Forthcoming at the *American Journal of Political Science*

**Adjusting for Confounding with Text Matching***

Margaret E. Roberts†, Brandon M. Stewart‡ and Richard A. Nielsen§

May 26, 2019¶

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 covariates 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 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 academia and the effects of censorship on Chinese social media users.

9,836 words (text and references)

---

*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, Simone Zhang, audiences at the Princeton Text Analysis Workshop, Princeton Politics Methods Workshop, Microsoft Research, Text as Data Conference, 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 supported, in part, by The Eunice Kennedy Shriver National Institute of Child Health & 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.

†Associate Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093-0521, (858) 284-0616, meroberts@ucsd.edu, MargaretRoberts.net

‡Assistant Professor, Department of Sociology, Princeton University, 149 Wallace Hall, Princeton, NJ 08544, 757-636-0956, bms4@princeton.edu, brandonstewart.org

§Associate Professor, Department of Political Science, Massachusetts Institute for Technology, 77 Massachusetts Avenue, E53 Room 455, Cambridge, MA, 02139, 617-254-4823, rnielsen@mit.edu, http://www.mit.edu/~rnielsen/research.htm

¶The first draft of this paper was circulated in the summer of 2015 under the title “Matching Methods for High-Dimensional Data with Applications to Text”
1 Introduction

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 a 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 a sensitive topics or do they avoid them? These are counterfactual, causal questions about how a treatment (the experience of censorship) shapes subsequent outcomes.

Experimentally manipulating censorship would allow us to make credible causal inferences about the effects of experience of censorship, but is impractical and unethical outside of a lab setting. Inferring causal effects in observational settings is challenging due to confounding. Users who are censored have different preferences which drive them to write different things. This in turn affects both their rate of censorship as well as future outcomes. We argue that conditioning on the text of the censored social media post and other user-level characteristics can substantially decrease or eliminate confounding leading to credible causal inference even with observational data. Intuitively, if we can find nearly identical posts, one of which is censored while the other is not, we can compare downstream online behavior to estimate credible estimates of the effects of censorship.

Traditional matching methods are not suited to the task of conditioning on the text of documents. Analysts typically represent text quantitatively using thousands or even millions of dimensions (e.g., the columns of a document term matrix, collections of word embeddings, etc). Common matching techniques such as propensity score matching (Rosenbaum and Rubin, 1983) and coarsened exact matching (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,” while 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 address confounding captured in text effectively and transparently. We see our contribution as fourfold. First, we introduce a framework for using the content of text to address confounding in observational data. Second, we develop 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 simulation studies. Finally, we demonstrate how to apply text-matching for causal inference through two applications.¹

A strength of our matching approach relative to other conditioning strategies is that analysts can evaluate the quality of the adjustment by reading treated documents alongside their matches. This allows analysts to draw on their substantive knowledge to recognize when sources of confounding remain in the matched text even if they cannot formalize that knowledge 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).

In our first application, we validate our method by replicating and extending the work of Maliniak, Powers and Walter (2013) who find that perceived gender of International Relations scholars affect citations to their articles. In order to address confounding, the authors controlled for article content with hand-coded variables. We show that applying TIRM allows us to recover a similar effect estimate without using the hand-coded data. This suggests that TIRM is 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

¹Our primary contribution is to pose the problem of text-based confounding and offer one possible, but not necessarily definitive, solution. Since our paper started circulating in July 2015, Mozer et al. (2018) have introduced another approach to matching text which we hope is the first of many innovations in this area.

²If the analyst has a clear balance metric that captures confounding in text, they can directly optimize it using any of a variety of methods (Hainmueller, 2011; Diamond and Sekhon, 2013; 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 the dependence on that assumption by facilitating human validation.

³A similar analysis is performed in Roberts (2018), who uses exact matches to accomplish this goal.
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 much more likely to be censored after censorship, 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 setting that 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 applicant race on the probability of college admission. And scholars of International Relations might wish to control for the content of international agreements when estimating the determinants of international cooperation. Our approach could also apply to non-text data in computer vision, population genetics, biological microarrays, and other areas where a generative model of pre-treatment covariates can be reliably estimated, or when the observed data are noisy measures of a latent confounder (Kuroki and Pearl, 2014).

The paper proceeds as follows. In Section 2, we describe text-based confounding adjustment, why we opt for a matching approach, and define basic notation. We highlight the importance of conditioning on both a density estimate and a measure of propensity for treatment. In Section 3 we present Topical Inverse Regression Matching (TIRM) as a way to jointly estimate the density and a measure of treatment propensity. We also offer an approach to balance checking and discuss the method’s strengths, limitations, relation to prior work, and performance in simulation. In Section 4, we detail our two applications: a validation study demonstrating the effect of perceived author gender on academic article citations (Section 4.2) and a study estimating the effect of being censored on the reactions of bloggers (Section 4.3). Section 5 concludes with a discussion of future directions.

2 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 plausible counterfactuals by finding pairs of similar posts by different authors where one post was censored and the other was not. In short, we are looking for censorship mistakes. We find that being censored increases the probability of future censorship, but 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 \( p \) pre-treatment covariates \( X = (X_1, X_2, \ldots, X_p) \) to improve similarity in the distribution of covariates within treatment and control groups, a condition called balance. The selection on observables identification strategy assumes conditional ignorability: \( T_i \perp \perp Y_i(0), Y_i(1) | X \). Under this assumption, balancing the distribution of observed covariates \( X \) across treatment groups provides us a way to estimate a causal effect of interest. 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, \( X \) is low-dimensional, with \( p \ll n \). However, in the cases we consider the potential dimensionality of \( p \) is very large. For example, censorship in social media posts may be associated with 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 \( W \) whose typical element \( W_{ij} \), contains the

\(^4\)Typically, exact matches on all observed covariates \( X \) are not possible, so we match treated units to the closest control units and then estimate the \( ATT \) within matches. In cases where treated units within some minimum distance of control units, researchers typically remove these units, in which case the quantity estimated is the Feasible Sample Average Treatment Effect on the Treated, henceforth \( FSATT \) or the \( ATT \) for the set of sampled treated units for which matches can be found.
number of times the \( j \)th word appears within the text associated with unit \( i \). The \( W \) 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 is a simplification that largely ignores word order or word combinations. These assumptions can be weakened to include word order or hierarchy by modifying \( W \) to include these features.\(^5\)

In some non-text settings, the \( p \) variables of \( X \) under which selection on observables would hold are known to the researcher because the treatment assignment mechanism is transparent. This is not the case for our applications with the matrix \( W \). We believe that some words in the text affect treatment, but are unsure which ones. If we attempt to match on all words, we will not identify any matches unless two texts have *identical* word frequencies, a possibility that becomes vanishingly small as \( v \) grows large.

Our approach is to estimate a low-dimensional summary of the variables in \( W \) that we can use to address confounding (optionally in addition to other pre-treatment confounders \( X \)). We denote this low-dimensional summary of the text as a function called \( g \). We assume that conditional ignorability is possible given some low-dimensional function \( g \) and pretreatment covariates. We also make standard assumptions of positivity and Stable Unit Treatment Values (SUTVA), such that in full we require:

**Assumption 1** (Conditional Ignorability). \( T_i \perp \perp Y_i(0), Y_i(1)|g(W), X \)

**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 \( t \in T \).\(^6\)

What should this \( g(W) \) function look like? We propose that there are two different ways in

---

\(^5\)Conditioning on text is appropriate when it is pre-treatment and doing so will address confounding. If the text is a result of treatment, rather than its cause, conditioning on text will induce post-treatment bias. If the text is an instrument, meaning it explains the outcome only through the treatment, conditioning on the text can actually amplify the bias of unconditioned confounders (Pearl, 2011). Consider the following directed acyclic graph which contains some small, but unavoidable confounding \( U \). Conditioning on \( X \) reduces bias but conditioning on \( Z \) would amplify bias from \( U \). If the text is some combination of \( X \) and \( Z \), we want to condition on it only if the confounding \((X)\) was large relative to the instrument \((Z)\).

