events are defined at the most granular level so that each media organization publishes at most one article about each event.

Proposition 1. *Suppose* $E_i^k = E_j^l$ *implies* $s(A_i^k, A_j^l) > s(A_i^k, A_j^{l'})$ *for all* $l' \neq l$ *and* $s(A_i^k, A_j^l) > s(A_i^{k'}, A_j^l)$ *for all* $k' \neq k$.

Then for all (k, l) *such that* $(k, l) \notin \{(k_1, l_1), \ldots, (k_N, l_N)\}$, $E_i^k \neq E_j^l$ *holds.*

In words, the assumption says for each article A_i^k by media i, among media-j articles, the only article that potentially covers the same event is the one most similar to it (and similarly for each article A_j^l by media j). Under this assumption, the initial pairing procedure picks up precisely the pairs of articles that potentially cover the same event, and hence $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l] =$

$\mathbb{E}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n})$.

When events are defined at a broader level, the assumption in Proposition 1 does not hold. In fact, it is easy to see that no pairing procedure can pick up all pairs of articles that potentially cover the same event while ensuring these pairs only involve different articles.[3] In this case the initial pairing procedure, under a mild assumption, picks up no pair that cannot potentially cover the same event while pairing articles covering the same event in descending order of similarity, as the following proposition shows.

Proposition 2. *Suppose* $E_i^k = E_j^l$ *and* $E_i^k \neq E_j^{l'}$ *imply* $s(A_i^k, A_j^l) > s(A_i^k, A_j^{l'})$ *and* $E_i^k = E_j^l$ *and* $E_i^{k'} \neq E_j^l$ *imply* $s(A_i^k, A_j^l) > s(A_i^{k'}, A_j^l)$.

Then for all (k,l) *such that* $(k,l) \notin \{(k_1,l_1),\ldots,(k_N,l_N)\}$*, one of the following holds: (1)* $E_i^k \neq E_j^l$*; (2)* $E_i^k = E_j^l$*, and for some* l'*,* $(k,l') \in \{(k_1,l_1),\ldots,(k_N,l_N)\}$ *with* $s(A_i^k, A_j^{l'}) > s(A_i^k, A_j^l)$*; (3)* $E_i^k = E_j^l$*, and for some* k'*,* $(k',l) \in \{(k_1,l_1),\ldots,(k_N,l_N)\}$ *with* $s(A_i^{k'}, A_j^l) > s(A_i^k, A_j^l)$.

In words, the assumption states for each article A_i^k by media i, among media-j articles, the ones that cover the same event are more similar to it than those covering different events (and similarly for each article A_j^l by media j). Under this assumption, if a pair of articles covering the same event is not selected in this initial pairing procedure, then at least one of them is paired with another more similar article covering the same event. In this case, the selection of pairs among articles covering the same event induces a wedge between $\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$ and $\mathbb{E}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n})$. The wedge depends on f and the number of articles associated with each event. In the case of media bias where $f(Y_i^k, Y_j^l) \equiv Y_i^k - Y_j^l$,

[3]To see this, suppose media i's articles A_i^k and $A_i^{k'}$ and media j's articles A_j^l and $A_j^{l'}$ cover the same event. Then any pairing procedure yielding all pairs of articles that potentially cover the same event, which necessarily includes (A_i^k, A_j^l), $(A_i^k, A_j^{l'})$, $(A_i^{k'}, A_j^l)$, $(A_i^{k'}, A_j^{l'})$, features the same articles in candidate pairs.

the wedge will be smaller if the ideological leanings of a media organization's multiple articles on the same event are similar. The wedge will also be smaller if media i and media j publish a similar number of articles on events.

Metadata and substantive knowledge about the articles, such as timestamps and authors, can be incorporated in this initial pairing stage. For example, with a compatible definition of an event, researchers may plausibly assume that articles on different days cover different events. In the main application, I make this assumption to utilize the published date by pairing articles within a day. For other applications, researchers can also incorporate metadata into the similarity measure and implement a multidimensional version of the initial pairing procedure.

3.5 Estimation

The candidate pairs of news articles $(A_i^{k_1}, A_j^{l_1}), \ldots, (A_i^{k_N}, A_j^{l_N})$ yielded by the initial pairing procedure potentially but not necessarily cover the same event, and do not involve the same article in two distinct pairs. The method proceeds by an intuitive Bayes' rule calculation. To do so, let $m_n \equiv m(A_i^{k_n}, A_j^{l_n}) \equiv \mathcal{I}(E_i^{k_n} = E_j^{l_n})$ denote the indicator for articles $A_i^{k_n}$ and $A_j^{l_n}$ covering the same event, $s_n \equiv s(A_i^{k_n}, A_j^{l_n})$ the similarity measure as before, and $y_n \equiv f(Y_i^{k_n}, Y_j^{l_n})$ the outcome of interest. In the case of media bias, $y_n = Y_i^{k_n} - Y_j^{l_n}$, the difference in ideological leanings between articles $A_i^{k_n}$ and $A_j^{l_n}$. While s_n and y_n are observable for all candidate pairs, m_n is unobservable without hand-coding. With these notations, Table 2 displays the data structure after the initial pairing procedure.

	1	2	3	...	n	$n+1$	...	N
m	NA	NA	NA	...	NA	NA	...	NA
s	s_1	s_2	s_3	...	s_n	s_{n+1}	...	s_N
y	y_1	y_2	y_3	...	y_n	y_{n+1}	...	y_N

Table 2: Data Structure after Initial Pairing Procedure

Using Bayes' rule, I can calculate $\mathbb{E}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n})$ as

$$\begin{aligned}\mathbb{E}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n}) &= \mathbb{E}(y|m=1) \\ &= \frac{\int y \cdot f(y, m=1)\,\mathrm{d}y}{\Pr(m=1)} \\ &= \frac{\int y \cdot \int f(y, m=1, s)\,\mathrm{d}s\,\mathrm{d}y}{\int f(m=1, s)\,\mathrm{d}s} \\ &= \frac{\int\int y \cdot \Pr(m=1|s, y) f(s, y)\,\mathrm{d}s\,\mathrm{d}y}{\int \Pr(m=1|s) f(s)\,\mathrm{d}s},\end{aligned} \tag{1}$$

with a natural estimator given by its sample analogue

$$\hat{\mathbb{E}}(f(Y_i^{k_n}, Y_j^{l_n})|E_i^{k_n} = E_j^{l_n}) = \hat{\mathbb{E}}(y|m=1) = \frac{\sum_{n=1}^{N} y_n \hat{\Pr}(m_n = 1|s_n, y_n)}{\sum_{n=1}^{N} \hat{\Pr}(m_n = 1|s_n)}. \tag{2}$$

