AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Adjusting for Confounding with Text Matching

Margaret E. Roberts University of California, San Diego
Brandon M. Stewart Princeton University
Richard A. Nielsen Massachusetts Institute for Technology

Abstract: *We identify situations in which conditioning on text can address confounding in observational studies. We argue that a matching approach is particularly well-suited to this task, but existing matching methods are ill-equipped to handle high-dimensional text data. Our proposed solution is to estimate a low-dimensional summary of the text and condition on this summary via matching. We propose a method of text matching, topical inverse regression matching, that allows the analyst to match both on the topical content of confounding documents and the probability that each of these documents is treated. We validate our approach and illustrate the importance of conditioning on text to address confounding with two applications: the effect of perceptions of author gender on citation counts in the international relations literature and the effects of censorship on Chinese social media users.*

Verification Materials: The materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/HTMX3K.

Social media users in China are censored every day, but it is largely unknown how the experience of being censored affects their future online experience. Are social media users who are censored for the first time flagged by censors for increased scrutiny in the future? Is censorship "targeted" and "customized" toward specific users? Do social media users avoid writing after being censored? Do they continue to write on sensitive topics or do they avoid them?

Experimentally manipulating censorship would allow us to make credible causal inferences about the effects of experiencing censorship, but this is impractical and unethical outside of a lab setting. Inferring causal effects in *observational* settings is challenging due to confounding. The types of users who are censored might have different opinions that drive them to write differently than the types of users who are not censored. This in turn might affect both the users' rate of censorship as well as future behavior and outcomes. We argue that conditioning on the text of censored social media posts and other user-level characteristics can substantially decrease or eliminate confounding and allow credible causal inferences with observational data. Intuitively, if we can find nearly identical posts—one of which is censored while the

Margaret E. Roberts is Associate Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093-0521 (meroberts@ucsd.edu). Brandon M. Stewart is Assistant Professor and Arthur H. Scribner Bicentennial Preceptor, Department of Sociology, Princeton University, 149 Wallace Hall, Princeton, NJ 08544 (bms4@princeton.edu). Richard A. Nielsen is Associate Professor, Department of Political Science, Massachusetts Institute for Technology, 77 Massachusetts Avenue, E53 Room 455, Cambridge, MA 02139 (rnielsen@mit.edu).

We thank the following for helpful comments and suggestions on this work: David Blei, Naoki Egami, Chris Felton, James Fowler, Justin Grimmer, Erin Hartman, Chad Hazlett, Seth Hill, Kosuke Imai, Rebecca Johnson, Gary King, Adeline Lo, Will Lowe, Chris Lucas, Walter Mebane, David Mimno, Jennifer Pan, Marc Ratkovic, Matt Salganik, Caroline Tolbert, and Simone Zhang; audiences at the Princeton Text Analysis Workshop, Princeton Politics Methods Workshop, the University of Rochester, Microsoft Research, the Text as Data Conference, and the Political Methodology Society and the Visions in Methodology conference; and some tremendously helpful anonymous reviewers. We especially thank Dustin Tingley for numerous insightful conversations on the connections between STM and causal inference and Ian Lundberg for extended discussions on some technical details. Dan Maliniak, Ryan Powers, and Barbara Walter graciously supplied data and replication code for the gender and citations study. The JSTOR Data for Research program provided academic journal data for the international relations application. This research was supported, in part, by the Eunice Kennedy Shriver National Institute of Child Health and Human Development under grant P2-CHD047879 to the Office of Population Research at Princeton University. The research was also supported by grants from the National Science Foundation RIDIR program, award numbers 1738411 and 1738288. This publication was made possible, in part, by a grant from the Carnegie Corporation of New York, supporting Richard Nielsen as an Andrew Carnegie Fellow. The statements made and views expressed are solely the responsibility of the authors.

American Journal of Political Science, Vol. 64, No. 4, October 2020, Pp. 887–903

 DOI: 10.1111/ajps.12526

other is not—from similar users, we can compare downstream online behavior to obtain credible estimates of the effects of censorship.

Traditional matching methods are not suited to the task of conditioning on the text of documents. Quantitative analysts typically represent text using thousands or even millions of dimensions (e.g., the columns of a document term matrix, collections of word embeddings). Common matching techniques such as propensity score matching (Rosenbaum and Rubin 1983) and coarsened exact matching (CEM; Iacus, King, and Porro 2011) were developed for applications with fewer variables (dimensions) than observations in the data set. For example, Rubin and Thomas (1996, 249) note that in "typical examples" of matching, the number of variables is "between 5 and 50," whereas the number of observations is much larger. As the number of variables increases, the "curse of dimensionality" makes it difficult to find units similar on all dimensions. This poses a well-known problem for exact matching, which requires observations to match on all covariates and often fails to find any matches in high-dimensional settings (Rubin and Thomas 1996, 250). Existing methods that relax the requirement of exact matching (Iacus, King, and Porro 2011) or perform dimension reduction (e.g., by matching on Mahalanobis distances or propensity scores) can suffer from poor efficiency or fail entirely with high-dimensional data.

We propose a text-matching strategy that allows analysts to effectively and transparently address confounding captured in text. We make four contributions. First, we introduce a framework for using the content of text to address confounding in observational data. Second, we propose a general text-matching adjustment strategy that involves balancing both a low-dimensional density estimate of the data and a metric that captures the probability of treatment. This approach produces matches that capture aspects of text related to treatment and facilitates qualitative comparison and evaluation. Third, we design a specific procedure, topical inverse regression matching (TIRM), to match on a jointly estimated propensity for treatment and density estimate. We show that this procedure has strong performance in a simulation study.[1] Finally, we demonstrate how to apply text matching through two applications.

A strength of matching relative to other conditioning strategies is that analysts can evaluate the quality of the adjustment by reading treated documents alongside their matches. Comparing documents allows analysts to use substantive knowledge to recognize improved textual similarity in the matched sample even if they cannot formalize that knowledge a priori in a balance metric to be minimized in the matching procedure.[2] This type of human validation is an essential part of making comparisons in a high-dimensional and complex setting such as text (Grimmer and Stewart 2013).

Our first application extends the work of Maliniak, Powers, and Walter (2013), who find that perceived gender of international relations scholars affects citations to their articles. To address confounding, the authors controlled for article content with hand-coded variables. We show that applying TIRM recovers similar effect estimates without using the hand-coded data, suggesting that TIRM can be a viable alternative to measuring text confounders by hand. In our second application, we estimate how censorship affects the online experience of social media users in China.[3] We match censored social media users to uncensored social media users who write similar social media posts and have very similar censorship histories. We find that censored social media users are more likely to be censored again in the future, suggesting that either censors are flagging users when they are censored, that social media users write about more sensitive topics after censorship, or both.

In each application, we address confounding by conditioning on text, a solution we believe is broadly applicable across social science. Scholars of American politics could match legislative bills with similar content to estimate the effect of veto threats on repositioning in Congress. Scholars of race and ethnicity might match students with similar college admissions profiles and essays to estimate the effect of perceived race on college admissions. And scholars of international relations might condition on the content of international agreements when estimating the determinants of international cooperation. Our approach could also apply to nontext data in computer vision, population genetics, biological microarrays, and other areas where a generative model of pretreatment covariates can be reliably estimated, or when the observed

[1]Our primary contribution is to pose the problem of text-based confounding and offer TIRM as one possible solution. Since our paper started circulating in July 2015, Mozer et al. (2020) and Veitch, Sridhar, and Blei (2019) have introduced alternative approaches to text confounding. We hope that there will be further developments in this area.

[2]If the analyst can confidently define an a priori balance metric that captures confounding in text, they can directly optimize it using standard methods (Diamond and Sekhon 2013; Hainmueller 2011; Imai and Ratkovic 2014) while trading off between balance and sample size (King, Lucas, and Nielsen 2017). Our approach cannot obviate the need to (implicitly) choose a balance metric, but it does weaken reliance on the balance metric by facilitating human validation.

[3]A similar analysis is performed in Roberts (2018), who uses exact matches.

data are noisy measures of a latent confounder (Kuroki and Pearl 2014).

The article proceeds as follows. In the second section, we describe text-based confounding adjustment, explain why we opt for a matching approach, and define basic notation. We highlight the importance of conditioning on both a density estimate and a propensity score. In the third section, we present topical inverse regression matching as a way to jointly estimate the density and a propensity score. We also offer an approach to balance checking and discuss the method's strengths, limitations, and relation to prior work. In the fourth section, we detail our two applications: a validation study demonstrating the effect of perceived author gender on academic article citations and a study estimating the effect of being censored on the reactions of Chinese social media users. The last section concludes with a discussion of future directions.