We do not recommend our approach when text is post-treatment or text is unrelated to the outcome except through treatment.

\(^6\)D’Amour et al. (2017) provide a reassessment of positivity in the high-dimensional context and show that this assumption can often fail. This is another way in which high-dimensional data are a challenge for current matching approaches, but not one we take on here.
which a text might confound treatment assignment and that \( g(W) \) should reflect both of these possibilities. First, 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 to the subsequent behavior of the social media user. Matching on topics (using a topic model to estimate \( g \)) is one way to ensure that matched documents are substantively similar.\(^7\) As we discuss in more detail below, ensuring that documents are substantively similar also facilitates human validation of the matched documents.

While matching on topics helps ensure that the general substance of a treated and control document might be the same, the topic model may miss nuances of treatment assignment. In the censorship example, if both documents discuss a large group of people getting together, one might talk about a political protest while the other discusses a parade. Therefore, the matching method should not only ensure that the documents have similar content, but also match on a measure that captures the propensity to receive treatment.

The probability of treatment is a sufficient balancing score, so why match on estimated topics as well? The key advantage of matching relative to alternative weighting or regression strategies is that analysts can compare matched documents and assess whether likely confounding remains.\(^8\)

When matching only on the estimated propensity score from text, it is difficult to tell whether matched pairs are dissimilar because the text model has failed or because wildly different types of documents can have similar probabilities of treatment. For example, in our censorship application, posts related to protests and pornography might have similarly high probabilities of censorship. Do unexpected matches of pornography with protest indicate that our model is flawed? Or can these posts serve as adequate counterfactuals for each other? It is hard for a human to tell by evaluating matched documents. To avoid this problem, we match on both topics and the probability of treatment.

\(^7\)Other ways of matching the substance of the text could be matching on the principal components of a document or using a similarity metric like cosine similarity to ensure that documents have similar words.

\(^8\)Although popular because of widely cited introductions (Ho et al., 2007; Sekhon, 2009), matching lacks some 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.
3 Topical Inverse Regression Matching

There are hundreds of matching approaches, each with strengths and weaknesses. Similarly, there are many different approaches to modeling text. Here we propose one method for matching on text, Topical Inverse Regression Matching (TIRM) because it includes the two important attributes we discussed above that we believe should be present in all high-dimensional matching: it matches on a coarsened representation of the text to ensure that the resulting matches are substantively similar and it uses information about how the text relates to treatment assignment. In this section, we describe the TIRM method and we show that matching on both information about how the text relates to treatment assignment and a coarsened representation of the text is preferable to matching only on one or the other in simulations and applications.

3.1 Estimation

TIRM is one approach to achieving both of these general goals we described above. In order to capture both the amount that the potentially confounding document discusses a topic and the subtle ways in which the words in the document could be related to treatment, TIRM uses output from the Structural Topic Model (STM) (Roberts, Stewart and Airoldi, 2016). For each document, STM provides an estimate of the topic distribution of the document. As we describe below, by including the treatment vector as a content covariate, STM also provides an estimate how word use differs by treatment group both within and across topics. Using these estimates, we can 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.

3.1.1 TIRM Step 1: Estimate STM

STM is a logistic-normal topic model which can include document-specific covariates affecting both topic prevalence and topical content. While prior work has focused on topic prevalence, here we leverage the topical content covariate to capture the relationship between individual words and propensity to treatment. For space, we provide a basic overview of the model here, but refer readers to more specifics in Roberts, Stewart and Airoldi (2016).

9Note that other approaches could also include these two attributes, and we see TIRM as only one of many possible solutions, no single instance of which is likely to be optimal in all cases.
Table 1: Overview of the TIRM Method

<table>
<thead>
<tr>
<th>Step</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Estimate a Structural Topic Model including the treatment vector as a content covariate.</td>
<td>Reduces dimension of the text.</td>
</tr>
<tr>
<td>2. Extract each document’s topics calculated as though treated (part of $g(W)$).</td>
<td>Ensures semantic similarity of matched texts.</td>
</tr>
<tr>
<td>3. Extract each document’s projection onto the treatment variable (part of $g(W)$).</td>
<td>Ensures similar treatment probability of matched texts.</td>
</tr>
<tr>
<td>4. Use a low-dimensional matching method to match on $g(W)$ and estimate treatment effects using the matched sample.</td>
<td>Standard matching.</td>
</tr>
</tbody>
</table>

As with the latent Dirichlet allocation (LDA) model (Blei, Ng and Jordan, 2003), we adopt a simple ‘bag of words’ language model that for each token $l$ in a document $i$, first samples a topic $z_{i,l}$ from the document-specific distribution $\theta_i$ and then samples the observed word from a topic specific distribution over the vocabulary. Unlike LDA, the distribution over the vocabulary is document-specific such that $B_i$ is a $k$ by $v$ matrix. We represent $z_{i,l}$ as a one-hot-encoding column vector so that $z_{i,l}B_i$ returns a $v$ length vector giving the distribution over the vocabulary for the particular token’s topic. Thus each token is generated by,

\begin{align}
    z_{i,l} &\sim \text{Multinomial}_k(\theta_i), & \text{for } l = 1 \ldots L_i, \\
    w_{i,l} &\sim \text{Multinomial}_v(z_{i,l}B_i), & \text{for } l = 1 \ldots L_i,
\end{align}

STM allows each document to have an individual prior for $\theta_i$ based on topic prevalence covariates but for notational simplicity we consider the shared global prior:

\[
    \theta_i \sim \text{LogisticNormal}_{k-1}(\mu, \Sigma),
\]

where $\theta_i$ is on the $k - 1$ dimensional simplex and $\Sigma$ is a global $k - 1$ by $k - 1$ covariance matrix. Including topic prevalence covariates allows the model to share information about the value $\theta_i$ across different documents with similar covariate values and enters the model by parametrizing $\mu$ (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 $B_i$ as

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

(4)

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

In the process of estimating the topic model, the user selects the number of topics $k$ which 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 essential that the topics themselves are interpretable, only that they capture the density accurately. However, interpretable topics can be particularly useful as they allow the analyst to understand the dimensions along which they should expect documents to be similar and can facilitate selective matching on topics (where some topics are more tightly matched or others are not matched at all).

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

### 3.1.2 TIRM Step 2: Extract Topic Proportions

The parameter $\theta_i$ provides a measure of the document’s topical content. To ensure topics are comparable irrespective of treatment and control differences, we re-estimate 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 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 (such as treating years 9-12 of schooling as ‘high-school’), topics coarsen across variables (such as treating ‘tax’, ‘tariff’ and ‘economic’ as part of an economy
When we have estimated more topics, $\theta_i$ will be longer, facilitating a more fine-grained match of substantive content. On the other hand, fewer topics will result in a shorter $\theta_i$, which will treat more words as equivalent and create a more coarse match.

### 3.1.3 TIRM Step 3: Extract Treatment Projection

While the refit $\theta_i$ 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). The Supplemental Information B provides additional details, properties and alternative strategies. Here we briefly overview the form of the projection for document $i$ which we denote $\rho_i$.