Notice that in Equation 2, both $\Pr(m=1|s)$, the probability of two articles covering the same event conditional on a given level of similarity, and $\Pr(m=1|s,y)$, the probability conditional on similarity s and outcome y, need to be estimated from the data. To estimate these quantities, the researcher selects a subset of news articles, say $\{1, \ldots, n\}$, and determines whether each pair covers the same event, and Table 3 displays the data structure with this labeled subset.

With the labeled subset in place, $\Pr(m=1|s)$ and $\Pr(m=1|s,y)$ can be estimated either nonparametrically by local polynomials or regression splines,

	1	2	3	...	n	$n+1$	...	N
m	m_1	m_2	m_3	...	m_n	NA	...	NA
s	s_1	s_2	s_3	...	s_n	s_{n+1}	...	s_N
y	y_1	y_2	y_3	...	y_n	y_{n+1}	...	y_N

Table 3: Data Structure with Labelled Subset

or parametrically by logistic regressions. When estimating these quantities nonparametrically, imposing plausible assumptions such as monotonicity in s and y can improve efficiency.

While selecting the subset to label via simple random sampling suffices to give a consistent estimate, stratified sampling is preferable for estimation efficiency. This consideration is particularly relevant when a large fraction of candidate pairs have similarity close to one or zero, as these pairs almost always or never cover the same event, respectively.

3.6 Comparison to Deterministic Matching

Another way to understand my method is to compare it to a deterministic matching procedure employed in the literature (Mozer et al., 2020). The deterministic matching procedure also starts by computing a similarity measure $s(A_i^k, A_j^l)$ between each pair of news articles. It proceeds by, for each article A_i^k, finding the article by media j that is most similar to it, i.e., A_j^l such that $s(A_i^k, A_j^l) > s(A_i^k, A_j^{l'})$ for all $l' \neq l$. The deterministic matching procedure declares (A_i^k, A_j^l) to be a match if and only if $s(A_i^k, A_j^l) \geq \underline{s}$ for some prespecified threshold $\underline{s}$, which could be $-\infty$. It estimates the quantity of interest by taking the sample analog among matched pairs. In other words, the deterministic matching procedure estimates

$\mathbb{E}[f(Y_i^k, Y_j^l)|E_i^k = E_j^l]$ by $\mathbb{E}[f(Y_i^k, Y_j^l)|s(A_i^k, A_j^l) > \max(s(A_i^k, A_j^{l'}), \underline{s}), \forall l' \neq l]$.[4]

The deterministic matching procedure has the advantage of saving the time-consuming process of hand-coding. Bias, however, is built in the procedure as it neglects the correlation between similarity and the outcome of interest. In other words, among pairs of news articles covering the same event, the ones with higher similarity are not a random subset. In the extreme, in the case of media bias, the almost identical ones are by construction very similar in terms of ideological leaning.

The threshold $\underline{s}$ is a critical tuning parameter in the implementation of the deterministic matching procedure. In practice, researchers have set $-\infty$ or arbitrary percentiles of the distribution of text similarity between all pairs of articles as the threshold. The choice of $\underline{s}$, however, governs the direction and magnitude of the bias of the estimate. A stringent matching criterion with a high threshold leads to misses of pairs of news articles covering the same event. Among true matches, only those that are sufficiently similar are detected. A slack matching criterion with a low threshold instead picks up false matches, i.e., pairs of articles covering different stories, biasing the estimate to the quantity unconditional on the event covered.

In the application, it will become clear that the distributions of similarity between pairs of articles are smooth. For an intermediate range of similarity, some pairs cover the same event while others do not. Instead of a zero-one coding depending on whether the similarity measure is a larger than an arbitrary threshold, my method assigns a probability to each candidate pair covering the

[4]This expression assumes the researcher starts with media organization i and for each article A_i^k, finds the article by media j that is most similar to it. If the researcher starts with media j instead, then the expression becomes $\mathbb{E}[f(Y_i^k, Y_j^l)|s(A_i^k, A_j^l) > \max(s(A_i^{k'}, A_j^l), \underline{s}), \forall k' \neq k]$. Alternatively, Mozer et al. (2020) starts with a sample of articles from each media organization.

same event and incorporate these probabilities to obtain a consistent estimate via a Bayes' rule calculation. Unlike the deterministic matching procedure, it statistically takes into account the correlation between similarity and the outcome of interest conditional on whether two articles cover the same event.

4. Application: Media Bias in the U.S.

4.1 Media Bias in the U.S.

Scholars, journalists, political elites, and voters have continuously engaged in debates about partisan bias in the news. According to Pew (2018), 68% of U.S. adults think news organizations tend to favor one side when presenting the news on political and social issues, and only 30% believe all sides are dealt with fairly.

In a comprehensive review of the partisan media bias in presidential election campaigns, D'Alessio and Allen (2000) theorize media bias to be of three types: gatekeeping bias, coverage bias, and statement bias. A closely related definition, yet more applicable to political news coverage generally, is put forth by Groeling (2013). Based on his categorization, "selection bias" refers to choosing news stories that favor one party over the other. In contrast, "presentation bias" involves composing news stories that skew the content of those resulting stories. If a viewer follows the complete news coverage from a single media outlet, then they are subject to the total bias. However, as I argued above, the impact of these two components will, in general, interacts with the way different consumers of news acquire political information. Some previous works have looked at "presentation bias" or "selection bias" in particular topic domains: Morris and Francia (2010) compare Fox News and CNN's coverage of the 2004 national party conventions;

Larcinese et al. (2011) focus on coverage of economic statistics and compare which stories are produced by which news organizations.