Using Text to Address Confounding

We begin by describing the setting for which we develop our approach. To fix ideas, we use one of our applications—the effects of experiencing government censorship on Chinese social media users—as a running example. In this example, our goal is to answer two questions. First, are Chinese social media users who have a post censored more likely to be censored in subsequent posts? Second, does censorship decrease the number of future posts by a user? To answer both questions, we match censored bloggers to uncensored bloggers with similar posts and similar histories of censorship and posting. We use matching to identify censorship mistakes: similar posts by different authors where one post was censored and the other was not. We find that censorship increases the probability of future censorship, but it does not have an effect on the number of posts the user writes. This provides evidence that censorship is targeted toward users who are recently censored, but that it does not induce a chilling effect on the number of posts written.

We adopt the following notation for the confounding problem. We start with a data set of n units. Each unit i is assigned treatment T_i, which takes a value of 1 for treated units and 0 for control. Under the potential outcomes framework, the outcome variable Y_i takes on the value $Y_i(1)$ when unit i is treated and $Y_i(0)$ when unit i is a control. In the censorship case, the units are individual Chinese social media users, the treatment T_i is censorship, and the outcome Y_i is the subsequent censorship rate of the social media user.

Because we have observational data, T_i is not randomly assigned and treated and control groups may not be comparable. A common practice is to match on p pretreatment covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ to improve similarity in the distribution of covariates within treatment and control groups, a condition called balance. If we assume conditional ignorability, $T_i \perp\!\!\!\perp Y_i(0), Y_i(1) | \boldsymbol{X}$, then balancing the distribution of observed covariates $\boldsymbol{X}$ across treatment groups provides a way to estimate the average causal effect of T on Y If we were able to exactly match each treated unit to control units, we would estimate the average treatment effect on the treated by averaging the difference between the value of Y_i for the treated unit and the value of Y_i for the matched control units.[4]

In most matching applications, $\boldsymbol{X}$ is low-dimensional, with $p \ll n$. We consider cases where the potential dimensionality of p is very large. For example, censorship in social media posts may be a function of thousands of particular words, particular combinations of words, hierarchies of words, and so on. As is common in the text analysis literature (Grimmer and Stewart 2013), we represent each document in a sparse count matrix $\boldsymbol{W}$ whose typical element, W_{ij}, contains the number of times the jth word appears within the text associated with unit i. The $\boldsymbol{W}$ matrix has dimension n (number of documents) by v (number of unique words in the corpus). The data are high-dimensional in the sense that v is always large relative to n. This approach largely ignores word order and combinations, though $\boldsymbol{W}$ can be modified to include these features.

As with conditioning in low-dimensional settings, conditioning on the text represented by $\boldsymbol{W}$ is appropriate when $\boldsymbol{W}$ is pretreatment and conditioning on it will address confounding. However, if the text is a result of treatment, rather than its cause, conditioning on the text will induce posttreatment bias. Even when the text is pretreatment, conditioning on it can cause problems. For example, if the text is an instrument, meaning it explains the outcome only through the treatment, conditioning on the text can amplify the bias of unconditioned confounders (Pearl 2011). Consider the following directed acyclic graph, which contains some small, but unavoidable, unobserved confounding U.

[4]Typically, exact matches on all observed covariates $\boldsymbol{X}$ are not possible, so we match treated units to the closest control units and then estimate the average treatment effect on the treated (ATT) within matches. In situations where some treated units are not similar enough to any control units, researchers typically remove the unmatchable treated units. If so, the quantity estimated is the feasible sample average treatment effect on the treated (FSATT)—the ATT for the set of sampled treated units for which matches can be found.

$$\begin{tikzcd}
Z = \text{Instrument} \arrow[dr] & & U \arrow[dl] \arrow[dr] & \\
& T = \text{Treatment} \arrow[rr] & & Y = \text{Outcome} \\
V \arrow[ur] \arrow[urrr] & & &
\end{tikzcd}$$

Conditioning on V reduces bias, but conditioning on Z would amplify bias from U. If the text contains some combination of V and Z that cannot be separated, we want to condition on the text only if the confounding (V) was large relative to the instrument (Z). We do not recommend our approach when text is posttreatment or text is unrelated to the outcome except through treatment.

In nontext settings with a transparent treatment assignment mechanism, the researcher may be confident about which p variables of $\boldsymbol{X}$ are necessary for the assumption of selection on observables to hold. This is not generally the case for text matching; some words in the text affect treatment, but we are unsure of which ones. If we attempt to match all words, only texts with *identical* word frequencies will match, a possibility that becomes vanishingly small as v grows large.

Our approach is to create a low-dimensional density estimate of the variables in $\boldsymbol{W}$ that we can use to address confounding (in concert with any nontext confounders in $\boldsymbol{X}$). Our function for creating this low-dimensional summary of $\boldsymbol{W}$ is called g, and we refer to the resulting summary as $g(\boldsymbol{W})$. We make the standard conditional ignorability assumption with respect to $g(\boldsymbol{W})$ and $\boldsymbol{X}$, along with standard assumptions of positivity[5] and stable unit treatment values (SUTVA):

Assumption 1 (Conditional Ignorability). $T_i \perp\!\!\!\perp Y_i(0), Y_i(1) | g(\boldsymbol{W}), \boldsymbol{X}$.

Assumption 2 (SUTVA). *For all individuals i, $Y_i(T) = Y_i(T_i)$.*

Assumption 3 (Positivity). *For all individuals i, $Pr(T_i = t) > 0$ for all $t \in \mathcal{T}$.*

For Assumption 1 to be plausible, $g(\boldsymbol{W})$ must retain the aspects of the text that confound causal inference (much in the way we might assume that a "years of education" variable blocks confounding through education). Yet, to be useful, $g(\boldsymbol{W})$ must be a dramatic simplification of $\boldsymbol{W}$. What should this g function look like? The answer depends on the application and the way in which text confounds the treatment and outcome. To develop a useful general method, we note that in many applications, text confounds treatment assignment in two ways: through topics that affect treatment assignment and through individual words that affect treatment assignment, regardless of topic. A useful g function for many applications is one that retains information about both of these channels and eliminates other information about the text.

The first likely channel of text confounding is through topics; treatment assignment might be related to the amount that a document discusses one or more topics. In the censorship application, for example, censorship might be related to the topic(s) that the document discusses, which could also be correlated with the subsequent behavior of the social media user. If topics of documents confound treatment assignment, then methods that summarize the topical content using density estimates become strong candidates for our g function.

The second likely channel of text confounding is through individual words that affect the probability of treatment, regardless of topic. This channel of confounding is especially likely if treatment assignment is made by humans using forms of word searching. For example, if Chinese censors sift through large numbers of posts for the keyword "protest," they might flag posts that use this word incidentally, but are not about political protest. A more subtle version arises if treatment assignment is a function of the specific terms in which a particular topic is discussed, rather than merely the amount of that topic. For example, Chinese censors might encounter two posts with equal attention to the topic of large public marches. However, if one post talks about a public march in terms of a political protest while the other talks in terms of a holiday parade, the first is more likely to be censored. To account for this channel of text confounding, we might turn to methods such as propensity scores that can summarize how individual features affect the probability of treatment.

We prefer g functions that address text confounding from both of these channels. In the next section, we introduce a method called topical inverse regression matching (TIRM) that addresses confounding from both topics and individual words by combining elements of a topic model with elements of a propensity score model.

A distinguishing advantage of using text to adjust for confounding is that reading documents provides the analyst with a richer depiction of the units under study than what is encoded in the data matrix. This is in contrast to, for example, survey data where all known information about each respondent is encoded in the data matrix. Our matching approach leverages this advantage: Reading matched documents is a powerful tool for assessing the suitability of our g function.[6] Analysts should read

[5] D'Amour et al. (forthcoming) show that this assumption may fail in high-dimensional data, presenting another challenge for current matching approaches that we do not take on here.

[6] Although matching has many strengths and is a popular approach (Ho et al. 2007; Sekhon 2009), we acknowledge that it lacks some

matched documents to evaluate two potential points of failure: (1) the g function has obscured an important distinction (e.g., the difference between a parade and a protest), and (2) the matching is too permissive.

Matching on both topics and propensity scores aids in this manual evaluation. Theoretically, the propensity score alone is a sufficient balancing score (Rosenbaum and Rubin 1983). However, propensity score matches are difficult to manually assess because propensity score matching approximates a fully randomized experiment rather than a block-randomized experiment (King and Nielsen 2019); documents with similar treatment probabilities may look very different. For example, propensity scores might match posts about protests to posts about pornography because they have equal probability of censorship. However, it may be difficult to distinguish this from a failure of the g function or the matching algorithm, even for an expert analyst, so we seek topically similar matches.