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 simply the sum of the document’s weighted word counts and then normalized by document length. For a given level of the content covariate denoted by $t$,

$$\rho_{i,t} = \frac{1}{L_i} \left( \sum_{l=1}^{L} w_{i,l}^{(cov)} \kappa_{l,c}^{(cov)} + \sum_r w_{i,l}^{(int)} \frac{I(z_{i,l} = r)}{k^{(int)}} \kappa_{l,r,c}^{(int)} \right)$$

where each $w_{i,l}$ is a one-hot encoding which indicates the observed word at token $l$. In practice, we don’t observe the true topic indicators $z_{i,l}$ and so we use their posterior means. Each document has a projection value for each level of the treatment (thus in our binary treatments, two values) which 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.

### 3.1.4 TIRM Step 4: Matching and Treatment Effect Estimation

In the final step, we match on both the STM projection and the estimated topic proportions from the STM, which ensures that matches are both topically similar and have similar within-

---

10 Topic models are even more nuanced than this example suggests because topics are estimated in a context-specific way so that if the word ‘bat’ is included with ‘pitcher’ and ‘batter’ it can be assigned to a baseball topic and if with ‘cave’ and ‘fly’ it can be associated with an animal topic.

11 If we were to omit topic-covariate interactions (an option in the stm package), the projection would take the simple form $\frac{1}{L_i} \sum_l w_{i,l}^{(cov)} \kappa_{l,c}^{(cov)}$ which has the advantage that the weights don’t depend on the topics directly. This projection was explored in work by Rabinovich and Blei (2014) on the inverse-regression topic model where it is used as a summary for prediction.
topic probabilities of treatment. For example, this assures that matched blog posts in our Chinese censorship data will be on the same topics and have the same within-topic probability of treatment. Outside of matching on topic proportions and the STM projection, researchers can also match on other pre-treatment covariates in this step. For example, in our Chinese censorship example we also match on the day the post was written, previous posting rate and previous censorship rate. While our framework allows any matching algorithm to be used at this final stage, we generally prefer using CEM if pruning treated units is acceptable.

In most matching algorithms there are choices the analyst must make beyond the choice of which variables to match on. For CEM, this involves choosing the level of coarsening. Software packages implementing CEM make automated choices using histogram binning algorithms. In addition to automated binning, for topics, we have had success with bins which capture whether a topics is substantially present or absent; for example, by using two bins — one ranging from 0 to .1 and another ranging from .1 to 1. With a moderate to large number of topics this is still a quite demanding criterion for a match as CEM requires matched units to share the same bin across all variables. We use the automated binning with five levels or more for the projections. A larger number of bins with more fine-grained distinctions allows for closer matches (and thus possibly less bias) at the expense of pruning more treated units. These choices are necessarily application-specific, depending particularly on how different documents are from each other and how many units the analyst has available for matching.

Using the matched sample, we fit a model predicting the outcome as a function of treatment status and possibly other controls (Ho et al., 2007) to estimate the effect of treatment. Inference following matching procedures has been subject to substantial debate. Initially, some applied researchers tried bootstrap confidence intervals. Abadie and Imbens (2008) show that the bootstrap results in inconsistent estimates of the treatment effect standard error, and propose an asymptotically consistent alternative for some matching settings (Abadie and Imbens, 2006). Yet standard practice in political science is to use the standard errors from the analysts 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 the analyst is willing to change their axiom of inference from simple random sampling to stratified sampling. We encourage further work on this issue, but it is beyond the scope of this article to resolve.
3.2 Balance Checking

Balance checking — confirming that matched units are in fact similar on pre-treatment confounders — is important for assessing whether matching is successful. However, checking balance can be difficult with high-dimensional confounders measured from text. This is because there is no reason to believe there is a universally best metric for text similarity. We recommend several procedures to check balance following text matching.

First, we check whether words that predict treatment in the unmatched sample are balanced in the matched sample. We also recommend verifying that the distribution of topics in treated and control documents are 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. However, our hope is that the model has picked up something deeper about the text which cannot be directly evaluated in a balance check of the topics themselves. We deal with this uncertainty 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, 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 improves balance in a metric which respects that ordering. In this project we use simple graphical diagnostics (see Supplemental Information Figure 5) but future work could develop formal hypothesis tests.

Finally, we recommend close reading of matched documents. A crucial advantage of our matching approach is that it allows experts to directly scrutinize the claim that matched text are sufficiently similar. Evaluation through reading is subjective, but can help analysts judge whether they believe that the texts are sufficient for identification and whether balance has been achieved in practice. Tables 4 in the Supplemental Information and 6 provide some examples.

\[12\]

Like all topic models, STM estimates the proportion of a document allocated to each topic. This means that two documents of different lengths can be matched together because they allocate a similar proportion of their length to the same topics. If document length is a concern, it can be explicitly included as an additional confounder in the matching algorithm.
though we recommend examining more pairs than we have space to present there.

### 3.3 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 which 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 \texttt{stm} software (Roberts, Stewart and Tingley, 2017).

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 documents. We have found topic models to work well for this purpose and 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 which are more attuned to particular types of data (for example 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 in the context of the analyst’s application. Another limitation is that like other conditioning approaches to causal inference, TIRM also requires that the selection on observables assumption is met. In the applications below, we show that applying TIRM to text data makes this crucial assumption substantially more plausible when confounders are potentially in the text itself. Finally, we emphasize that when using approaches which drop units the estimand is changing. Particularly when dropping treated units it is necessary to carefully 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.

### 3.4 Related Work

Before moving to applications of TIRM, we briefly mention how it relates with other approaches to similar problems. To date, the matching literature has focused on the problems of high-dimensional data for estimating propensity scores (Hill, Weiss and Zhai, 2011; Belloni, Chernozhukov and Hansen, 2014). While helpful in other settings, we do not find these approaches useful for text matching because it produces 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). The most closely related work is the recent study by Mozer et al. (2018) which builds on our framework to propose an alternative text matching approach.

There is little work in the matching framework that proposes matching on a density estimate, as we do here. Price et al. (2006) reduces the dimensionality of genotype data with an eigen decomposition, but follows 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).

Finally, Egami et al. (2017) provide a framework for causal inference with text as treatment or outcome, complimenting our discussion of text-based confounding. See the Supplemental Information for further related work.

### 4 Applications and Simulations

To demonstrate the effectiveness of our text matching approach, we present a two-part validation study in Sections 4.1 and 4.2 that builds on previous work by Maliniak, Powers and Walter (2013) to estimate the effect of perceived author gender on the citation counts of academic journal articles in the discipline of International Relations, controlling for the text of the original article. Maliniak, Powers and Walter (2013) control for text confounding using hand-coded categories based on the text of international relations articles. In Section 4.1 we use this hand-coded data to produce a simulated dataset that we use to study the performance of our proposed estimator in recovering a known truth. In Section 4.2, we show that we recover a result using text matching that is similar to the Maliniak, Powers and Walter (2013) analysis without the use of their hand-coded data. In Section 4.3 we demonstrate the use of our methods in our motivating example.
of studying the effects of government censorship on Chinese social media users.

4.1 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? Maliniak, Powers and Walter (2013) say yes. Obtaining credible answers to this question is not straightforward with observational data. On average, male and female authorial teams 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 the 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 TRIP Journal Article Database to control for the broad sub-field of each article, the issue areas covered, the general methodology, paradigm, and epistemology. They find that academic articles by female authors in IR have lower citation counts than articles by men or groups of men and women, 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. 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.

13 We are estimating the effect of perceived author gender in the minds of other authors making citation decisions. This is a more tractable question for causal inference than whether an article would be cited more if the author's gender could somehow be directly manipulated.

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

15 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.