Adopting a different approach, Budak et al. (2016) recruit a large number of Amazon Mechanical Turk workers to assign partisanship scores to a sample (9%) of news articles pertaining to political events, where the pool of political news articles was identified via a supervised learning algorithm from all news articles published by 15 news sites in 2013.[5] The scores are on a scale of one to five, capturing the articles' favorability toward the Democratic party and the Republican party, respectively. They find that the reportings of U.S. politics by major news outlets online are considerably more similar than generally thought.

4.2 Text-Based Measure of Article Ideological Leanings

In their seminal work, Gentzkow and Shapiro (2010) measure the ideological leanings of newspapers by comparing phrase frequencies in the newspapers with phrase frequencies in congressional speeches by congresspeople. Martin and Yurukoglu (2017) adopt this approach to measure ideologies of major cable news channels (CNN, Fox News, and MSNBC). The text-based measure of ideological leanings at the media level has the advantage of scalability. It also avoids the subjectivity problem involving a priori determinations of what constitutes favorable coverage of a particular party associated with traditional content analyses Groeling (2013).

Fitting a predictive model of ideologies based on phrase frequencies in congressional speeches and applying it to news articles, however, is considerably

[5] Budak et al. (2016) describe the sampling procedure as: "[s]pecifically, for every day in 2013, we randomly selected two political articles, when available, 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."

more difficult at the article level than at the media level. Instead, I build on the core idea of this text-based approach—comparing the news articles to congressional speeches—but construct a different text-based measure of ideological leanings at the article level for media organizations. In particular, I consider the posterior probability that an observer with a neutral prior assigns to the speaker being a Republican after hearing the utterance of a single partisan phrase as if the article were a speech given by a congressperson. This measure is inspired by an idea recently proposed by Gentzkow et al. (2019), where they define partisanship in congressional speeches to be the ease at which an observer could infer a congressperson's party from a single utterance.

To do so, I first downloaded the congressional record of the 115th Congress, extracted non-procedural speeches given by U.S senators and representatives, and processed them to obtain phrase counts of each congressperson. Details of these steps are in Section A.2. Let q_p^D and q_p^R be the probabilities that a speaker affiliated with the Democratic party and the Republican party uses phrase p, respectively. Denote $\rho_p \equiv q_p^R/(q_p^R + q_p^D)$ to be the posterior belief that an observer with a neutral prior assigns to a speaker being Republican if the speaker utters phrase p. Following Gentzkow and Shapiro (2010) and Martin and Yurukoglu (2017), I restrict attention to 2,000 phrases with the highest χ^2-statistic among those satisfying minimal occurrence conditions in the congressional record and

news articles.[6][7] I then followed a similar procedure to obtain phrase counts of each political news article. Denoting by q_{ip}^k the probability that article A_i^k uses phrase p among the 2,000 partisan phrases, the proposed measure Y_i^k of ideological leaning of article k published by media i takes the form $Y_i^k \equiv \sum_{j=1}^{1000} q_{ip}^k \rho_p$.

4.3 Implementation

The total media bias $\mathbb{E}(Y_i^k - Y_j^l)$, unconditional on the event under coverage, can be straightforwardly estimated by taking the sample analog $\sum_{k\in K} Y_i^k - \sum_{l\in L} Y_j^l$. To obtain the presentation bias $\mathbb{E}[Y_i^k - Y_j^l | E_i^k = E_j^l]$, the difference in ideological leanings between media i and j conditional on covering the same event, my method proceeds in two steps laid out in Section 3. I shall next explain the particular choices made in the pairing and estimation steps.

Following the same procedure used to get the phrase counts, I obtain the word counts of each political news article. I choose the cosine distance over the term frequency-inverse document frequency (TF-IDF) weighted term-document matrix (TDM) representation as the similarity measure between each pair of articles.[8] To take advantage of available article metadata on the initial publication

[6]The Gentzkow-Shapiro χ^2-statistic is a test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans, if phrase frequencies in congressional speeches by congresspeople are drawn from party-specific multinomial distributions. To be precise, let f_p^D and f_p^R denote the total number of times Democrats and Republicans use phrase p, respectively, and $f_{\sim p}^D$ and $f_{\sim p}^R$ denote the total occurrence of phrases that are not phrase j spoken by Democrats and Republicans, respectively. Then

$$\chi_j^2 = \frac{(f_p^R f_{\sim p}^D - f_p^D f_{\sim p}^R)^2}{(f_p^R + f_p^D)(f_p^R + f_{\sim p}^R)(f_p^D + f_{\sim p}^D)(f_{\sim p}^R + f_{\sim p}^D)}.$$

[7]The minimal occurrence conditions imposed on phrases are: (1) spoken at least ten times during the 115th congress, and (2) used at least three times by each media organizations during the sample period.

[8]To be precise, let f_{iw}^k be the total number of times word w appears in article A_i^k, known as the

date, I form candidate pairs within a day in the initial pairing procedure.

To estimate $\hat{\mathbb{E}}(Y_i^{k_n} - Y_j^{l_n} | E_i^{k_n} = E_j^{l_n}) = \sum_{n=1}^{N} y_n \hat{\Pr}(m_n = 1|s_n, y_n) \cdot [\sum_{n=1}^{N} \hat{\Pr}(m_n = 1|s_n)]^{-1}$, I draw a random sample and label whether each sample candidate pair covers the same event. To get $\hat{\Pr}(m_n = 1|s_n)$ and $\hat{\Pr}(m_n = 1|s_n, y_n)$, I adopt logistic specifications, including the difference in ideological leanings, the difference squared and their interactions with text similarity in the latter case.

4.4 Results

I compare ideological leanings between political news articles published online by three pairs of media organizations: cable news CNN and Fox News, newspapers New York Times and Wall Street Journal, and news websites HuffPost and Breitbart between May 2018 and April 2019. I also conduct two comparisons across media types, between the *New York Times* with CNN and Fox News.