Text matching is only useful when conditioning on $g(\boldsymbol{W})$ (and other available variables $\boldsymbol{X}$) is sufficient to block confounding. Not only does the text need to contain relevant information about confounding, but also the summary g needs to capture that information. We expect that TIRM will often be a sufficiently rich representation of the text, but if confounding is based on textual aspects besides topics and words (e.g., sentiment, word order, punctuation, spacing, rhyme, font), then analysts should modify $\boldsymbol{W}$ or g accordingly. We prefer topic models for interpretability, but analysts with large data sets might find it easier to fit a simpler model, such as principal components analysis, instead. This would then require fitting a separate model for the propensity score (e.g., regularized logistic regression).

Topical Inverse Regression Matching

There are many matching approaches and many approaches to modeling text, each with strengths and weaknesses. Here, we propose one method for matching on text, topical inverse regression matching (TIRM), that includes the two attributes we believe should be present in most high-dimensional matching. First, it matches on a *coarsened* representation of the text to ensure that the resulting matches are substantively similar. Second, it uses information about how the text relates to *treatment assignment*. Our general methodology relies only on the ability to extract these two quantities: a low-dimensional

attractive theoretical properties. For example, Abadie and Imbens (2006) show that matching estimators are not $N^{1/2}$ consistent and do not attain the Hahn (1998) semiparametric efficiency bound.

Table 1 Overview of the TIRM Method

Step	Rationale
1. Estimate a structural topic model including the treatment vector as a content covariate.	Reduces dimension of the text
2. Extract each document's topics calculated as though treated (part of $g(\boldsymbol{W})$).	Ensures semantic similarity of matched texts
3. Extract each document's projection onto the treatment variable (part of $g(\boldsymbol{W})$).	Ensures similar treatment probability of matched texts
4. Use a low-dimensional matching method to match on $g(\boldsymbol{W})$ and estimate treatment effects using the matched sample.	Standardizes matching

representation of the text and the treatment propensity. TIRM uses a structural topic model (STM; Roberts, Stewart, and Airoldi 2016) as the measurement model to jointly estimate the topics (the low-dimensional representation) and a projection of information about treatment (the treatment assignment model). Although we explain our methodology in terms of STM below, other methodology could be substituted.

Estimation

STM will always estimate the topic distribution of each document. By including the treatment indicator as a content covariate in STM, we show below how to calculate a projection that captures information about treatment propensity not captured in the topics. Matching on this projection and the topic profile of the documents ensures that we will find documents that are topically similar to each other and have a similar probability of receiving treatment. Table 1 provides an overview of the complete procedure.

TIRM Step 1: Estimate STM. STM is a logistic-normal topic model that can incorporate document-specific covariates affecting both *topic prevalence* and *topical content*. Whereas prior work has focused on topic prevalence, we leverage the topical content covariate to capture the relationship between individual words and propensity to treatment. We provide a basic overview of the model but refer readers to details in Roberts, Stewart, and Airoldi (2016).

As with the latent Dirichlet allocation (LDA) model (Blei, Ng, and Jordan 2003), we adopt a "bag of words" language model: For each token l in a document i, first sample a topic $z_{i,l}$ from the document-specific distribution θ_i and then sample the observed word from a topic-specific distribution over the vocabulary. Unlike LDA, STM allows a distribution over the vocabulary that is document-specific; $\boldsymbol{B}_i$ is a $k \times v$ matrix. We represent $z_{i,l}$ as a one-hot-encoding column vector so that $z_{i,l}\boldsymbol{B}_i$ returns a v-length vector giving the distribution over the vocabulary for the particular token's topic. Thus, each token is generated by

$$z_{i,l} \sim \text{Multinomial}_k(\theta_i), \qquad \text{for } l = 1 \ldots L_i; \tag{1}$$

$$w_{i,l} \sim \text{Multinomial}_v(z_{i,l}\boldsymbol{B}_i), \qquad \text{for } l = 1 \ldots L_i. \tag{2}$$

STM allows each document to have an individual prior for θ_i based on topic prevalence covariates, but for notational simplicity we consider a shared global prior:

$$\theta_i \sim \text{LogisticNormal}_{k-1}(\mu, \Sigma), \tag{3}$$

where θ_i is on the $k-1$ dimensional simplex and Σ is a global $k-1 \times k-1$ covariance matrix. Including topic prevalence covariates allows the model to share information about the value θ_i across different documents with similar covariate values. This enters the model by parameterizing μ (see Roberts, Stewart, and Airoldi 2016).

We form the document-specific topic distributions over words by a combination of a baseline word prevalence and sparse deviations from the baseline due to the topic, the content covariate, and the topic–content covariate interaction. For TIRM, we use the treatment status T_i as the content covariate and model row r (the topic) and column c (the vocabulary word) of the matrix $\boldsymbol{B}_i$ as

$$B_{i,r,c} = \frac{\exp\left(m_c + \kappa^{(\text{topic})}_{r,c} + \kappa^{(\text{cov})}_{T_i,c} + \kappa^{(\text{int})}_{T_i,r,c}\right)}{\sum_c \exp\left(m_c + \kappa^{(\text{topic})}_{r,c} + \kappa^{(\text{cov})}_{T_i,c} + \kappa^{(\text{int})}_{T_i,r,c}\right)}, \tag{4}$$

where m_c is the baseline propensity of the word, $\kappa^{\text{topic}}_{r,c}$ is the sparse topic-specific deviation from the baseline for vocabulary entry c, $\kappa^{\text{cov}}_{T_i,c}$ is the deviation from the baseline due to the treatment status, and $\kappa^{(\text{int})}_{T_i,r,c}$ is the deviation from the baseline due to the treatment and topic interaction. The parameters in STM are estimated using variational expectation-maximization.

The analyst selects the number of topics k that sets the granularity of the summary. As we increase the number of topics, matches will be harder to find, but more substantively similar. It is not strictly necessary that the topics be interpretable as long as they provide an accurate density estimate. However, interpretable topics can be useful because they allow the analyst to understand the dimensions along which matched documents are similar. Interpretable topics can also facilitate selective matching on substantively important topics in confounding known to be captured in a limited subset of the document.

In the next two steps, we extract quantities of interest from this model that capture topic similarity and propensity to receive treatment. We can then apply standard matching techniques on those summary measures.

TIRM Step 2: Extract Topic Proportions. The parameter θ_i provides a measure of the document's topical content. To ensure topics are comparable irrespective of treatment and control differences, we reestimate the topics for all control documents as though they were treated. This choice is consistent with an estimand of the (feasible-sample) average treatment effect on the *treated*. Two different words with the same topic are stochastically equivalent under the model and thus can be matched together. In this way, matching on topics can be seen as a high-dimensional analog to the coarsening step in coarsened exact matching (Iacus, King, and Porro 2011). Where coarsened exact matching coarsens within a variable (e.g., treating years 9–12 of schooling as *high school*), topics coarsen across variables (e.g., treating *tax, tariff,* and *economic* as part of an economy topic). When we have estimated more topics, θ_i will be longer, facilitating a more fine-grained match of substantive content. On the other hand, fewer topics will result in a shorter θ_i, which will treat more words as equivalent and create a coarser match.

TIRM Step 3: Extract Treatment Projection. While the refit θ_i captures information about the topic of the document, we want to extract the model-based information about whether or not the unit is treated. The core insight is to treat the topic model with content covariates as a multinomial inverse regression conditional on the latent variables and derive a projection in the style of Taddy (2013; see also Rabinovich and Blei 2014). Appendix B in the supporting information (SI) provides additional details, properties, and alternative strategies. Here, we overview the form of the projection for document i, which we denote ρ_i.

For each treatment level, the STM content covariate model learns a weight for each word ($\kappa^{(cov)}$ in Equation 4) and for each topic–word combination ($\kappa^{(int)}$). The projection for a given document is the sum of the document's weighted word counts normalized by document length.

For a given level of the content covariate denoted by t,

$$\rho_{i,t} = \frac{1}{L_i}\left(\sum_{l=1}^{L}\left(w'_{i,l}\underbrace{\kappa^{(cov)}_{t,c}}_{\text{weight}} + \sum_{r}^{k} w'_{i,l}\underbrace{\overbrace{I(z_{i,l}=r)}^{\text{topic indicator}}\kappa^{(int)}_{t,r,c}}_{\text{topic-specific weight}}\right)\right), \quad (5)$$

where each $w_{i,l}$ is a one-hot encoding that indicates the observed word at token l. In practice, we do not observe the true topic indicators $z_{i,l}$, so we use their posterior means. Each document has a projection value for each level of the treatment that can then be included along with the topics in the matching. Given the model and the latent variables, this term captures the information about treatment assignment not contained in the topics.

TIRM Step 4: Matching and Treatment Effect Estimation. In the final step, we match on both the STM projection and the estimated topic proportions, which ensures that matches are both topically similar and have similar within-topic probabilities of treatment. Though our framework is compatible with most matching algorithms at this final stage, we generally prefer using CEM if pruning treated units is acceptable. Researchers can also match on other pretreatment covariates that are thought to be confounders in this step. In our Chinese censorship example, we also match on the day the post was written, previous posting rate, and previous censorship rate.