16 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.
In order to design a simulation which 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.

\[ T_i \sim \text{Bernoulli}\left(\pi = .1X_i + .25(1 - X_i)\right) \]  
\[ Y_i \sim \text{Normal}\left(\mu = .5X_i - .2T_i, \sigma^2 = .09\right). \]

where \( Y, T \) and \( X \) are, respectively, the outcome, 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 the observed text \( W_i \). We then use only information about the text \( W_i \) in TIRM to adjust for the unobserved confounder \( X_i \), to see if TIRM can use the text to identify the form of confounding coming from the human labeled category ‘quantitative methodology.’ We believe this simulation is a difficult test of TIRM’s performance because it must recover a process of unobserved confounding from real texts. The Supplemental Information section C offers additional details on the simulation and discusses some of the strengths and weaknesses of this design.

We produce 1000 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 a treatment effect estimate and 95% confidence interval for five estimators: the true linear model using the unobserved quantitative methodology variable, the TIRM model,\(^{17}\) 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 1000 simulations.

The TIRM model performs substantially better than matching only on the topics or the projection particularly in terms of bias and coverage. TIRM has on average substantially fewer matched treated units and the resulting higher variance of the estimator yield slightly worse mean squared error than matching on the topics alone. While TIRM’s 86% coverage rate on

\(^{17}\)For each simulated data set we apply the TIRM procedure using 15 topics and matching with CEM on the projection scores using 8 automatically generated bins and the topics using two bins each (0-.1 and .1-1).
Figure 1: Results from 100 simulations using real texts and hand-coded ‘quantitative methodology’ variable along with a simulated treatment and outcome (according to Equations 6 and 7). The plot shows the estimates and 95% confidence intervals from 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). Colors indicate whether the interval does or does not cover the truth (denoted by a dashed grey line at -.2). TIRM achieves the best coverage out of the models we evaluate.
this example does not match the nominal 95%, it performs well given that the true confounder has been withheld from the model.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Bias</th>
<th>Coverage</th>
<th>Avg. # Treated</th>
<th>Avg. # Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Model</td>
<td>0.00018</td>
<td>0.00032</td>
<td>0.955</td>
<td>619.6</td>
<td>3201.0</td>
</tr>
<tr>
<td>TIRM</td>
<td>0.00157</td>
<td>-0.02318</td>
<td>0.866</td>
<td>217.4</td>
<td>692.0</td>
</tr>
<tr>
<td>Topic Matching</td>
<td>0.00144</td>
<td>-0.03377</td>
<td>0.547</td>
<td>548.8</td>
<td>2142.8</td>
</tr>
<tr>
<td>Matching on Projection</td>
<td>0.00312</td>
<td>-0.03220</td>
<td>0.456</td>
<td>496.2</td>
<td>2693.4</td>
</tr>
<tr>
<td>Naive Difference-in-Means</td>
<td>0.01299</td>
<td>-0.11278</td>
<td>0</td>
<td>619.6</td>
<td>3201.0</td>
</tr>
</tbody>
</table>

Table 2: Demonstration of TIRM estimator performance and relevant benchmark estimators across 1000 simulations. 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.

4.2 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? 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 show that we can recover similar effects with TIRM using far less manual effort.

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 compare the performance of TIRM to results from other approaches: matching only on propensity scores, matching only on topics, exact matching on the original human-coded data.

It is no surprise that we find women’s and men’s writings are not entirely the same in the unmatched data set. Word usage by women and men in IR varies in predictable ways. For example, women are more likely to write about gender (“women”, “gender”, and “children”), while men or male/female co-authored articles are more likely to use words associated with statistical methods (“model”, “estimate”, and “data”). We also find substantial imbalance in the human coded substance variables, as well as in estimated topics from STM. We evaluate whether TIRM has adequately reduced these differences in the matched sample in a variety of ways.

Our first approach to checking balance in the matched sample produced by TIRM is to focus on 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. Figure 2 shows the relationship between the difference in word occurrence by gender and the mutual information of each word in each dataset. If perfect balance on all words were possible, we would hope to see every word lined up vertically on the $x = 0$ line (and shaded blue 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 this is unimportant because balancing these words does not substantially reduce 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 a variety of 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 Table 4 in the Supplemental Information 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 topics estimated by the topic model well. Matching on topics only (without the TIRM projection) performs slightly better on this metric, but that is to be expected because TIRM is trying to achieve balance on the estimated topics and the probability of treatment as the same time. Finally, we check string kernel similarity of the matched data sets and find that TIRM out performs the alternatives and offers substantial improvement over the raw data. Details for all of these balance checks are in the Supplemental Information D.

We re-estimate the models of Maliniak, Powers and Walter (2013) using the TIRM matched

---

18 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 (2018) for a longer treatment of mutual information.
sample and find that gender differences in citations are even more pronounced than those originally reported. We find an average 16 fewer citations to articles written by women, when compared to as-identical-as-possible articles written by men or mixed-gender groups. Most of
this effect seems concentrated among the highest citation-earning articles. Because this example is focused on the method, rather than the finding, we refer readers to the Supplemental Information D for further results and a sensitivity analysis.

4.3 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; Marolt, 2011; MacKinnon, 2011), with technologies ranging from manipulating search results to blocking foreign websites. Even as we learn more about the types of content which is censored in practice and the technical infrastructure enabling censorship (Bamman, O’Connor and Smith, 2012; King, Pan and Roberts, 2013, 2014), we continue to 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 as the targeted use of artificial intelligence for the purpose of censorship has become a widespread cause of concern, as it further complicates the ability for the public to hold the government accountable for censorship decisions (Morozov, 2012; Tufekci, 2014; Roberts, 2018).

Our goal is to estimate the causal effect of experiencing censorship on social media users in China on their subsequent censorship and posting rate. Chinese censorship is quite thorough, but occasionally misses sensitive social media posts. 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 a dataset of 4,685 social media users on Weibo from the Weiboscope dataset (Fu, Chan and Chau, 2013).\textsuperscript{19} Fu, Chan and Chau (2013) collected posts from Weibo in real time and then revisited these posts later to observe whether they were censored. Our goal is to identify posts that are about the same event and have a similar ex-ante probability of being 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. We match closely on 100 topics in order to retrieve posts that are highly similar in content. For

\textsuperscript{19}The Weiboscope dataset is too large to conduct this analysis on, and therefore we selected only users who had been censored at least once in the first half of the data to ensure that both treated and control users write on somewhat similar topics to begin with. We then study their behavior only in the last half of the data.
comparison, we compare balance from matching with TIRM to matching only on the topics or only on the projection.\textsuperscript{20}

We find that the TIRM matching is effective at identifying social media posts about the same topic or event, but with different censorship statuses. As we show in Figure 3, TIRM matching outperforms other matching strategies in reducing the difference between topics in censored and uncensored posts. Matching only on topics is a close second, and matching on the projection only slightly outperforms the unmatched dataset. We find that TIRM also outperforms matching only on topics or only on the projection in terms of string kernel similarity of documents in the matched data sets produced by each method (details in the Supplemental Information Figure 8).

We greatly prefer that matched units have not only a similar propensity to be censored but are also similar in their content. Matching only on the projection allows for balance to be achieved in expectation but it is difficult for humans to evaluate the balance by examining the matches.\textsuperscript{21} TIRM allows us to evaluate the closeness of matched documents, as shown in Table 3.

After determining that the propensity score and topic proportions estimated by TIRM produce the best text matches, we add additional covariates: the previous post rate, previous censorship experience and date. We match using CEM to yield a final matched sample of 879 posts.

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. All things equal, 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 probability of future censorship by almost two fold, but does not decrease the number of posts written by the censored user. This suggests one of two scenarios. First, 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 censored. Alternatively, social media users may chafe against censorship and respond by posting increasingly sensitive content that is more likely to be censored.\textsuperscript{22} Either

\textsuperscript{20}We provide more specific details about the data, the topic model, and the matching process in the Supplemental Information.