Table 4: Total Differences in Ideological Leanings

	N	N with Score	Mean Score	Std. Error	Median Score
CNN	11276	11225	0.394	0.001	0.383
Fox News	8734	8707	0.404	0.001	0.395
NYT	4093	4091	0.349	0.001	0.337
WSJ	2185	2183	0.361	0.002	0.348
HuffPost	8317	8272	0.376	0.001	0.366
Breitbart	12081	11738	0.429	0.001	0.421

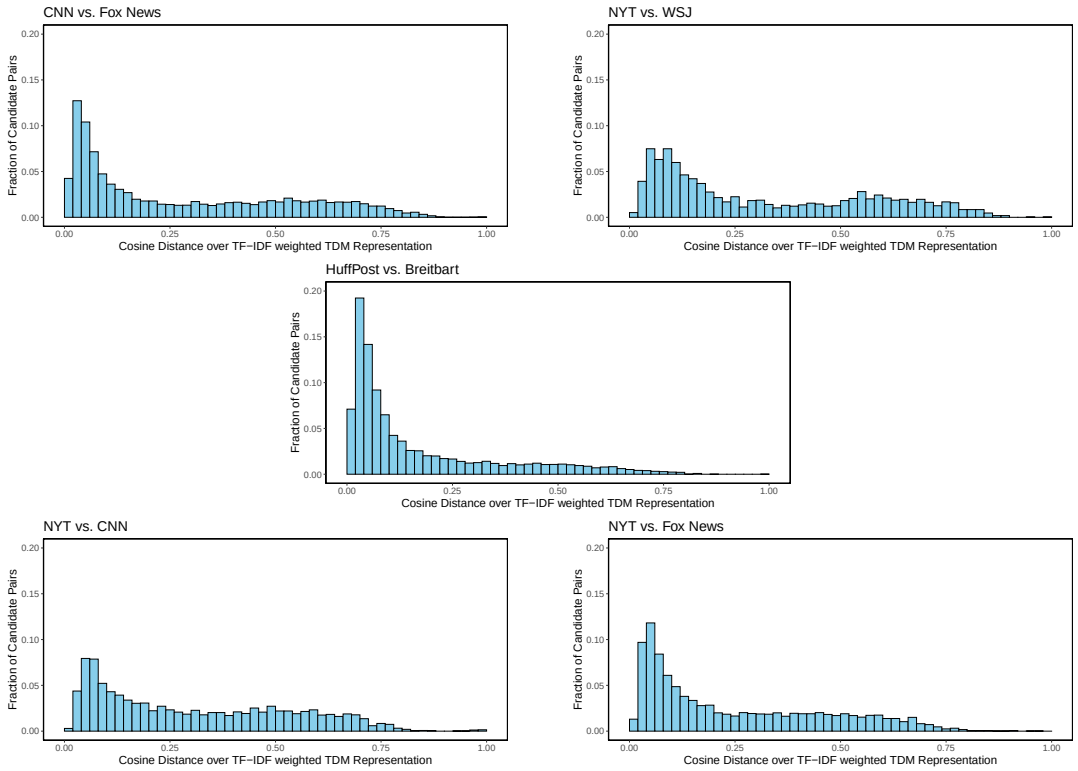
"term frequency". Denote by D_{total} the total number of articles and by D_w the number of articles containing word w. Then the "inverse document frequency" idf_w is given by $\log(D_{\text{total}}/D_w)$. Multiplying term frequency f_{iw}^k by inverse document frequency idf_w yields TF-IDF $\tilde{f}_{iw}^k$, which can be collected to a vector $\tilde{\boldsymbol{f}}_i^k$. The cosine distance over the TF-IDF weighted TDM representation is computed as $s(A_i^k, A_j^l) \equiv \cos(\tilde{\boldsymbol{f}}_i^k, \tilde{\boldsymbol{f}}_j^l)$.

Table 4 displays the summary statistics of the ideological leanings of political news articles according to the text-based measure. Newspapers, the *New York Times* and the *Wall Street Journal*, have fewer but longer articles on their websites than both cable news networks and online news sites. Consistent with previous research, I find, on average, articles published by Fox News, the *Wall Street Journal*, and Breitbart are ideologically located to the right of CNN, New York Times, and HuffPost, respectively. This pattern is not driven by outliers given a similar pattern in terms of median scores of ideological leanings. An average *Wall Street Journal* news article is located on the left side of the political spectrum according to the ideology measure, which is consistent with some previous research (Groseclose and Milyo, 2005; Flaxman et al., 2016), but not consistent with others (Gentzkow and Shapiro, 2010).[9]

To quantify presentation bias, I start by forming candidate pairs of political news articles from each pair of media organizations according to the cosine distance over the TF-IDF weighted TDM representation. Figure 1 displays the distributions of text similarity of these candidate pairs. The three histograms on the top show that the distribution of candidate pairs from HuffPost vs. Breitbart has the largest mass on low similarity, and the distribution associated with NYT vs. WSJ has the largest mass on high similarity, while CNN vs. Fox News lies somewhere in between. This order is aligned with the magnitude of the total difference in ideological leanings, indicating a correlation between text similarity and difference in ideological leanings. This observation is corroborated by

[9]The context is worth pointing out here. First, the conventional wisdom of the *Wall Street Journal*'s right-leaning comes primarily, if not entirely, from its opinion pieces (e.g., Budak et al. 2016). I, however, am interested in the partisan bias in each media organization's news contents and *not* opinion contents in this paper. Second, as discussed earlier, cable news networks like CNN and online news sites like HuffPost tend to cover many more news stories than newspapers, and some of these minor news stories are less partisan. This concern, however, is minimized as I focus mainly on pairs of media organizations of the same type.