Most matching algorithms require analyst input beyond the choice of matching variables. For CEM, analysts must choose the values at which to "coarsen" or "bin" the matching variables. Popular CEM software uses histogram binning algorithms to make automated choices, but we prefer coarsening that is meaningful for text. Our default is to coarsen in a way that roughly corresponds to the presence or absence of each topic; for example, by using two bins—one ranging from 0 to 0.1 and another ranging from 0.1 to 1. This may seem overly permissive, but we have found that this coarsening creates a relatively demanding criterion with many topics because CEM requires matches to share the same bin across all variables. We generally use automated binning with five levels or more for the projections. A larger number of bins with more fine-grained distinctions allows for closer matches but prunes more units. These choices are necessarily application-specific. They depend on the similarity and number of documents available for matching.

Using the matched sample, we fit a model predicting the outcome as a function of treatment status and possibly other controls to estimate the effect of treatment. Inference following matching procedures has been subject to substantial debate, and we do not break new ground on this issue. Standard practice in political science is to use the standard errors from the analyst-preferred post-matching analysis model without correction (Ho et al. 2007), which we do here. Iacus, King, and Porro (2019) show that this results in accurate inference, provided analysts are willing to change their axiom of inference from simple random sampling to stratified sampling. Some analysts have tried to construct alternative estimates of uncertainty using the bootstrap method; Abadie and Imbens (2008) show that the bootstrap method; results are inconsistent estimates of the treatment effect standard error, and they propose an asymptotically consistent alternative for some matching settings (Abadie and Imbens 2006). None of these approaches account for uncertainty in the estimation of the representation of the text. We encourage further work on accounting for uncertainty in text matching, but it is beyond the scope of this article.

Balance Checking

Balance checking—confirming that matched units are in fact similar on pretreatment confounders—is important for assessing whether matching is successful. However, we see no reason to believe there is a universally best balance metric for text similarity, and therefore, checking balance after text matching is not straightforward. We recommend several procedures.

First, we check whether words that predict treatment in the unmatched sample are balanced in the matched sample. Second, we verify that the distribution of topics in treated and control documents is similar in the matched sample. TIRM is designed to jointly minimize both of these, so if these checks reveal that matches are not adequately similar, then technical problems may be to blame, or else good matches may not exist in the data.

The TIRM procedure is designed to maximize balance on the term frequencies.[7] However, our hope is that the procedure has picked up more general similarities in the texts. We assess this in two ways: first, by automated balance checking using a metric not directly optimized by the procedure, and second, by manual evaluation of document pairs.

For the automated balance checking we turn to string kernels, which measure similarities in sequences of characters (Spirling 2012). String kernels retain word order information that we typically discard in text models, so confirming that matching improves the string kernel similarity of treated and control texts builds confidence that even though our procedure omits word order, it still

[7]Two documents of very different lengths can be matched together if they allocate a similar proportion of their length to the same topics. If document length is a confounder, it can be included as a covariate in the matching algorithm.

improves balance in a metric that respects that ordering. In this project, we use simple graphical diagnostics (see SI Figure 5), but future work could develop formal hypothesis tests.

Finally, we check balance manually by reading matched documents. A crucial advantage of our matching approach is that it allows experts to directly scrutinize the claim that matched texts are sufficiently similar. Evaluation through reading is subjective, but it can help analysts judge whether they believe the texts are sufficient for identification and whether balance has been achieved in practice. SI Tables 4 and 6 provide some examples, though we recommend examining more pairs than we have space to present there.

Strengths and Limitations

TIRM is just one solution to the text-matching problem, but it satisfies our desiderata of producing human-verifiable matches. TIRM also estimates both document topic proportions and within-topic propensities for treatment, and, as a result, increased weight is given to words that predict treatment while the resulting matches are topically similar. This allows TIRM to prioritize variables that are related to treatment assignment while approximating a blocked design on the full set of confounders. The method is easy to apply and can be estimated with the existing `stm` software (Roberts, Stewart, and Tingley 2019) or through a new `R` package called `textmatching`.

One limitation of TIRM is that it requires an adequate density estimate for a complex data-generating process. Loosely speaking, the matches are only useful if the topic model is a sufficiently accurate summary of the confounding in the documents. We have found topic models to work well for this purpose. Analysts can always evaluate the quality of their model by substantively interpreting the topics, verifying that they are coherent, and considering whether documents with similar topic proportions are really similar upon close reading. The density estimate of STM can also be replaced by other density estimators that are more attuned to particular types of data (e.g., network or genetic data) or simply because better alternatives to topic models are developed in future years.

TIRM also inherits limitations that are common to other matching methods. In general, matching for causal inference requires the stable unit treatment value assumption (SUTVA; Rubin 1980), which requires any interference between units to be properly modeled. Interference between units is especially likely in applications of high-dimensional matching involving text because the purpose of writing is often to influence or respond to the writing of others. Violations of SUTVA should be carefully considered based on the context of the analyst's application. Like other conditioning approaches to causal inference, TIRM also requires that the selection on observables assumption is met. The core idea of text matching is that the documents themselves contain information about confounding and that the learned representation is sufficient to capture this confounding. If there are other pathways for confounding beyond the text that are not adjusted for, the procedure could be badly biased. Finally, the estimand can change as matching drops observations, particularly treated observations; if so, it is important to characterize the group to which the estimated effect applies (King, Lucas, and Nielsen 2017; Rubin 2006, 221–230). Dropping too many units can also result in a loss of efficiency.

Related Work

Before moving to applications of TIRM, we briefly mention how it relates to other approaches for similar problems. The matching literature has considered the problems of high-dimensional data, but mainly for estimating propensity scores (Belloni, Chernozhukov, and Hansen 2014; Hill, Weiss, and Zhai 2011). We do not find these approaches useful for text matching because they produce matches that have high probabilities of treatment for very different textual reasons. Our approach of augmenting propensity scores with information about topic balance is most similar to the covariate-balancing propensity scores of Imai and Ratkovic (2014). Mozer et al. (2020) and Veitch, Sridhar, and Blei (2019) directly build on our framework to propose alternative text adjustment approaches, and the related literature is reviewed in Keith, Jensen, and O'Connor (2020).

There is little work in the matching framework that proposes matching on a density estimate, as we do here. Price et al. (2006) reduce the dimensionality of genotype data with an eigen decomposition, but follow it with regression adjustment. A recent working paper by Kallus (2018) balances covariate representations based on a deep neural network with applications to image data (which shares some structural similarities to text data). Johansson, Shalit, and Sontag (2016) and Louizos et al. (2017) consider learned representations optimized for causal inference.

Finally, Egami et al. (2018) provide a framework for causal inference with text as treatment or outcome, complimenting our discussion of text-based confounding. See the supporting information for further related work.

Applications and Simulations

To demonstrate the effectiveness of our text-matching approach, we present a two-part validation study in the next two subsections that builds on previous work by Maliniak, Powers, and Walter (2013). Their original study estimates the effect of perceived author gender on the citation counts of academic journal articles in the discipline of international relations. They condition on the text of the articles using variables hand-coded by research assistants based on close reading. First, we use this hand-coded data to produce a simulated data set that we use to study the performance of our proposed estimator. Next, we show that we recover a result using text matching that is similar to the Maliniak, Powers, and Walter (2013) analysis without their hand-coded data. Finally, we demonstrate the use of our methods in our motivating example of studying the effects of government censorship on Chinese social media users.

The Gender Citation Gap: Data and Simulation

If an international relations (IR) article published under a woman's name were instead published in the same venue under the name of a man with the same scholarly credentials, would it be cited more?[8] Maliniak, Powers, and Walter (2013) say yes.[9] Obtaining credible answers to this question is not straightforward with observational data. On average, authorial teams of different gender compositions tend to write on different topics within IR, use different methods, and have different epistemological commitments. Because these factors may affect citation counts, it is possible that lower citation counts for all-female authorial teams reflect bias against certain topics and approaches, rather than against perceived gender of the authors. Maliniak, Powers, and Walter (2013) address this challenge using information from the Teaching, Research, and International Policy (TRIP) Journal Article Database to control for the broad subfield of each article, the issue areas covered, the general methodology, paradigm,[10] and epistemology. They find that academic articles by female authors in IR have lower citation counts than articles by men or mixed-gender author teams, even after accounting for a range of potential text and non-text confounding.

We revisit the question of whether a gender citation gap exists in IR to illustrate the benefits to text matching with TIRM. With the help of JSTOR's Data for Research Program, we supplement the data from Maliniak, Powers, and Walter (2013) with the full text of 3,201 articles in the IR literature since 1980, 333 of which are authored solely by women.[11] We have two goals. The first is to show that TIRM can recover treatment effects in simulated data. Because simulating realistic text data is hard, we base our simulation off of the real text of articles in the IR literature, but simulate treatment effects and confounding. This subsection reports the results of these simulations. Our second goal is to demonstrate how text matching would have allowed comparable adjustment for text-based confounding without the time-consuming process of hand-coding the articles.