\textsuperscript{21}This is true for non-text settings as well. Text is different though in two ways: (1) we are less convinced that we have the right representation of the text and thus that traditional balance statistics are meaningful, and (2) we have access to the texts themselves which provide rich additional information about the unit which is generally not accessible for other observational designs.

\textsuperscript{22}Backlash against censorship is commonly known as a “Streisand effect”.
Figure 3: Topic balance comparison: Top panel shows the balance for unmatched, projection-only matched, topic-only matched, and TIRM-matched for the most unbalanced topics in the unmatched dataset. 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.
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*ck 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.

I personally think, it is not a shame of Zhou Jun, but a shame of Chinese media!

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

5 Conclusion

Scholars across the social sciences are finding an increasing number of ways to use text as data. In this paper, we have proposed conditioning on text to address confounding using matching.

25
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 which captures propensity to treatment. To assist applied researchers wishing to make causal inferences with high-dimensional data, we will provide supplemental \texttt{R} code which leverages the \texttt{stm} package (Roberts, Stewart and Tingley, 2017) to implement the matching procedures described in this paper. This general strategy may have application 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 where 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.

As the vast and rapidly growing body of scholarship on matching shows, identifying a single optimal method and articulating all of the theoretical properties of a new matching method is beyond the scope of a single article. We hope that the text matching problem will be the subject of future research (a process already begun by Mozer et al. 2018) and as a field we will continue to develop methods for handling confounding with text and other forms of high-dimensional data.
References


Supplemental Information – Online Only

Adjusting for Confounding with Text Matching

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen

May 26, 2019
# Table of Contents

A Additional Related Work 1

B Treatment Projection 1

C Simulation Details 6

D The Gender Citation Gap: Balance Checking, Results, and Sensitivity Analysis 7

E Details of Chinese Social Media User Analysis 14
A Additional Related Work

In this short section we highlight some additional related work that we were unable to cite in the main text because of space constraints.

One of the most exciting frontiers in causal inference right now is the use of machine learning methods for causal inference including many methods which could be used to adjust for high-dimensional covariates in experimental and observational settings (Van der Laan and Rose, 2011; Bloniarz et al., 2015; Hazlett, 2015; Sales, Hansen and Rowan, 2015; Athey and Imbens, 2016; Athey, Imbens and Wager, 2016; Hartford et al., 2016; Chernozhukov et al., 2017; Ratkovic and Tingley, 2017). Most of these approaches leverage models of the outcome (see also, Rubin and Thomas, 2000; Hansen, 2008). By contrast our approach is focused on the analysis of observational data, falls in the matching framework, and does not use a separate regression model of the outcome data. There has been some work on the particular problem of using high-dimensional data for estimating propensity scores (Schneeweiss et al., 2009; Westreich, Lessler and Funk, 2010; Hill, Weiss and Zhai, 2011; Belloni, Chernozhukov and Hansen, 2014).

Although not about matching, Taddy (2013a) offers an approach that is conceptually related to ours: considering how to select documents for manual coding in supervised learning. Ideally, manual coding should use an optimally space filling design, but this is impractical in high-dimensions. Taddy proposes a topic model followed by a D-optimal space filling design in the lower-dimensional topic space. Both of these approaches share our intuition that if two features in a high-dimensional covariate set commonly co-occur, then they can be treated interchangeably to identify appropriate counterfactual cases.

In balance checking with string kernels we primarily use visual diagnostics here. However, there is a framework for formal hypothesis tests using the Minimum Mean Discrepancy framework developed in Gretton et al. (2012) (Gretton et al., 2007; Sejdinovic et al., 2013; Szabó et al., 2015, see also,). For practical purposes these tests would need to be made more computationally efficient (Zhang et al., N.d.) and altered to reflect the appropriate null hypothesis for a balance test (Hartman and Hidalgo, 2018). Those developments are beyond the scope of this paper.

B Treatment Projection

In this appendix section we derive the the treatment projection discussed in Section 3.1.3. As a reminder of the notation: \( w_{i,l} \) is a one-hot-encoding vector indicating the observed word in doc-
ument $i$ at token $l$. $z_{i,l}$ is a categorical variable indicating the topic of that token and $\kappa$ are word weights (with parenthetical superscripts indicating whether they correspond to parameters for the topics, content covariate, or interaction between topics and content covariates). Equation 5 from Section 3.1.3 provides the projection which is reproduced here:

$$\rho_{i,t} = \frac{1}{L_i} \left( \sum_{l=1}^{L} w_{i,l}^{(\text{cov})} \kappa_{t,c}^{(\text{cov})} \text{weight} + \sum_{r} w_{i,l}^{(\text{inte})} \kappa_{t,r,c}^{(\text{inte})} \text{topic indicator} \right)$$

(8)

Although notationally dense, the projection has the straightforward interpretation of summing up two weights for each word appearing in the document and normalizing by document length. The first weight is specific to the entry of the vocabulary (e.g. parade and protest have different weights) while the second weight is specific to the entry of the vocabulary and that token’s topic (e.g. parade has one weight under Topic 3 but a different weight under Topic 2).

**Connections to Inverse Regression** We arrive at this projection by noting that conditional on the token-level latent variables $z$ (which denote the topic of a given word token), the structural topic model with content covariates has the form of the multinomial inverse regression (MNIR) in Taddy (2013b). In that work, Taddy derives a projection for the MNIR model and proves that it satisfies classical sufficiency for the outcome such that given the model and the parameters, the treatment indicator is independent of the words given the projection (to use our example). Given this low-dimensional representation, Taddy (2013b) then fits the low-dimensional forward regression which predicts the treatment indicator using the projection. We don’t actually need this final step because we are matching on the projection itself (this is roughly analogous to the practice of matching on the linear predictor in the propensity score).

**Rationale for Projection** The rationale for using the projection is the same as in Taddy’s work: efficiency. As explained in the rejoinder to the original paper (Taddy, 2013c), using the inverse regression provides efficiency gains relative to the forward regression which derive from assuming a generative model for the words. Given the generative model, the variance on the projection decreases in the number of words rather than the number of documents. Even when the generative model does not hold, this can provide substantial gains in practice.
**Advantages of Joint Estimation**  In our setting, the inverse regression formulation affords us two additional advantages. First, we can allow words to have different weights depending on their context (as captured through the topics). For example, in the censorship example we can allow for the possibility that certain words may always increase your odds of censorship while others are only sensitive in particular situations. Second, we avoid redundant information between the topics and the words.

**Properties**  In Taddy (2013b) there are no topics or interactions between topics and covariates. Here we observe that his Propositions 3.1 and 3.2 establishing sufficiency of the projection conditional on the parameters of the model (including the document level random effects), extend to our setting as well by conditioning on the token-level latent variables $z_{i,l}$. We show that our model can be written in that form, following closely on Taddy (2013b, page 758)

\[
\begin{align*}
  w_{i,l} &\sim \text{Multinomial}(q_{i,l}, 1) \\
  q_{i,l,c} &= \frac{\exp(\eta_{i,l,c})}{\sum_c \exp(\eta_{i,l,c})} \\
  \eta_{i,l,c} &= \frac{m_c}{\text{baseline}} + \left( \sum_r I(z_{i,l} = r) \kappa_{r,c}^{(\text{topic})} \right) + \sum_a I(T_i = a) \kappa_{a,c}^{(\text{cov})} + \\
  &\quad \sum_a I(T_i = a) \left( \sum_r I(z_{i,l} = r) \kappa_{a,r,c}^{(\text{int})} \right)
\end{align*}
\]

The data $w_{i,l}$ and $q_{i,l}$ are $v$-length column vectors representing the one-hot encoding vector of the observed data in token $l$ and the probability vector that draws it respectively. All other terms are scalars. Entries in the vocab are indexed by $c$ and runs to $v$, the levels of the content covariate is indexed by $a$ and runs to $b$, the topics are indexed by $r$ and runs to $k$.

