Supplemental Information (Online Only)

Adjusting for Confounding with Text Matching

January 14, 2020

Table of Contents

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, Forthcoming; 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 (2013*a*) 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.

There are several recent lines of work considering embeddings and density estimates as proxies for unobserved confounding (Veitch, Wang and Blei, 2019; Sridhar and Getoor, 2019; Wang and Blei, 2019; Yao et al., 2019; Tran and Blei, 2017). These are also connected to broader examinations of the role of text representations in causal inference (Egami et al., 2017; Wood-Doughty, Shpitser and Dredze, 2018).

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 Discrep-

ancy framework developed in Gretton et al. (2012) (see also, Gretton et al., 2007; Sejdinovic et al., 2013; Szabó et al., 2015). 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 document i at token l. $z_{i,l}$ is a categorical variable indicating the topic of that token and κ 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:

$$\rho_{i,t} = \frac{1}{L_i}\left(\sum_{l=1}^{L}\left(w'_{i,l}\underbrace{\kappa_{t,c}^{(cov)}}_{\text{weight}} + \sum_{r}^{k} w'_{i,l}\underbrace{\overbrace{I(z_{i,l}=r)}^{\text{topic indicator}}\,\kappa_{t,r,c}^{(int)}}_{\text{topic-specific weight}}\right)\right) \tag{8}$$

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 (2013*b*). 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 (2013*b*) 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, 2013*c*), 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 (2013*b*) 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 (2013*b*, page 758)

$$
\begin{aligned}
w_{i,l} \sim & \text{Multinomial}(q_{i,l}, 1) \\
q_{i,l,c} = & \frac{\exp(\eta_{i,l,c})}{\sum_c \exp(\eta_{i,l,c})} \\
\eta_{i,l,c} = & \underbrace{m_c}_{\text{baseline}} + \underbrace{\left(\sum_r^k I(z_{i,l} = r) \kappa_{r,c}^{(\text{topic})} \right)}_{\text{topic}} + \underbrace{\sum_a^b I(T_i = a) \kappa_{a,c}^{(\text{cov})}}_{\text{treatment}} + \\
& \underbrace{\sum_a I(T_i = a) \left(\sum_r^k I(z_{i,l} = r) \kappa_{a,r,c}^{(\text{int})} \right)}_{\text{treatment-topic-interaction}}
\end{aligned}
$$

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 $\kappa_r^{(\text{int})}$ to indicate the v-by-a matrix of interaction parameters for the r-th topic. We can now write the v-length vector η_l as

$$\begin{aligned}\eta_l &= m + \kappa^{(\text{topic})\prime} z_l + \kappa^{(\text{cov})\prime} T + \sum_{c=1}^{k} \kappa_r^{(\text{int})\prime} T \\ &= m + \kappa^{(\text{topic})\prime} z_l + \Phi' T\end{aligned}$$

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

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

$$\begin{aligned}\exp\left(w_l'\eta_l - A(\eta_l)\right) &= \exp\left(w_l'\alpha\right)\exp\left((w_l'\Phi)T - A(\eta)\right) \\ &= h(w)g(\Phi' w, T)\end{aligned}$$

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 (2013*b*) 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 (2013*b*), 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 θ_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 (2013*b*) 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

$$\beta_{i,r,c}\frac{\beta_{r,c}\text{exp}(\Phi_c y_i)}{\sum_c \beta_{r,c}\text{exp}(\Phi_c y_i)}$$

where β_r is a draw from a Dirichlet distribution, Φ_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 (β_r) multiplied by a distortion factor ($\text{exp}(\Phi_c y_i)$) 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,

$$B_{i,r,c} \propto \overbrace{\text{exp}(m_c + \kappa^{(\text{topic})}_{r,c})}^{\beta_{r,c}} \underbrace{\text{exp}(\kappa^{(\text{cov})}_{T_i,c})}_{\text{exp}(\Phi_c y_i)} \overbrace{\text{exp}(\kappa^{(\text{int})}_{y_d,k,v})}^{\text{no equivalent}}.$$

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 ($\frac{w'\Phi}{L}$) 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 `stm` package's command `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 largely 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 the treatment projection.

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. Because we use 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. For balance checking in this application we are able to include comparisons based on human-coding of the articles from the Teaching Research and International Policy (TRIP) article database. The collection of this data involved a large-scale human coding effort that will not be generally available in other applications, but provides us a chance to probe the performance of our method. We do however note that this 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. However if TIRM is performing well, then we believe it should improve balance on the human-coded variables measuring article content. When presenting comparisons to matches based on human coding we are matching on 35 variables which includes the methods variables (see Fig 6), issue areas (see Fig 6), paradigms (realist, liberal, constructivist, atheoretical, nonparadigmatic, marxist) and epistemic orientation variables (positivist orientation, material variables, ideational variables) from the TRIP dataset.