In order to design a simulation that has a credible joint distribution of confounder, treatment, and outcome, we use the observed text of the articles and simulate both treatment and outcomes. To avoid assuming that the topics themselves exactly capture the necessary confounding, we use one of the observed hand-coded categories, "quantitative methodology," as the true unobserved confounder in our simulation. Using the *real* article text and hand-coding, we *simulate* a treatment and outcome using the following simple model.

$$T_i \sim \text{Bernoulli}(\pi = .1X_i + .25(1 - X_i)), \qquad (6)$$

$$Y_i \sim \text{Normal}\big(\mu = .5X_i - .2T_i, \sigma^2 = .09\big), \qquad (7)$$

where Y, T, and X are, respectively, the outcome, the treatment, and a binary confounder indicating whether the article is hand-coded as using quantitative methodology. This defines a joint distribution over the outcome (Y_i), treatment (T_i), unobserved binary confounder (X_i taking on the value of whether each article is coded as using quantitative methodology), and observed text ($\boldsymbol{W}_i$). We then use only information about the text ($\boldsymbol{W}_i$) in TIRM to adjust for the unobserved confounder (X_i) to see whether TIRM can use the text to identify the form of confounding coming from the human-labeled category "quantitative methodology." This is a difficult test of TIRM's performance because it must recover a process of

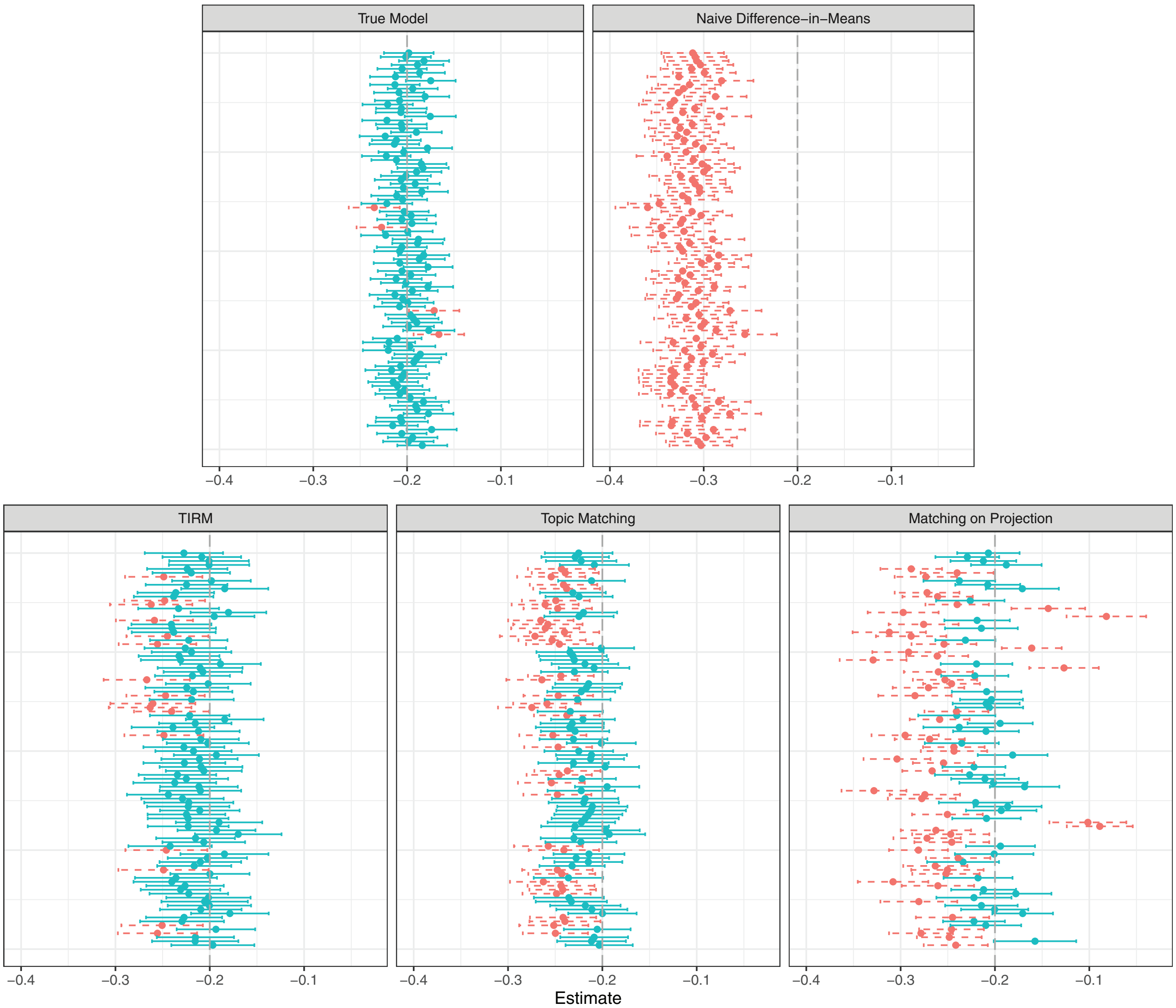
[8]We are estimating the effect of *perceived* author gender in the minds of other authors making citation decisions.

[9]The finding is critiqued in Zigerell (2017) and defended in Maliniak, Powers, and Walter (2017).

[10]Scholarship in international relations is sometimes organized into "paradigms," or schools of thought about which factors are most crucial for explaining international relations. The predominant paradigms are realism, liberalism, and constructivism, though others exist.

[11]We analyze more articles than Maliniak, Powers, and Walter (2013) because the TRIP database has coded more articles since 2013. However, we are missing data for a few articles used by Maliniak, Powers, and Walter (2013) because they are not in JSTOR's data set.

FIGURE 1 Gender Citation Gap Simulation



Note: The plot shows 100 simulations generated using real texts, the real hand-coded "Quantitative Methodology" variable, a simulated treatment, and a simulated outcome (according to Equations 6 and 7). Each panel shows the estimates and 95% confidences intervals from one of five different estimators, including the benchmark correct linear model specification using the unobserved confounder (True Model), our proposed estimator (TIRM), matching on only the topic proportions from the TIRM procedure (Topic Matching), matching on only the projection from the TIRM procedure (Matching on Projection), and the completely unadjusted estimator (Naive Difference-in-Means). Line types indicate whether the interval covers the truth (denoted by a dashed gray line at −.2). TIRM achieves the best coverage out of the models we evaluate.

unobserved confounding from real texts. SI Appendix C offers additional details on the simulation and discusses some of the strengths and weaknesses of this design.

We produce 1,000 simulated data sets for analysis. In Figure 1, we show the results of each model for a subset of 100 simulations. For each simulated data set, we plot the treatment effect estimate and 95% confidence interval for five estimators: the true linear model using the unobserved quantitative methodology variable, the TIRM model,[12] matching on the topics only from the TIRM model, matching on the projections from the TIRM model, and the unadjusted difference-in-means estimator. Table 2 provides summary statistics for all 1,000 simulations.

[12]For each simulated data set, we apply the TIRM procedure using 15 topics and matching with CEM on the projection score for treatment using eight automatically generated bins and the topics using two bins each (0−.1 and .1−1).

TABLE 2 Demonstration of TIRM Estimator Performance and Relevant Benchmark Estimators across 1,000 Simulations

	MSE	Bias	Coverage	Avg. # Treated	Avg. # Matched
True Model	0.00018	0.00032	0.955	619.6	3,201.0
TIRM	0.00099	−0.02214	0.826	392.8	1,447.2
Topic Matching	0.00144	−0.03377	0.547	548.8	2,142.8
Matching on Projection	0.00312	−0.03220	0.456	496.2	2,693.4
Naive Difference-in-Means	0.01299	−0.11278	0.000	619.6	3,201.0

Note: Results are compared on mean squared error (MSE), bias, coverage of the 95% confidence interval, average number of treated units in the matched set, and average number of treated and control units in the matched set.

The TIRM model performs substantially better than matching only on the topics or the projection, particularly in terms of bias and coverage. Although TIRM's 82% coverage rate on this example does not match the nominal 95%, it performs well given that the true confounder has been withheld from the model.

The Gender Citation Gap: Female IR Scholars Are Cited Less Often

We now return to the substantive question motivating our reanalysis of the data from Maliniak, Powers, and Walter (2013): Are female IR scholars cited less often due to perceived gender? Maliniak, Powers, and Walter (2013) control for differences in the writing of men and women using qualitative variables painstakingly hand-coded by research assistants over more than a year. We recover similar effects with TIRM using far less manual effort.