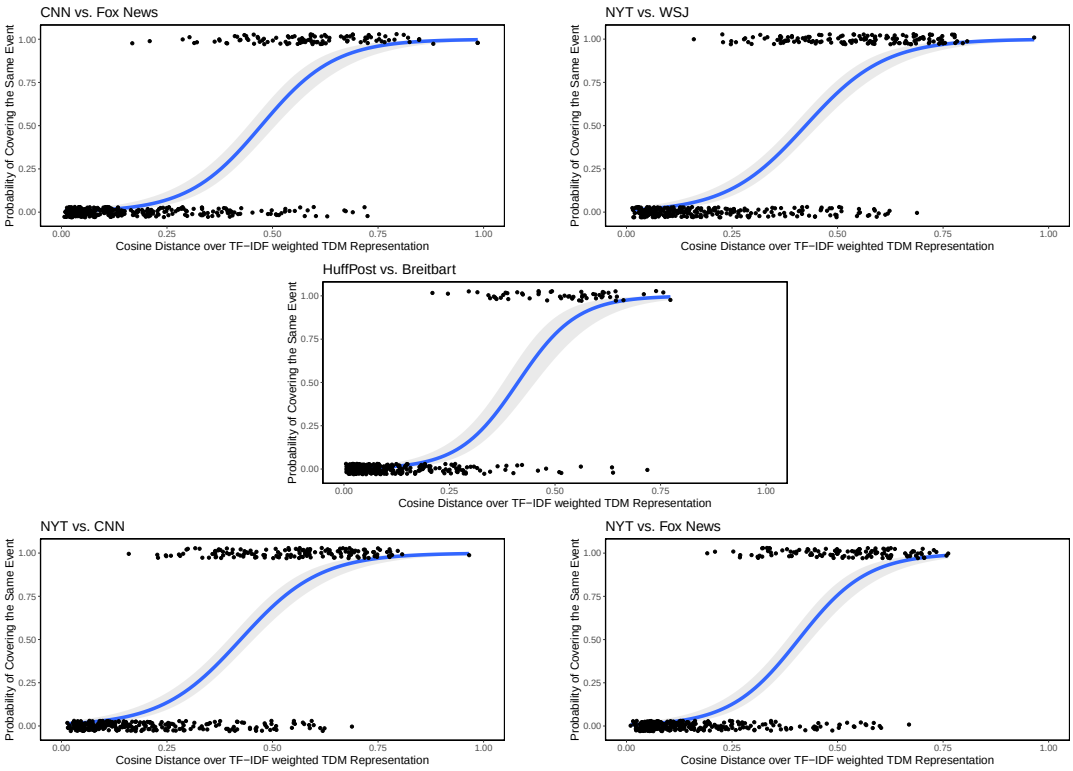
Figure 1: Distribution of Similarity of Candidate Pairs



Note: The figure displays the distributions of text similarity between candidate pairs of political news articles from five pairs of media organizations (CNN vs. Fox News, NYT vs. WSJ, HuffPost vs. Breitbart, NYT vs. CNN, NYT vs. Fox News). Text similarity is defined as the cosine distance over the TF-IDF weighted TDM representation.

comparing NYT against CNN and Fox News, respectively, shown on the bottom of the same figure. Moreover, all histograms suggest smooth distributions of text similarity between candidate pairs of political news articles. This feature implies a threshold such that pairs of articles with similarity above it are coded as covering the same event, if at all possible to pick systematically, is likely to be arbitrary. Furthermore, the correlation between text similarity and the outcome variable further implies the threshold, however picked, dictates the estimates in a deterministic matching procedure.

Figure 2: Labelled Candidate Pairs and Estimated Probability of Same Event



Note: The figure displays text similarity and the assigned label to each sampled candidate pair of political news articles. The assigned label indicates whether a candidate pair covers the same event. The probability of a candidate pair of articles covering the same event conditional on a given level of similarity is estimated and overlaid in the figure.

Instead of picking an arbitrary threshold, I proceed by sampling and labeling a subset of candidate pairs to estimate the probability of a pair covering the same event conditional on any given level of similarity. Figure 2 displays the text similarity of each candidate pair in the sample and whether the pair covers the same event. Using the labeled subset as training data, I estimate the probability of a candidate pair of articles covering the same event $\Pr(m = 1|s)$ and $\Pr(m = 1|s, y)$ via logistic regressions, with the fitted curve for $\hat{\Pr}(m = 1|s)$ also shown in Figure 2. For pairs of articles with high text similarity, they almost always cover

the same event. For articles written in very different words, corresponding to text similarity smaller than 0.25, the covered events are almost never the same. There is an intermediate range of text similarity, however, where some pairs falling in this range cover the same event while others do not. This observation highlights the importance of assigning a probability, instead of zero or one, to each candidate pair in order to statistically account for the correlation between text similarity and the outcome variable in the estimation for presentation bias.

Table 5: Differences in Ideological Leanings due to Presentation Bias

	Difference due to Presentation	Bootstrap Std. Error	Total Difference
CNN vs. Fox News	0.003	0.001	0.010
NYT vs. WSJ	0.016	0.001	0.012
HuffPost vs. Breitbart	0.035	0.004	0.054
NYT vs. CNN	0.042	0.002	0.045
NYT vs. Fox News	0.056	0.004	0.055