We can rewrite this in a more compact notation by suppressing the dependence on $i$ and writing $m$ as a $v$-length column vector, $T$ as a $b$-length column vector (one-hot encoding), and $z_l$ as a $k$-length column vector (one-hot encoding). We use $\kappa^{(\text{topic})}$ to denote the $v$-by-$k$ matrix of topic parameters and $\kappa^{(\text{cov})}$ to denote the $v$-by-$a$ matrix of covariate parameters and $\kappa^{(\text{int})}_{r}$ to indicate the $v$-by-$a$ matrix of interaction parameters for the $r$-th topic. We can now write the
\( \eta_l \) as

\[
\eta_l = m + \kappa^{(\text{topic})} z_l + \kappa^{(\text{cov})} T + \sum_{c=1}^{k} \kappa^{(\text{int})}_c T
\]

\[
= m + \kappa^{(\text{topic})} z_l + \Phi' T
\]

where \( \Phi = \kappa^{(\text{cov})} + \sum_{r=1}^{k} \kappa^{(\text{int})}_r \) collects the \( v \)-by-\( a \) matrix of word weights.

Rewriting the components that do not depend on \( T \) as \( \alpha = m + \kappa^{(\text{topic})} z_l \), we can write the likelihood in exponential family form

\[
\exp \left( w'_l \eta_l - A(\eta_l) \right) = \exp \left( w'_l \alpha \right) \exp \left( (w'_l \Phi) T - A(\eta) \right)
\]

\[
= h(w) g(\Phi' w, T)
\]

where \( A(\eta) = \log \left( \sum_c \exp(\eta_c) \right) \) is the log-partition function. The form of the model is now the same as in Taddy (2013b) and the remainder of his proof follows, with standard sufficiency results for the exponential family implying that \( p(w_l | \Phi' w_l, T) = p(w_l | \Phi' w_l) \). Proposition 3.2 follows analogously and establishes that the reduction still holds when we normalize for document length.

**Limitations** As in Taddy (2013b), it is worth emphasizing that sufficiency only holds conditional on the latent variables (in our setting \( z_{i,l} \)). We subsequently include the document level topics \( \theta_i \) when we are doing our matching. This is to say that we don’t invest too heavily in the sufficiency result, rather the results simply provide an intuition for why these word weights would be helpful in understanding the propensity to be treated.

**Alternatives** Taddy (2013b) leaves open the question of the best way to use the latent structure in the projection and there is likely still further work to be done on this point. Rabinovich and Blei (2014) introduce a simpler projection in the context of their inverse regression topic model (IRTM). The model structure is similar to the Structural Topic Model with a content covariate and the predecessor of both models, the Sparse Additive Generative model of text (Eisenstein, Ahmed and Xing, 2011). In IRTM, the probability of observing word \( c \) from topic
\( r \) in document \( i \) is given by

\[
\beta_{i,r,c} \frac{\beta_{r,c} \exp(\Phi_c y_i)}{\sum_c \beta_{r,c} \exp(\Phi_c y_i)}
\]

where \( \beta_r \) is a draw from a Dirichlet distribution, \( \Phi_c \) is a draw from a Laplace distribution and \( y_i \) is a continuous covariate (NB: we have adopted their notation from equation 1 of their paper except for the indices we have changed to match ours). Thus, their representation of a topic is a background topic distribution \( \beta_r \) multiplied by a distortion factor \( \exp(\Phi_c y_i) \) where the argument of \( \exp() \) is sparse, leading the distortion factor to often be 1. We can rewrite the STM model to look more like this in order to clarify the connections,

\[
B_{i,r,c} \propto \beta_{r,c} \exp(m_c + \kappa_{r,c}^{(\text{topic})}) \exp(\kappa_{T_i,c}^{(\text{cov})}) \exp(\kappa_{y_i}^{(\text{int})}) \exp(\Phi_c y_i).
\]

The first section is an alternate way of representing the background topic distribution. In IRTM, each topic \( r \) is a draw from a Dirichlet distribution, in STM it is a log-linear model with a dense vector shared by all topics and a sparse, topic-specific deviation. The second second is mostly the same with the distinction that in IRTM the covariate is continuous in 1-dimension and in STM it is categorical. The third chunk in STM which captures the topic-covariate interaction has no equivalence in the IRTM. The IRTM performs MAP estimation using a stochastic variational EM algorithm.

Rabinovich and Blei (2014) compute the analogous projection from their model \( \frac{w^T \Phi}{L} \) and find that using the projection alone is not very effective for prediction. They instead opt to compute the MAP estimate of the content covariate. Performing the joint optimization of the content covariate and the topics is complicated and Rabinovich and Blei (2014) employ a coordinate ascent estimation strategy that would be even more complicated (and slow) in our setting, but it is a direction to possibly explore in future work. Because we are already conditioning on the topics, we would expect better performance than the initial Rabinovich and Blei (2014) tests on the simple projection.
C Simulation Details

In this appendix we provide the technical details of the simulation which we summarized in Section 4.1. To review the motivating logic, we wanted to avoid simulating new documents because simulating text from the TIRM model itself would make strong and unrealistic assumptions about the world. We know of no way to provide a realistic model for treatment assignment, the outcome, and the confounder at the same time while still also knowing the true causal effect. We opted instead to use the structure of the female IR scholarship application in Section 4.2 to simulate a confounded treatment assignment and outcome using the real texts and hand-coding of a plausible confounder (a binary variable indicating quantitative research methodology). In this appendix we provide additional details on the simulation including the rationale behind our choices and interpretation of our results. We also describe the factors that make the simulation challenging enough to be interesting and those that are still unrealistically simplistic.

We conducted a thousand simulations with each simulation taking approximately 40 minutes to run. Each of the 1000 simulations is linked to a unique seed and can be run independently using the code in our replication archive.

Simulating the Data We preprocessed the 3201 articles in the JSTOR data using the default settings of the \texttt{stm} package’s command \texttt{textProcessor}. We then limited to words which appear in at least 25 documents in order to shrink the vocabulary to a manageable size. We then construct a model where articles on quantitative methodology are treated 10\% of the time and non-quantitative articles are treated 25\% of the time. We then simulate an outcome using the actual hand-coded variable on quantitative methodology ($X_i$) as an unobserved confounder which along with the treatment generates the outcome using a linear model:

\[
T_i \sim \text{Bernoulli} \left( \pi = .1X_i + .25(1 - X_i) \right) \\
Y_i \sim \text{Normal} \left( \mu = .5X_i - .2T_i, \sigma^2 = .09 \right).
\]

Estimation We mirror the estimation choices in our application on this data, fitting a structural topic model with 15 topics using the treatment as a content covariate and matching topics in two bins (less than .1 topic proportion in the document or more than .1 topic proportion in the document) and eight automatically generated bins for both the treatment and control
Strengths and Weaknesses of the Simulation Strategy  The structure of the simulation preserves both the real documents and hand-coding of a category from those documents which might plausibly represent a confounder. The simulation is hard (and thus meaningful) because the confounder is not available to our matching model and there is no guarantee that it is recoverable from the word-count data. We also induced a substantial amount of confounding and noise (as evidence by the strikingly poor performance of the unadjusted estimator). We have relatively few treated units (about 20% of the sample on average).