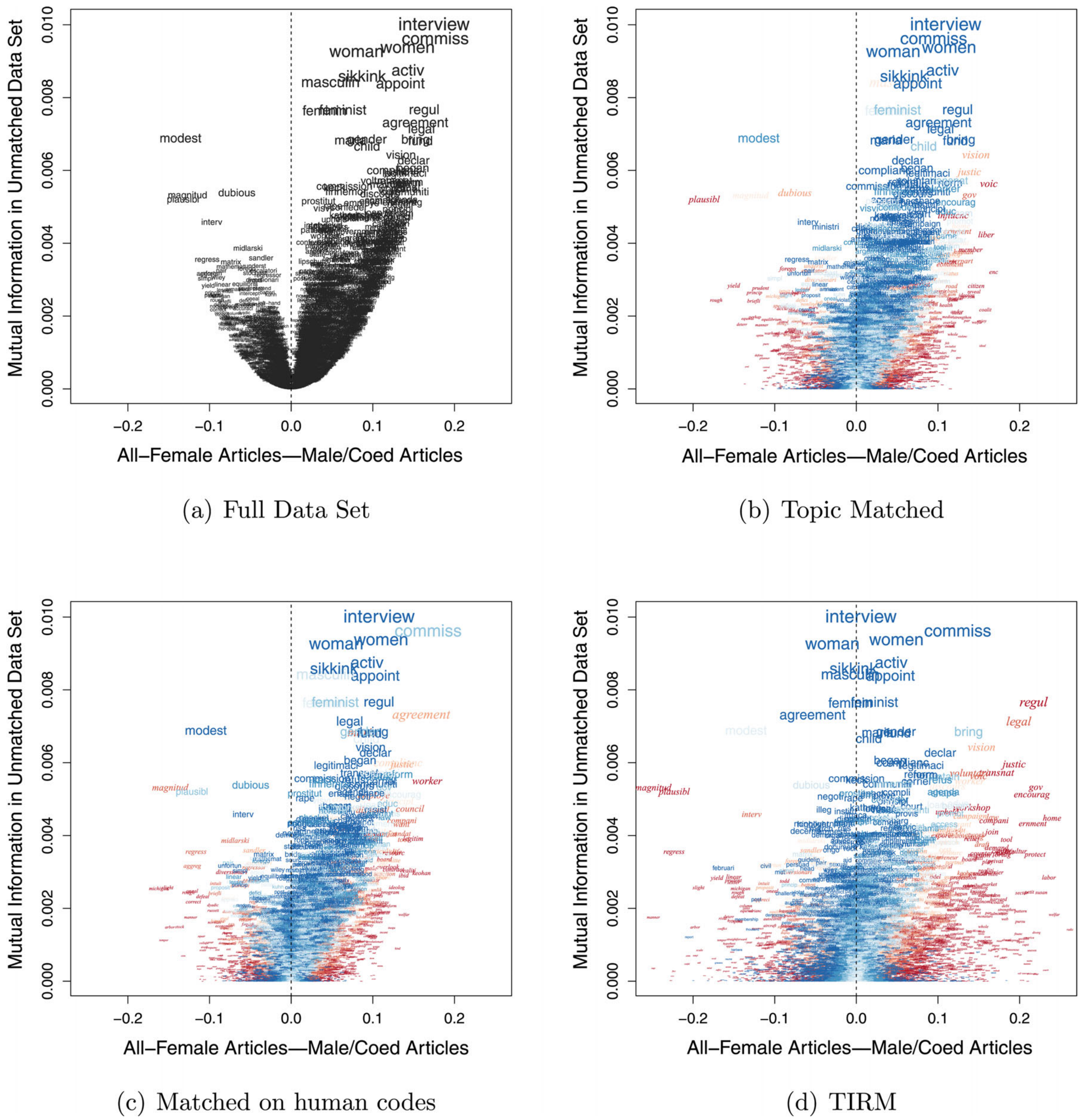
Women's and men's writings are not the same, on average, in the unmatched data set. Among other differences, women are more likely to write about gender (*women, gender*, and *children*), whereas men or male/female coauthored articles are more likely to use words associated with statistical methods (*model, estimate*, and *data*). We also find substantial imbalance in the human-coded variables measuring each article's research approach, paradigm, methodology, and issue area.

We apply TIRM to the data, specifying 15 topics in the STM portion of the algorithm, using all-female authorship as the treatment variable. This is a relatively small number of topics to span the diversity of the IR literature, so this model recovers relatively broad topics. We use CEM to match on the TIRM output, along with three nontext variables from Maliniak, Powers, and Walter (2013): (1) whether at least one author is tenured; (2) whether the article appeared in the *American Political Science Review*, *American Journal of Political Science*, or *Journal of Politics*; and (3) article age. We drop treated units with no available matches, so it is important to note how the treated pool changes after matching. On average, women's articles in the *matched* sample are 3 years older and 8 percentage points more likely to have a tenured author than women's articles in the *full* sample. The matched sample is also heavier on topics related to security and political economy and lighter on international organizations and institutions.

We compare the performance of TIRM to alternative approaches: matching only on propensity scores, matching only on topics, and exact matching on the original human-coded data. We first look for improvements in balance among the words that are most imbalanced in the unmatched data set. We identify the words that have high *mutual information* with author gender in the raw data set as well as the matched data from TIRM, matching only on topics, and human matching.[13] Figure 2 shows the relationship between the difference in word occurrence by gender and the mutual information of each word in each data set. If perfect balance on all words were possible, we would hope to see every word lined up vertically on the $x = 0$ line (and in sans-serif typeface accordingly). However, since not all words can be balanced, balance on words with high mutual information is most important. TIRM—shown in the bottom-right panel—outperforms the others in balancing the high mutual information words. Many high mutual information words such as *interview* that were previously imbalanced are now lined up down the center. TIRM makes the imbalance substantially worse on words with low mutual information, but because these words have low imbalance

[13]We calculate the mutual information for an individual word w as the difference between the entropy of category k, $H(k)$, and the entropy of category k when conditioning on a word's appearance in the document, $H(k|w)$. $H(k) - H(k|w)$ can be calculated as follows: $H(k) - H(k|w) = \sum_{t=0}^{1}\sum_{s=0}^{1} P(k = t, w = s)\log_2 \frac{P(k=t,w=s)}{P(k=t)P(w=s)}$. See Grimmer (2010) and Manning, Raghavan, and Schütze (2008) for a longer treatment of mutual information.

FIGURE 2 Relationship between Mutual Information and Difference in Word Occurrence



(a) Full Data Set (b) Topic Matched (c) Matched on human codes (d) TIRM

Note: (all female - male/coed) in a) the full data set; b) topic matched; c) matched on human codeing; and d) TIRM. In panels b, c, and d, words for which matching decreased the absolute value of the difference in word appearance are in sans-serif typeface and words for which matching increased the absolute value of the difference in word appearance are in italicized serif typeface. Darker words changed more in either direction. Word size is proportional to each word's mutual information

to begin with, we anticipate that they will not substantially increase bias. This analysis highlights the benefits of the treatment model because it can identify and address the most imbalanced words—exact matching on human coding and matching only on topics do not perform as well as TIRM.

We also check balance in other ways. We manually examine pairs of matched documents to confirm these pairs match our intuitions about which articles in the IR literature are similar enough for comparison. We read more pairs than we can conveniently present, but see SI Table 4 for a few examples. We also evaluate TIRM's success at balancing the human-coded variables of article substance and the estimated topics from the topic model. We find that TIRM performs reasonably well at balancing the human-coded variables, particularly the most imbalanced ones. This is reassuring because in most applications, the purpose of TIRM is to substitute for painstaking human coding of texts. We also find that TIRM balances the STM topics well. Matching on topics only (without the TIRM projection) performs slightly better at balancing the STM topics, but that is to be expected; TIRM is trying to simultaneously balance the estimated topics *and* the probability of treatment. Finally, we check string kernel similarity of the matched data sets and find that TIRM outperforms the alternatives and offers substantial improvement over the raw data. Details of these balance checks are in SI Appendix D.

Maliniak, Powers, and Walter (2013, 906) report that women's articles receive 4.7 fewer citations on average. Our estimates from similar specifications using the TIRM matched sample show a slightly more pronounced gender gap of 6.5 citations, most of which seems concentrated among the highest citation-earning articles. Because this article is focused on the method, we refer readers to SI Appendix D for further results and a sensitivity analysis.

Government Censorship of Chinese Social Media Users

The Chinese government oversees one of the most sophisticated censorship regimes in the world (Esarey and Qiang 2008; MacKinnon 2011; Marolt 2011), with technologies ranging from manipulating search results to blocking foreign websites. Even as we learn more about the types of content that are censored and the technical infrastructure enabling censorship (Bamman, O'Connor, and Smith 2012; King, Pan, and Roberts 2013, 2014), we still know little about the subsequent impacts of censorship on individual social media users in China. Is censorship completely determined by the text of a particular post, or does censorship become more targeted toward users based on their previous censorship history? This issue is particularly important; the targeted use of artificial intelligence for censorship has become a cause of concern because it further complicates the ability of the public to hold the government accountable for censorship decisions (Morozov 2012; Roberts 2018; Tufekci 2014).