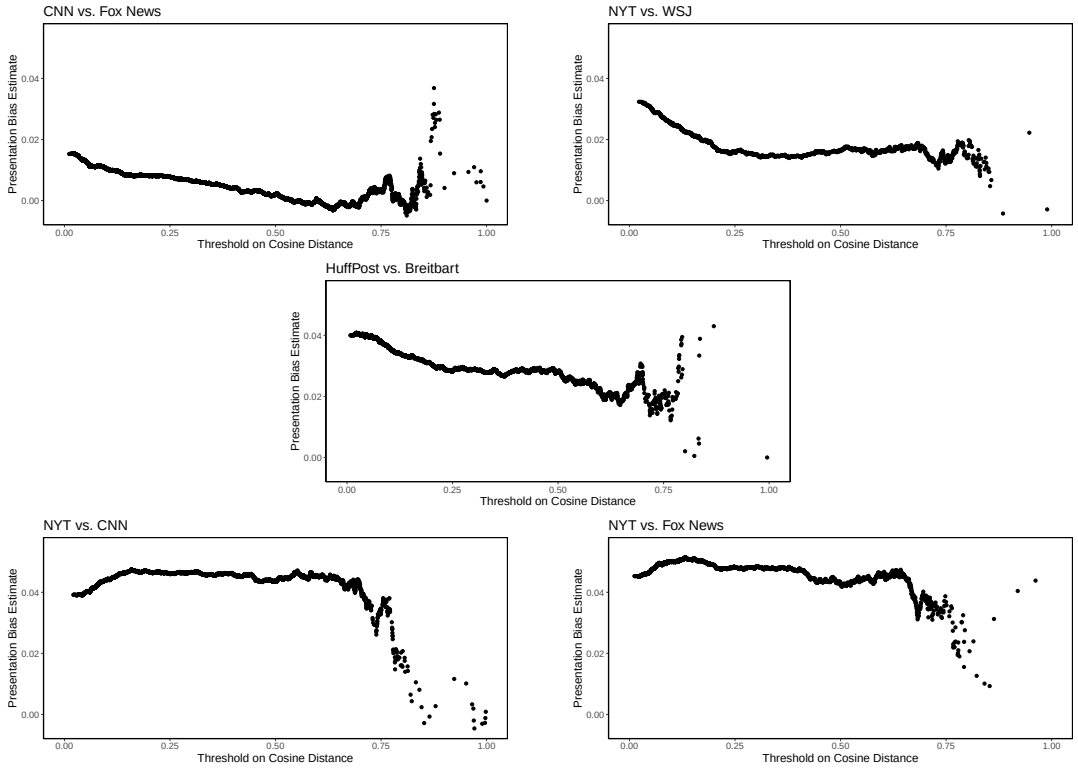
With $\hat{\Pr}(m = 1|s)$ and $\hat{\Pr}(m = 1|s, y)$ at hand, $\hat{\mathbb{E}}(Y_i^{k_n} - Y_j^{l_n}|E_i^{k_n} = E_j^{l_n})$ can be constructed via an intuitive Bayes' rule calculation given by Equation (1)-(2). Table 5 displays the estimates for the difference in ideological leanings conditional on covering the same event $\hat{\mathbb{E}}(Y_i^{k_n} - Y_j^{l_n}|E_i^{k_n} = E_j^{l_n})$, in comparison to the estimates for the unconditional ideological difference $\mathbb{E}(Y_i^k - Y_j^l)$. The fraction of the difference in ideological leanings attributable to presentation bias differs across media types. While presentation bias drives the ideological difference in entirety between newspapers the *New York Times* and the *Wall Street Journal*, it accounts only for a third of the total difference in ideological

leanings between cable news CNN and Fox News. As most articles end up in the printed version of the newspapers, the *New York Times* and the *Wall Street Journal* have fewer but longer articles. As a result, they both tend to focus the coverage on the most salient news stories, which limits the scope of selection bias. Longer articles with background information and analysis, meanwhile, allows more room for different framing and assembling of supplementary contents. By contrast, cable news CNN and Fox News publish more articles, providing more opportunities for selective coverage. However, since news articles by these major cable news networks commonly share basic information about the underlying event as well as direct quotes from politicians, the shorter article length leaves less room for partisan presentation. Online-only sites HuffPost and Breitbart differ a lot in both presentation and selection, with presentation bias contributes about two thirds of the total ideological gap.

It is also informative to see how a deterministic matching procedure will perform in this case. As described in Section 3.6, to compare two media organizations i and j, for each article published by media organization i, the media j article most similar to it is selected to form a pair. Then for a prespecified threshold, pairs with text similarity higher than the threshold are declared to be matches. In a deterministic matching procedure, the difference in ideological leanings conditional on covering the same event is calculated as the average ideological differences between matched pairs.

Figure 3 displays the deterministic matching estimates for presentation bias given different thresholds on cosine distance over the TF-IDF weighted TDM representation, where the threshold is a tuning parameter left to the researchers to decide. When the threshold is chosen to be larger than 0.7, the estimate is very

Figure 3: Deterministic Matching Estimates for Presentation Bias for Different Thresholds on Cosine Distance



Note: The figure displays the deterministic matching estimates for the difference in ideological leanings conditional on covering the event for different thresholds on text similarity between five pairs of media organizations (CNN vs. Fox News, NYT vs. WSJ, HuffPost vs. Breitbart, NYT vs. CNN, NYT vs. Fox News). Text similarity is defined as the cosine distance over the TF-IDF weighted TDM representation.

sensitive to the exact choice as the estimate is based on a relatively small number of pairs of articles. For some pairs of media organizations (CNN vs. Fox News, NYT vs. WSJ, and HuffPost vs. Breitbart), the estimates are decreasing in the threshold choice. In this case, a stringent threshold leads to the conclusion of a small presentation bias, whereas a slack threshold yields the opposite conclusion. For other pairs of media organizations (NYT vs. CNN and NYT vs. Fox News), the difference in ideological leanings does not seem much correlated with text

similarity. This observation is consistent with my finding that media bias is almost entirely driven by presentation bias for these comparisons.

5. Discussion and Conclusion

Text data such as corpora of news articles are organized by characteristics that are unavailable to researchers absent large-scale hand-coding. Estimating substantive quantities of interests depending on such hidden structure poses a methodological issue to applied researchers. In this paper, I proposed a method to estimate quantities of interest which are conditional on a hidden characteristic. The proposed method overcomes the difficulties of existing methods that are sensitive to researcher-specified tuning parameters and produce biased estimates.

I applied my method to engage in a substantive debate on the sources of media bias among U.S. media organizations. My results indicate that partisan bias enters political news coverage not always through how information for a news story is framed and presented by media organizations. Even though partisan framing and presentation is the dominant source of bias for newspapers such as the *New York Times* and the *Wall Street Journal*, selective coverage that reflects liberal or conservative talking points accounts for a significant fraction of overall bias for cable news websites and online news sites. My results shed light on the exposure of voters with different news consumption habits to the partisan bias of political news coverage.

The method can be readily applied to address other substantive questions in the study of media. In a separate paper, I study the difference in restrictions on online discussion of news stories by state-run media versus non-state media

in China. A naive analysis would conclude a heavier restriction by state media, but this neglect the fact that state media cover more politically sensitive news and thus have more appearances of discussion censoring. It is, therefore, crucial to condition on the same event to study restrictions on discussion of politics in authoritarian countries. The method also has potential in areas such as policy diffusion and judicial politics. For instance, scholars interested in leveraging the same policy proposals across state legislatures to scale the ideology of state legislators can estimate the probability of the same proposal being voted on given bill text similarity and incorporate it to obtain an unbiased estimate. The performance of the method proposed here relative to existing methods in these contexts warrants future work.