The simulation is also easy in some important ways that might give us pause in generalizing. The treatment and outcome model are quite simple. We use a binary unobserved confounder because it makes the simulation straightforward and easier to describe but one could imagine constructing a more complicated basis for unobserved confounding. The linear model for the outcome means that the treatment effect is constant. This helps remove any complications from the estimand changing as units are pruned, but also makes things substantially easier for models like TIRM which drop a large number of treated units.

For computational reasons, we compare performance of topic matching and matching on the projection using the fitted TIRM model. This effectively demonstrates what each of these components is contributing to our overall estimate but does not necessarily reflect what would happen if a topic model or balancing score were fit separately to the data. We conducted smaller tests (of 100 simulations) using separately estimated topic models and balancing scores alone and did not observe substantially different results from what we have reported here.

As a practical we are using real texts, we are also limited to only the 3201 articles in our database and thus were not able to study performance as sample size increases.

D  The Gender Citation Gap: Balance Checking, Results, and Sensitivity Analysis

This section reports details about the balance checking and analysis for the application estimating whether there is a gender citation gap in the IR literature.

Balance Checking:Balancing Estimated Topics

We check whether TIRM balanced the estimated topics adequately in Figure 4.
Figure 4: Matching Comparison for Topics

Balance Checking: Kernel Similarity

We then compare the matching based on human coding to TIRM using a string kernel similarity metric. Figure 5 shows the similarity between matched documents in the corpus matched using TIRM and corpus matched exactly on human codes. Overall, TIRM performs as well to the human-coding matching in producing semantically similar documents when measured with string kernel similarity.

Balance Checking: Comparing TIRM and Human Coding

As part of our evaluation of TIRM’s performance in the gender citation gap application, we compare the performance of TIRM to the large-scale human coding effort that produced the TRIP data. Human coding is not necessarily the gold-standard for this application because the coding system was not designed to facilitate estimation of our causal estimand, but if TIRM is performing well, then we believe it should improve balance on the human-coded variables measuring article content.
Figure 6 depicts this comparison where the rows along the y-axis correspond to non-mutually exclusive, human-coded categories from the article text: methodological categories on top and issue-area categories below. To the right of each category label, we plot a bar showing the imbalance of this category by gender of article author in the raw data set. Topic-only matching most reduces imbalance in the human-coded categories, suggesting that the topic model comes closest to mimicking the human coding process. TIRM performs almost as well on most categories, and better on some particularly imbalanced ones, like qualitative methods. This suggests that injecting information about treatment assignment into a topic model helps TIRM improve balance on the most likely confounders without sacrificing overall improvement in topic similarity. In contrast, projection-only matching makes balance worse on many human-coded categories.
<table>
<thead>
<tr>
<th>Method</th>
<th>Qualitative</th>
<th>Counterfactual</th>
<th>Methodology</th>
<th>Policy Analysis</th>
<th>Experimental</th>
<th>Descriptive Analysis</th>
<th>Formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean topic difference (Women−Men)

-0.1 0.0 0.1 0.2

Int'l Security
General
US Foreign Policy
Methodology
Philosophy of Science
IR Theory
History of Discipline
American
Political Theory
Health
Environment
Comparative Foreign Policy
Comparative
Int'l Law
Int'l Organization
Policy Analysis
Methodology
Qualitative
Comparative
Human Rights
IPE
Comparative Foreign Policy
Comparative
Int'l Law
Int'l Organization
Health
Environment

Figure 6: Automated Matching Comparison and Human Categories
Table 4: Summaries of several documents by female IR scholars (left) with matching documents by male IR scholars selected by TIRM (right).
Results

First, we re-estimate the models in Maliniak, Powers and Walter (2013) using the matched data set from automated text matching (Table 5). Following Maliniak, Powers and Walter (2013), we use a negative binomial model to estimate the effects of gender on citations. For robustness checks, we also include the additional covariates provided by Maliniak, Powers and Walter (2013) in the model. In the last model, we include all covariates included in their “Kitchen Sink” model. Some of the covariates are not identified because no variation exists within the matched data sets and so we do not include them in the table below.

We then re-estimate these models on subsets of the sample to determine whether the effect of gender on citations is evenly distributed across papers or concentrated in the subset of papers that earn high citations. We define high-citation earning based on the citations to articles by male and mixed-gender teams (control articles). If the control articles in a stratum have an average of 20 or more citations, we include that stratum in the high citation subset. If the average citation count of the control articles in a stratum is less than 20, we include that stratum in the low citation subset. The results Figure 7 show that the estimated effect of gender on citations is strong in the high citation subset and virtually non-existent in the low citation subset.

Sensitivity Analysis

Regression and matching approaches on observational data rely on the assumption that all confounding is due to the observable factors included in the matching and regression procedures. Sensitivity analysis offers a way to test how robust the findings are to violations of this assumption. We use Rosenbaum’s sensitivity analysis framework based on randomization inference (Rosenbaum, 2002), implemented by Keele (2010). This procedure compares the differences in outcomes between matched pairs in a data set as if treatment were randomly assigned and then calculates how large a confounder would be necessary to eliminate the observed difference. This is done by positing an odds ratio $\Gamma$ that corresponds to the magnitude of the potential unobserved confounder. When $\Gamma = 1$, there is no confounding. As $\Gamma$ increases from 1, the odds of units being treated increase. We do not know the true $\Gamma$, so we instead posit increasing values and see whether our results could be overturned by a relatively small unobserved confounder, or only by a very large one. Matched samples for which $\Gamma$ is higher are considered less sensitive.
Table 5: Maliniak, Powers, and Walter 2013 with Text Matching: Full Matched Sample

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female author(s)</td>
<td>-0.95***</td>
<td>-1.17***</td>
<td>-0.74**</td>
<td>-0.70***</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.28)</td>
<td>(0.29)</td>
<td>(0.22)</td>
</tr>
<tr>
<td>Article Age</td>
<td>0.12</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Article Age(^2)</td>
<td>-0.004**</td>
<td>-0.003***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tenured</td>
<td>1.59***</td>
<td></td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td></td>
<td>(0.22)</td>
<td></td>
</tr>
<tr>
<td>Tenured Female</td>
<td>0.50</td>
<td>-0.93*</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.50)</td>
<td>(0.38)</td>
<td></td>
</tr>
<tr>
<td>Coauthored</td>
<td>-0.42</td>
<td>-0.26</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.32)</td>
<td>(0.25)</td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>0.46**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AJPS</td>
<td></td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IO</td>
<td></td>
<td>1.16***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td></td>
<td>0.72**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISQ</td>
<td></td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JCR</td>
<td></td>
<td>0.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOP</td>
<td></td>
<td>-0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP</td>
<td></td>
<td>1.53***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Qualitative</td>
<td></td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Quantitative</td>
<td></td>
<td>1.14***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Formal</td>
<td></td>
<td>0.68**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Analytical</td>
<td></td>
<td>1.42***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Policy Analysis</td>
<td></td>
<td>-1.24***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Descriptive</td>
<td></td>
<td>-0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method: Counterfactual</td>
<td></td>
<td>2.98***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positivist</td>
<td></td>
<td>0.78***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.29)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Materialist</td>
<td></td>
<td>-0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ideational</td>
<td></td>
<td>-0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>3.57***</td>
<td>3.30***</td>
<td>2.60***</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.20)</td>
<td>(0.69)</td>
<td>(0.84)</td>
</tr>
<tr>
<td>Observations</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-766.29</td>
<td>-762.53</td>
<td>-754.52</td>
<td>-677.81</td>
</tr>
</tbody>
</table>

Note: Negative binomial generalized linear models. *p<0.1; **p<0.05; ***p<0.01
Figure 7: The estimated gender citation gap from 12 regression models on the matched data set selected by TIRM. The low citation subset contains matched strata for which the average citations to men’s articles was less than 20. The high citation subset contains matched strata for which the average citations to men’s articles was greater than or equal to 20.

to potential confounding.