Our goal is to estimate the causal effect of experiencing censorship on the subsequent censorship and posting rates of social media users in China. The Chinese censorship regime appears to attempt to thoroughly remove social media posts it deems sensitive (King, Pan, and Roberts 2013, 2014), but the censors miss some. We exploit these mistakes by using TIRM to identify pairs of nearly identical social media posts written by nearly identical users, where one is censored and the other is not. We can then observe the subsequent censorship rates of both users to estimate the causal effect of censorship on the treated units who remain in our sample.

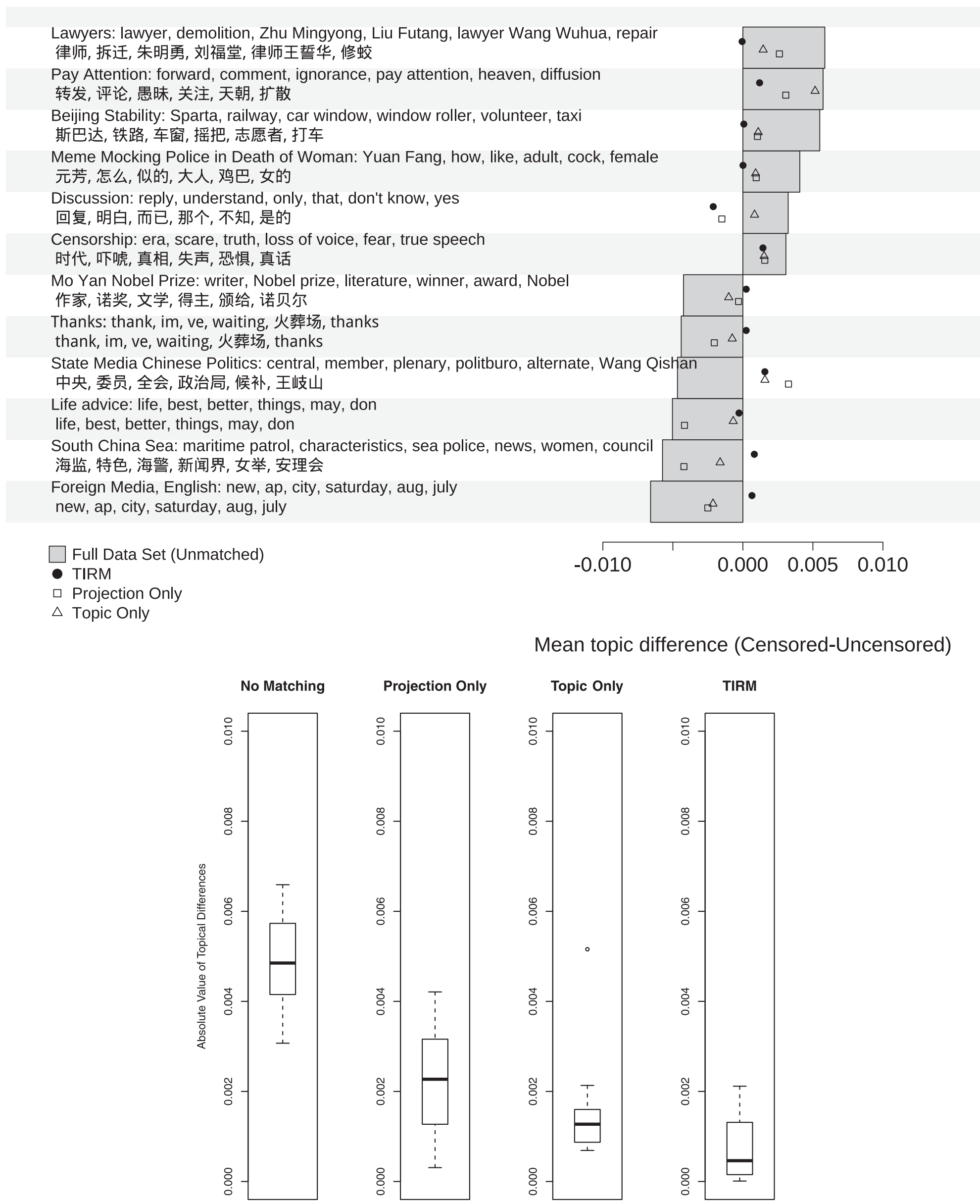
We use data on 4,685 social media users on Weibo from the Weiboscope data set (Fu, Chan, and Chau 2013).[14] Fu, Chan, and Chau (2013) collected posts from Weibo in real time and then revisited these posts later to observe whether they were censored. After processing the text to account for linguistic features of Chinese, we convert the text of social media posts into a matrix of document word counts and estimate TIRM using 100 topics. There may be confounding factors not captured in the text if, for example, the government uses nontext information to make censorship decisions. To address these possibilities, we also match on nontext confounders: previous post rate, previous censorship experience, and date. We assume these covariates capture pretreatment differences in user targeting. Our final matched sample contains 879 posts. Since we drop treated units without matches, we describe the differences between our matched and unmatched samples. On average, the matched sample has a slightly lower previous censorship rate, is more likely to discuss environmental protests such as those in Shifang and Qidong in 2012, and is less likely to discuss human rights lawyers and central politics. Additional details are in the supporting information.

For illustration, we compare matching with TIRM to matching only on the topics or only on the projection. We find that the TIRM is effective at identifying social media posts about the same topic or event, but with different censorship statuses. Figure 3 shows that TIRM matching outperforms other matching strategies in reducing the difference between topics in censored and uncensored posts. Topic matching is a close second, and matching only on the projection performs poorly. TIRM also outperforms the others at improving the string kernel similarity of documents in the matched data set (details in SI Figure 7). TIRM allows us to manually evaluate the similarity of the matched documents, some of which we show in Table 3.

Our first outcome measure is the censorship rate of each blogger after the matched post. Our second outcome is the rate of posting after censorship. The results show that censored users are likely to experience more censorship in the future as a result of their censorship experience. Having a post censored increases the

[14]Due to data size, we selected only users censored at least once in the first half of 2012 and study their behavior in the last half of 2012, and we restrict the control donation sample to posts made on the same day and with a cosine similarity to a censored post greater than 0.5.

FIGURE 3 Topic Balance Comparison



Note: The top panel shows the balance for unmatched, projection-only matched, topic-only matched, and TIRM-matched for the most unbalanced topics in the unmatched data set. The bottom panel shows the distribution of the absolute values of topical differences for all 100 topics under each type of matching. TIRM outperforms the other matching methods.

probability of future censorship significantly,[15] but it does not decrease the number of posts written by the censored user. This suggests one of two scenarios. This evidence is consistent with algorithmic targeting of censorship, where social media users are more likely to be censored after censorship because they are flagged by the censors. Alternatively, social media users may chafe against censorship and respond by posting increasingly sensitive content that is more likely to be censored. Either way, these results are consistent with Roberts (2018), who finds a similar pattern using exact matches, and indicate that users do not seem to experience a chilling effect of censorship. In

[15]The probability of a "Permission denied" post is not different between the two groups in the pretreatment period, but it is 0.009 for the censored group and 0.004 for the uncensored group in the period after treatment. The probability of a "Weibo does not exist" post is not different between the two groups in the pretreatment period, but it is 0.25 for the censored group and 0.20 for the uncensored group in the period after treatment.

TABLE 3 Translations of example social media posts that were censored (left) with matched uncensored social media posts selected by TIRM (right)

Censored Post	Uncensored Post
There may be even bigger plans: When the chaos escalate to a certain degree, there will be military control, curfew, Internet cutoff, and then complete repression of counterrevolution. There are precedent examples.	The person on the right (refers to the previous comment) knew too much. Bang (Sound effect for gunshot)! You knew too much! Bang! There may be even bigger plans: When the chaos escalate to a certain degree, there will be military control, curfew, Internet cutoff, and then complete repression of counterrevolution. There are precedent examples.
#Weitianxia#Shifang netizen's disclose: I saw police officers looking for restaurants to eat on the street and the intestine vermicelli restaurant owner immediately said they don't sell it to the police. Then everyone on the street came out and yelled, which was very impressive. Now many stores have signs saying that police tactical units are not allowed to enter. Shifang people said: F∗k you, you beat us up, bombed us, and still ask us to feed you, why don't you eat sh∗t?	Due to the lack of prior publicity procedures, some people are unfamiliar, uncomprehending and unsupportive of this program. To respond to the general public's request, the municipal party committee and the government researched and decided to stop the project. Shifang will not ever develop the Molybdenum Copper project in the future.
[17-year-old young athlete fails 3 attempts to lift The media calls it a shame of Chinese female weightlifters] According to Sina: Chinese female weightlifters faced a shameful failure of its Olympic history last night! During the female 53kg weightlifting competition, joined as the black horse, Zhou Jun, a 17-year-old young athlete from Hubei, failed in all 3 of her attempts and ended with no result, which ends her Olympic journey. Many media reported this using "the most shameful failure of Chinese female Olympic weightlifters" as the title.	[17-year-old young athlete fails 3 attempts to lift The media calls it a shame of Chinese female weightlifters] According to Sina: Chinese female weightlifters faced a shameful failure of its Olympic history last night! During the female 53kg weightlifting competition, joined as the black horse, Zhou Jun, a 17-year-old young athlete from Hubei, failed in all 3 of her attempts and ended with no result, which ends her Olympic journey. Many media reported this using "the most shameful failure of Chinese female Olympic weightlifters" as the title. I personally think, it is not a shame of Zhou Jun, but a shame of Chinese media!

the supporting information, we explore sensitivity to unobserved confounding and the choice of CEM coarsening.