References

Budak, C., S. Goel, and J. M. Rao (2016). Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly 80*(S1), 250–271.

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

Flaxman, S., S. Goel, and J. M. Rao (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly 80*(S1), 298–320. Publisher: Oxford Academic.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring Group Differences in

High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica 87*(4), 1307–1340.

Grimmer, J. and G. King (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences 108*(7), 2643–2650.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Groeling, T. (2013). Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science 16*(1), 129–151.

Groseclose, T. and J. Milyo (2005). A Measure of Media Bias. *The Quarterly Journal of Economics 120*(4), 1191–1237.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis 15*(3), 199–236.

Larcinese, V., R. Puglisi, and J. M. Snyder (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *Journal of Public Economics 95*(9), 1178–1189.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. The Regents of the University of California. ISSN: 0097-0433.

Martin, G. J. and A. Yurukoglu (2017). Bias in Cable News: Persuasion and Polarization. *American Economic Review 107*(9), 2565–2599.

Morris, J. S. and P. L. Francia (2010). Cable News, Public Opinion, and the 2004 Party Conventions. *Political Research Quarterly 63*(4), 834–849. Publisher: SAGE Publications Inc.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. Jason Anastasopoulos (2020). Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis*.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020). Adjusting for Confounding with Text Matching. *American Journal of Political Science*.

A. Appendix

A.1 Proofs

Proof of Proposition 1. Suppose for some $(k,l) \notin \{(k_1,l_1),\ldots,(k_N,l_N)\}$, it holds that $E_i^k = E_j^l$. Since the procedure continues until exhausting all articles published by at least one media organization, it is without loss of generality to assume $k \in \{k_1,\ldots k_N\}$. Let $k = k_n$.

If $l = l_p$ for some $p < n$, then by the construction of the procedure, $s(A_i^{k_p}, A_j^l) > s(A_i^{k'}, A_j^l)$ for all $k' \notin \{k_1,\ldots,k_{p-1}\}$. Since k_1, …, k_N are distinct by the construction of the procedure and $k = k_n$ with $p < n$, $k \notin \{k_1,\ldots,k_{p-1}\}$ and in turn $s(A_i^{k_p}, A_j^l) > s(A_i^k, A_j^l)$. But this inequality is contradicted by $s(A_i^k, A_j^l) > s(A_i^{k''}, A_j^l)$ for all $k'' \neq k$ implied by $E_i^k = E_j^l$.

If $l \neq l_p$ for all $p < n$, then by the construction of the procedure, $s(A_i^k, A_j^{l_n}) > s(A_i^k, A_j^{l'})$ for all $l' \notin \{l_1,\ldots,l_n\}$. In particular, $s(A_i^k, A_j^{l_n}) > s(A_i^k, A_j^l)$. But this inequality is contradicted by $s(A_i^k, A_j^l) > s(A_i^k, A_j^{l''})$ for all $l'' \neq l$ implied by $E_i^k = E_j^l$. □

Proof of Proposition 2. Take any (k,l) such that $(k,l) \notin \{(k_1,l_1),\ldots,(k_N,l_N)\}$. If $E_i^k \neq E_j^l$, i.e. (1) holds, then the proof is concluded. Now suppose $E_i^k = E_j^l$. Since the procedure continues until exhausting all articles published by at least one media organization, it is without loss of generality to assume $k \in \{k_1,\ldots k_N\}$. Let $k = k_n$.

If $l = l_p$ for some $p < n$, then by the construction of the procedure, $s(A_i^{k_p}, A_j^l) > s(A_i^{k'}, A_j^l)$ for all $k' \notin \{k_1,\ldots,k_{p-1}\}$. Since k_1, …, k_N are distinct by the construction of the procedure and $k = k_n$ with $p < n$, $k \notin \{k_1,\ldots,k_{p-1}\}$ and in turn $s(A_i^{k_p}, A_j^l) > s(A_i^k, A_j^l)$. Define $k' \equiv k_p$, then (3) holds.

If $l \neq l_p$ for all $p < n$, then by the construction of the procedure, $s(A_i^k, A_j^{l_n}) > s(A_i^k, A_j^{l''})$ for all $l'' \notin \{l_1, \ldots, l_n\}$. In particular, $s(A_i^k, A_j^{l_n}) > s(A_i^k, A_j^l)$. Define $l' \equiv l_n$, then (2) holds. □

A.2 Processing Congressional Record

I download the daily Congressional Record (the Record hereafter) of the 115th Congress from January 2017 to January 2019 using the govinfo Application Programming Interface made available by the Government Publishing Office. The Record contains four sections: Daily Digest, Senate, House of Representatives, and Extension of Remarks. To obtain speeches made by U.S. senators and congresspeople on the Senate floor and on the House floor respectively, I focus on the Senate section and the House of Representatives section of the Record.

I write a script that parses the Record to extract congressional speeches. Some parts of the Record are procedural or otherwise document information or activities containing no congressional speeches, and are hence dropped.[10][11] The