In the matched sample for the Maliniak data, we find that the result we report would be overturned with $\Gamma > 2.2$. This means that an unobserved confounder associated with female authorship by a factor of 2.2 would overturn our result. This is a middling value. Our results are not particularly sensitive, but it is possible to imagine an unobserved factor that might be this strongly correlated with the gender of authors and citation counts. By comparison, the sensitivity analysis of the unmatched data indicates that the difference in citations between treated and control documents would be overturned with $\Gamma > 1.16$, indicating that the unmatched result is highly sensitive to very modest levels of unobserved confounding.

E Details of Chinese Social Media User Analysis

This section contains details of the analysis of Chinese social media users. To create our sample, we identify all users within the Weiboscope dataset (Fu, Chan and Chau, 2013) who were
censored at least once in the first half of 2012. We subset the data to contain only these 4,685 users. We identify all posts that are censored for these users. Fu, Chan and Chau (2013) identify two types of error that could be censorship – one is posts that have a “permission denied” error after being removed and others that have a “Weibo does not exist” error after being removed. Fu, Chan and Chau (2013) determine from their own experiments that the “permission denied” error indicates censor removal all the time, but while the “Weibo does not exist” error is usually censorship, it could also be the user removing the post at their own discretion. To ensure that match posts indicate censorship, we only use “permission denied” posts as treated units and match to control posts that are neither “permission denied” nor “Weibo does not exist”.

After identifying posts that were censored with a “permission denied” message by these users, we subset to posts that are longer than 15 characters. We use 15 characters in order to ensure that there is enough information in the censored post to match the content of the post – many of the posts only include the text “reposting” which does not include enough information to ensure that matches are substantively similar. Because the post volume of the 4,685 users is sufficiently large, we restrict our pool of control posts to those that have a cosine similarity with the censored post of greater than 0.5 and were posted on the same day as the censored post. Thus, the potential control donation sample is all posts that are greater than 15 characters in length that have a cosine similarity to a censored post posted on the same day of greater than 0.5. This leaves us with 75,641 posts from 4,160 users across the last 6 months of 2012, with 21,503 censored posts and 54,138 uncensored posts from which to create matches.

We run a topic model on these 75,641 posts with 100 topics in order to ensure a close match and an indicator of censorship as the content covariate. We extract from this: 1) the estimated topic proportion for each post within our dataset and 2) the estimated projection for each post within our dataset.

To estimate the effect of censorship on future post rate and future censorship, we extract all posts for the users within the dataset for the four weeks before and the four weeks after censorship. For the four weeks before censorship, we calculate the number of posts each user wrote and the number of these that were censored or went missing. For the four weeks after censorship, we calculate the number of posts each user wrote and the number of these that were censored and went missing.

We then proceed with matching. Using coarsened exact matching, we match censored posts
Figure 8: Mean String Kernel Similarity for Matched Posts Randomly Sampled Within Each matched Dataset

to uncensored posts written on the same day that have similar topics and projection scores. In addition, we make sure that matched users have a similar previous posting rate and a similar previous censorship rate by matching on these variables. We also ensure that users are not matched to themselves within a strata.

Balance in terms of topics is described in the main text of the paper. Here we show that string kernel similarity results. Matched posts were randomly sampled within each matched dataset and their string kernel similarity was calculated. Posts matched via TIRM had the highest level of similarity, followed by topic matching, then propensity score matching, and last the unmatched dataset (Figure 8).

Further, we conducted a qualitative comparison of matched posts to ensure that TIRM was retrieving posts that were qualitatively similar. We provide some examples of matched posts in
Table 6.

<table>
<thead>
<tr>
<th>Censored Post</th>
<th>Uncensored Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>#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<em>ck you, you beat us up, bombed us, and still ask us to feed you, why don’t you eat sh</em>t?</td>
<td>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.</td>
</tr>
<tr>
<td>[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.</td>
<td>[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!</td>
</tr>
</tbody>
</table>

Table 6: Translations of example social media posts that were censored (left) with matched uncensored social media posts selected by TIRM (right).

After matching with TIRM, the matched data had 305 censored posts matched to 574 uncensored posts, with a total of 879 matched posts. There was no difference in the censorship rate before the match for matched users – both users who were censored in the matched post and not censored had a previous censorship rate of 0.003. Further, we also matched on the previous history of missing posts, and there was not difference in the history of “Weibo does not exist” messages between users who were censored in the matched post and not censored –
both had a previous missingness rate of 0.22. Last, there was no difference between the number of posts treated and control users posted before the match – on average, treated users post 661 posts in the four weeks before the match, while untreated users post 649. Balance tests for these pre-censorship non-text covariates are provided in Table 7.

<table>
<thead>
<tr>
<th></th>
<th>Mean of Treated</th>
<th>Mean of Control</th>
<th>Difference</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Censorship Proportion</td>
<td>0.003</td>
<td>0.003</td>
<td>&lt;0.0001</td>
<td>0.22</td>
</tr>
<tr>
<td>Previous Not Exist Proportion</td>
<td>0.220</td>
<td>0.220</td>
<td>&lt;0.0001</td>
<td>0.98</td>
</tr>
<tr>
<td>Previous Number of Posts</td>
<td>661.9</td>
<td>649.6</td>
<td>12.4</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 7: Balance Tests (T-tests) For Pre-Treatment Covariates

Even though the matched users posted very similar posts, and had very similar previous censorship rates, we find that their experience with censorship diverges after the match. The “Permission denied” rate of the treated users is approximately twice as large as the censorship rate of the untreated users after the match. Further, the rate of “Weibo does not exist” messages also increases for treated users after the match – treated users have on average 25% of their posts missing in the four weeks after the match, in comparison to control users who only have 20% of their posts missing in the four weeks after the match. This suggests that either treated users are affected by their experience with censorship in a way that inspires them to post more sensitive material after the match or that they are put on a list after being censored that increases their likelihood of future censorship.

Users are more likely censored after experiencing censorship; however, we do not see them posting less after experiencing censorship. On average, treated users post 600 posts in the four weeks after censorship, while untreated users post 588, an insignificant difference. This indicates that despite experiencing more censorship, treated users are not deterred in posting online after being censored.

We use a sensitivity analysis (Rosenbaum, 2002) to estimate the magnitude of unobserved confounding that would overturn the findings. We estimate \( \Gamma \), the factor by which a hypothetical unobserved confounder would have to be associated with treatment to erase the effect. For both outcomes with statistically detectable results, we find that an unobserved factor that increased the odds of treatment by roughly 1.6 would overturn the result (\( \Gamma = 1.62 \) and \( \Gamma = 1.64 \) respectively). It is possible to imagine such a factor, so our results are somewhat sensitive.

We conduct another sensitivity analysis to ensure that our results are not peculiar to the
tuning parameters of the matching algorithm. As shown in Figure 9, we compare the results across many matches of topical bin size – from 2 bins for each topic to 15 bins for each topic.\textsuperscript{23} The plots show a high level of consistency across bins – while the rate of censorship consistently increases for both “permission denied” and “Weibo does not exist” posts, the total number of posts written by censored and uncensored users is not different after the match.

\textsuperscript{23}We allow CEM to automatically construct the bins.
Figure 9: Sensitivity Analysis for Censorship Results, Varying Topical Bin Size. Numbers next to each confidence interval indicate sample size. Top Panel: Effect of Censorship on “Permission Denied” rate. Middle Panel: Effect of Censorship on “Weibo Does Not Exist” rate. Bottom Panel: Effect of Censorship on Posting Rate.
References