Conclusion

Scholars across the social sciences are finding an increasing number of ways to use text as data. In this article, we have proposed conditioning on text to address confounding using matching. We identify the core concerns for addressing confounding from text, provide a method for text matching, and introduce approaches to balance checking. Matching text is difficult because it is inherently high-dimensional; we address this concern with a simple approach matching on a density estimate and a projection that captures propensity to treatment. To assist applied researchers wishing to make causal inferences with high-dimensional data, we provide the `textmatching` package in R, which uses the `stm` package (Roberts, Stewart, and Tingley 2019) to implement the matching procedures described in this article. This general strategy may have applications in other types of high-dimensional data.

Our interest in text matching is born out of a practical necessity from the applications we present. There are an enormous number of research problems in which the content of texts potentially confounds causal inference in observational studies, and the different characteristics of our case studies reflect this diversity. We have used these methods in several languages, with corpora of varying size, and typical document lengths as short as a couple of sentences to roughly 10,000 words. These applications suggest that our solution has the flexibility to address the tremendous variety characteristic of social science data.

Text matching is not a panacea for observational causal inference, and perfectly balancing on text does not guarantee unbiased causal inference. In both our applications, we had to condition on additional nontext variables to block other paths of confounding. Credible causal inference requires careful design and domain knowledge. This domain knowledge may suggest other approaches to text adjustment, such as supervised measurement of a specific quantity. Regardless, we encourage applied users to leverage the greatest strength of text-based adjustment: the ability to carefully read documents to assess what the model can and cannot measure.

Social science has made tremendous progress in developing text-as-data methods, but we are only at the beginning of developments at the intersection of text and causal inference. We see several opportunities for pushing forward the text-based confounding literature. We hope that scholars will extend our work to a proposed alternative to TIRM, a task already started by Mozer et al. (2020) and Veitch, Sridhar, and Blei (2019). A central challenge is developing general-purpose methods for evaluating these new models. Future work could also introduce additional approaches to model criticism and balance checking to fill out the text-matching workflow. In addition to these aspects of practice, key aspects of theory remain to be addressed, including rigorous approaches to uncertainty, consistency results that allow our model to function as only an approximation to g, and implications of positivity assumptions in high dimensions.

References

Abadie, Alberto, and Guido W. Imbens. 2006. “Large Sample Properties of Matching Estimators for Average Treatment Effects.” *Econometrica* 74(1): 235–67.

Abadie, Alberto, and Guido W. Imbens. 2008. “On the Failure of the Bootstrap for Matching Estimators.” *Econometrica* 76(6): 1537–57.

Bamman, David, Brendan O’Connor, and Noah Smith. 2012. “Censorship and Deletion Practices in Chinese Social Media.” *First Monday* 17(3). https://journals.uic.edu/ojs/index.php/fm/article/view/3943

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. “Inference on Treatment Effects after Selection among High-Dimensional Controls.” *Review of Economic Studies* 81(2): 608–50.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3(jan): 993–1022.

D’Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Forthcoming. “Overlap in Observational Studies with High-Dimensional Covariates.” *Journal of Econometrics.*

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. “Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.” *Review of Economics and Statistics* 95(3): 932–45.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. “How to Make Causal Inferences Using Texts.” arXiv preprint arXiv:1802.02163.

Esarey, Ashley, and Xiao Qiang. 2008. “Political Expression in the Chinese Blogosphere: Below the Radar.” *Asian Survey* 48(5): 752–72.

Fu, King-wa, Chung-hong Chan, and Michael Chau. 2013. “Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy.” *IEEE Internet Computing* 17(3): 42–50.

Grimmer, Justin. 2010. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.” *Political Analysis* 18(1): 1–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3): 267–97.

Hahn, Jinyong. 1998. “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects.” *Econometrica* 66(2): 315–31.

Hainmueller, Jens. 2011. “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies.” *Political Analysis* 20(1): 25–46.

Hill, Jennifer, Christopher Weiss, and Fuhua Zhai. 2011. “Challenges with Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative.” *Multivariate Behavioral Research* 46(3): 477–513.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15(3): 199–236.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. “Multivariate Matching Methods That Are Monotonic Imbalance Bounding.” *Journal of the American Statistical Association* 106(493): 345–61.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2019. “A Theory of Statistical Inference for Matching Methods in Causal Research.” *Political Analysis* 27(1): 46–68.

Imai, Kosuke, and Marc Ratkovic. 2014. “Covariate Balancing Propensity Score.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1): 243–63.

Johansson, Fredrik, Uri Shalit, and David Sontag. 2016. “Learning Representations for Counterfactual Inference.” In *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA.

Kallus, Nathan. 2018. “DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training.” arXiv preprint arXiv:1802.05664.

Keith, Katherine, David Jensen, and Brendan O’Connor. 2020. “Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates.” arXiv preprint arXiv:2005.00649, University of Massachusetts Amherst.

King, Gary, Christopher Lucas, and Richard A. Nielsen. 2017. “The Balance-Sample Size Frontier in Matching Methods for Causal Inference." *American Journal of Political Science* 61(2): 473–89.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(1): 1–18.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2014. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science* 345(6199): 1251722. https://gking.harvard.edu/files/gking/files/chinasci2.pdf

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27(4): 435–54.

Kuroki, Manabu, and Judea Pearl. 2014. "Measurement Bias and Effect Restoration in Causal Inference." *Biometrika* 101(2): 423–37.

Louizos, Christos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. "Causal Effect Inference with Deep Latent-Variable Models." In *Advances in Neural Information Processing Systems*. 6446–56. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

MacKinnon, Rebecca. 2011. "China's 'Networked Authoritarianism.'" *Journal of Democracy* 22(2): 32–46.

Maliniak, Daniel, Ryan Powers, and Barbara F. Walter. 2013. "The Gender Citation Gap in International Relations." *International Organization* 67(4): 889–922.

Maliniak, Daniel, Ryan Powers, and Barbara Walter. 2017. "A Reply to 'Reducing Political Bias in Political Science Estimates.'" *PS: Political Science & Politics* 50(1): 184–85.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.

Marolt, Peter. 2011. "Grassroots Agency in a Civil Sphere? Rethinking Internet Control in China." In *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival*, ed. David Herold and Peter Marolt. New York: Routledge, 53–68.

Morozov, Evgeny. 2012. *The Net Delusion: The Dark Side of Internet Freedom*. New York: PublicAffairs.

Mozer, Reagan, Luke Miratrix, Aaron Russell 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*, 1–24. https://10.1017/pan2020.1

Pearl, Judea. 2011. "Invited Commentary: Understanding Bias Amplification." *American Journal of Epidemiology* 174(11): 1223–27.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38(8): 904–09.

Rabinovich, Maxim, and David Blei. 2014. "The Inverse Regression Topic Model." In *Proceedings of the 31st International Conference on Machine Learning*. 199–207.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton, NJ: Princeton University Press.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515): 988–1003.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91(2): 1–40.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rubin, Donald B. 1980. "Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomization Test.'" *Journal of the American Statistical Association* 75: 591–93.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, Donald B., and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52: 249–64.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Spirling, Arthur. 2012. "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* 56(1): 84–97.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503): 755–70.

Tufekci, Zeynep. 2014. "Engineering the Public: Big Data, Surveillance and Computational Politics." *First Monday* 19(7).

Veitch, Victor, Dhanya Sridhar, and David M. Blei. 2019. "Using Text Embeddings for Causal Inference." arXiv preprint arXiv:1905.12741.

Zigerell, L. J. 2017. "Reducing Political Bias in Political Science Estimates." *PS: Political Science & Politics* 50(1): 179–83.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Additional Related Work
Appendix B: Treatment Projection
Appendix C: Simulation Details
Appendix D: The Gender Citation Gap: Balance Checking, Results, and Sensitivity Analysis
Appendix E: Details of Chinese Social Media User Analysis