[10]For Senate, these parts include: ADDITIONAL COSPONSORS, ADJOURNMENT UNTIL [Time], ADMINISTRATION OF OATH OF OFFICE, AMENDMENTS SUBMITTED AND PROPOSED, AMENDMENTS SUBMITTED AND PROPOSED, APPOINTMENT, APPOINTMENT OF ACTING PRESIDENT PRO TEMPORE, APPOINTMENTS, ARMS SALES NOTIFICATION, AUTHORITY FOR COMMITTEES TO MEET, CERTIFICATE OF APPOINTMENT, CONCLUSION OF MORNING BUSINESS, CONFIRMATION, CONFIRMATIONS, CONGRESSIONAL RECORD, DISCHARGED NOMINATIONS, ENROLLED BILL PRESENTED, ENROLLED BILL SIGNED, EXECUTIVE AND OTHER COMMUNICATIONS, EXECUTIVE CALENDAR, EXECUTIVE MESSAGES REFERRED, EXECUTIVE REPORT OF COMMITTEE, EXECUTIVE REPORTS OF COMMITTEE, EXECUTIVE REPORTS OF COMMITTEES, EXECUTIVE SESSION, INTRODUCTION OF BILLS AND JOINT RESOLUTIONS, LEGISLATIVE SESSION, MEASURES PLACED ON THE CALENDAR, MEASURES REFERRED, MESSAGE FROM THE HOUSE, MESSAGE FROM THE PRESIDENT, MESSAGES FROM THE HOUSE, MESSAGES FROM THE HOUSE RECEIVED DURING ADJOURNMENT, MESSAGES FROM THE PRESIDENT, MORNING BUSINESS, NOMINATIONS, ORDER FOR ADJOURNMENT, ORDERS FOR [Date], PETITIONS AND MEMORIALS, PLEDGE OF ALLEGIANCE, PRAYER, PRESIDENTIAL MESSAGE, PRIVILEGES OF THE FLOOR, RECESS, RECOGNITION OF THE MAJORITY LEADER, RECOGNITION OF THE MINORITY LEADER, REPORTS OF COMMITTEES, RESERVATION OF LEADER TIME, STATEMENTS ON INTRODUCED BILLS AND JOINT RESOLUTIONS, SUBMISSION OF CONCURRENT AND SENATE RESOLUTIONS, SUBMITTED RESOLUTIONS, TEXT OF AMENDMENTS, and VOTE EXPLANATION.

[11]For House of Representatives, these parts include: ADDITIONAL SPONSORS, ADJOURNMENT, ANNOUNCEMENT BY THE SPEAKER, ANNOUNCEMENT BY THE SPEAKER PRO TEMPORE, BILL PRESENTED TO THE PRESIDENT, BILLS PRESENTED TO THE PRESIDENT, COMMUNICATION FROM THE CLERK OF THE HOUSE, COMMUNICATION FROM THE DEMOCRATIC LEADER, CONGRESSIONAL RECORD, CONSTITUTIONAL AUTHORITY STATEMENT, DELETIONS OF SPONSORS FROM PUBLIC BILLS AND RESOLUTIONS, DESIGNATION OF SPEAKER PRO TEMPORE, DESIGNATION OF THE SPEAKER PRO TEMPORE, DISCHARGE OF COMMITTEE, ENROLLED BILL SIGNED, ENROLLED BILLS SIGNED, EXECUTIVE COMMUNI-

beginning of speeches made by by U.S. senators and congresspeople is identified with one of the following: (1) a speaker's title and last name (or full name) in uppercase, followed by a period (e.g., Mr. CROWLEY.), or (2) a speaker's title, last name (or full name) in uppercase, and state, followed by a period (e.g., Mr. RYAN of Wisconsin.).[12] The end of congressional speeches is identified with one of the following: (1) the beginning of a new speech made by by U.S. senators and congresspeople as described above, (2) the beginning of a speech made by a titled speaker (e.g. The PRESIDING OFFICER.), (3) the end of a section of debate, (4) the insertion of a document in the Record, (5) the reading of a bill or other document, or (6) a vote.[13] The script detects congressional speeches according to these patterns, and drops purely procedural speeches that start with "Mr./Madam President/Speaker, I ask unanimous consent", "Mr./Madam Speaker, I move to suspend the rules", or "Mr./Madam Speaker, pursuant to House Resolution", or contain "The following Senators are necessarily absent" or "Had I been present, I would have voted", or otherwise consist of fewer than 300 characters. The script finally separates the speaker demarcations from the speech texts before further processing.

After extracting non-procedural speeches given by by U.S. senators and con-

CATIONS, ETC., EXPENDITURE REPORTS CONCERNING OFFICIAL FOREIGN TRAVEL, HOUR OF MEETING ON TOMORROW, HOUSE BILLS APPROVED BY THE PRESIDENT, LEAVE OF ABSENCE, MEMORIALS, MESSAGE FROM THE PRESIDENT, MESSAGE FROM THE SENATE, PERSONAL EXPLANATION, PETITIONS, ETC., PLEDGE OF ALLEGIANCE, PRAYER, PUBLIC BILLS AND RESOLUTIONS, PUBLICATION OF COMMITTEE RULES, RECESS, REPORTS OF COMMITTEES ON PUBLIC BILLS AND RESOLUTIONS, RESIGNATION FROM THE HOUSE OF REPRESENTATIVES, SENATE BILL REFERRED, SENATE BILLS APPROVED BY THE PRESIDENT, SENATE BILLS REFERRED, SENATE ENROLLED BILL SIGNED, and THE JOURNAL.

[12]Some U.S. senators and congresspeople have last names that constitute exceptions to this rule, e.g., Mr. McCONNELL, Mr. DeSANTIS, and Mr. LaMALFA. To consistently identify the beginnings of all speeches, I check whether the last letter of the last name is in uppercase.

[13]For Senate, titled speaker's positions include The PRESIDING OFFICER, The ACTING PRESIDENT pro tempore, The VICE PRESIDENT, and The PRESIDENT pro tempore. For House of Representatives, titled speaker's positions include The SPEAKER pro tempore, The Acting CHAIR, The ACTING CHAIR, The SPEAKER, The CHAIR, The VICE PRESIDENT, and The CLERK.

gresspeople, another script processes the speech texts in the following steps. First, the script converts the speech texts to lowercase. Second, all punctuation but hyphens and apostrophes are replaced with white spaces. Third, the script detects common stopwords and drops them.[14] Fourth, hyphens and apostrophes are dropped. Fifth, the script reduces words to their stems using Porter2 stemming algorithm (Porter 2009). Finally, the counts of bigrams are calculated for each speech.

Meanwhile, I identify the congressperson associated with each speech by matching the the speaker demarcations to information on congresspeople retrieved from voteview.com. A handful of speeches made by non-voting members and former members of the congress are first dropped. Almost all other speaker demarcations can be matched with VoteView data after correcting for typos and incongruent formats. I disambiguate the remaining speaker names by their gender and in three cases by manually inspecting the speeches.

[14]The set of stopwords is obtained from http://snowball.tartarus.org/algorithms/english/stop.txt retrieved on July 3, 2019.