Balance Checking: Balancing Estimated Topics

We check whether TIRM balanced the estimated topics adequately in Figure 4. Our primary comparison is between TIRM, projection matching (propensity score only), and topic matching. We find that topic matching performs best at balancing the estimated topics, as we would expect. Matching only on the projection generally does not improve balance on the topics – in fact balance on many topics gets notably worse. TIRM performs almost as well as topic matching, despite also balancing on the propensity score, which we see as evidence of TIRM's effectiveness in combination with the balance checks below. We also include a secondary comparison to the balance we obtain when we use matching on the human-coded variables. Matching on human-coded variables generally improves

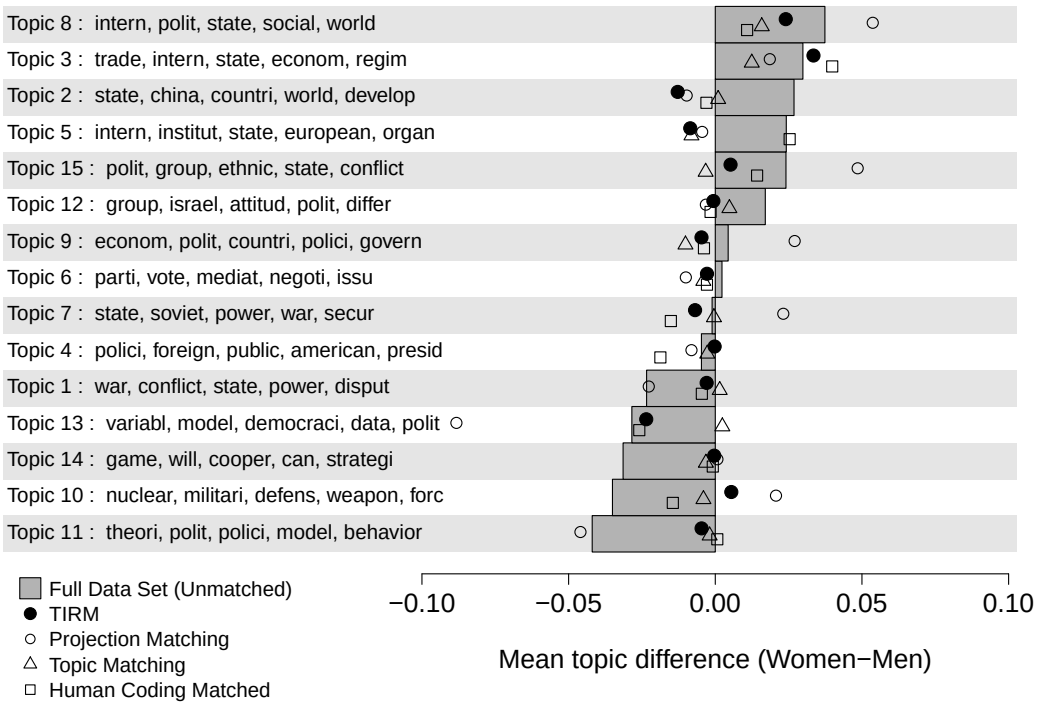

Figure 4: Balancing Estimated Topics

balance on the topic model topics, but not in all cases.

Balance Checking: Kernel Similarity

We also 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

Figure 6 evaluates balance on the human coded categories. 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

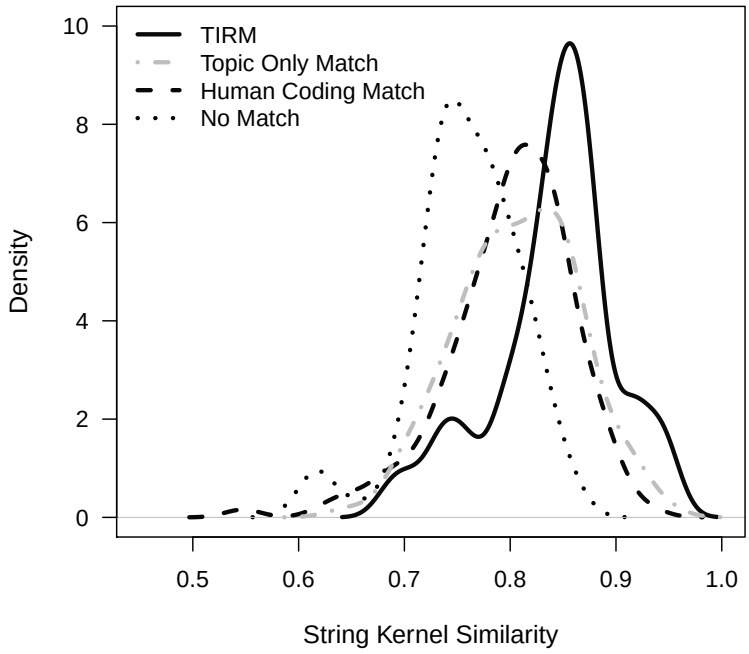

Figure 5: String Kernel Similarity Comparison

in the raw data set. TIRM performs reasonably well at balancing the human-coded categories, especially on the variables that were initially most imbalanced in the full sample: qualitative methodology, formal methodology, quantitative methodology, and the issue area of international security, suggesting that TIRM comes closest to mimicking the human coding process. Topic matching also substantially reduces imbalance in many of the human-coded categories. In contrast, projection-only matching makes balance worse on several of the most imbalanced human-coded categories (formal methods and quantitative methods) and does not improve balance much on several others (qualitative methods and international security).

It's true that the other methods also occasionally make imbalance worse on some of the human-coded variables. TIRM and topic matching both increase imbalance in articles using descriptive methodology, for example. However, we are more concerned about correcting extreme imbalances present in the full data set because we *ex ante* believe they will produce the greatest bias in our estimates. On these, TIRM and topic matching outperform projection matching.

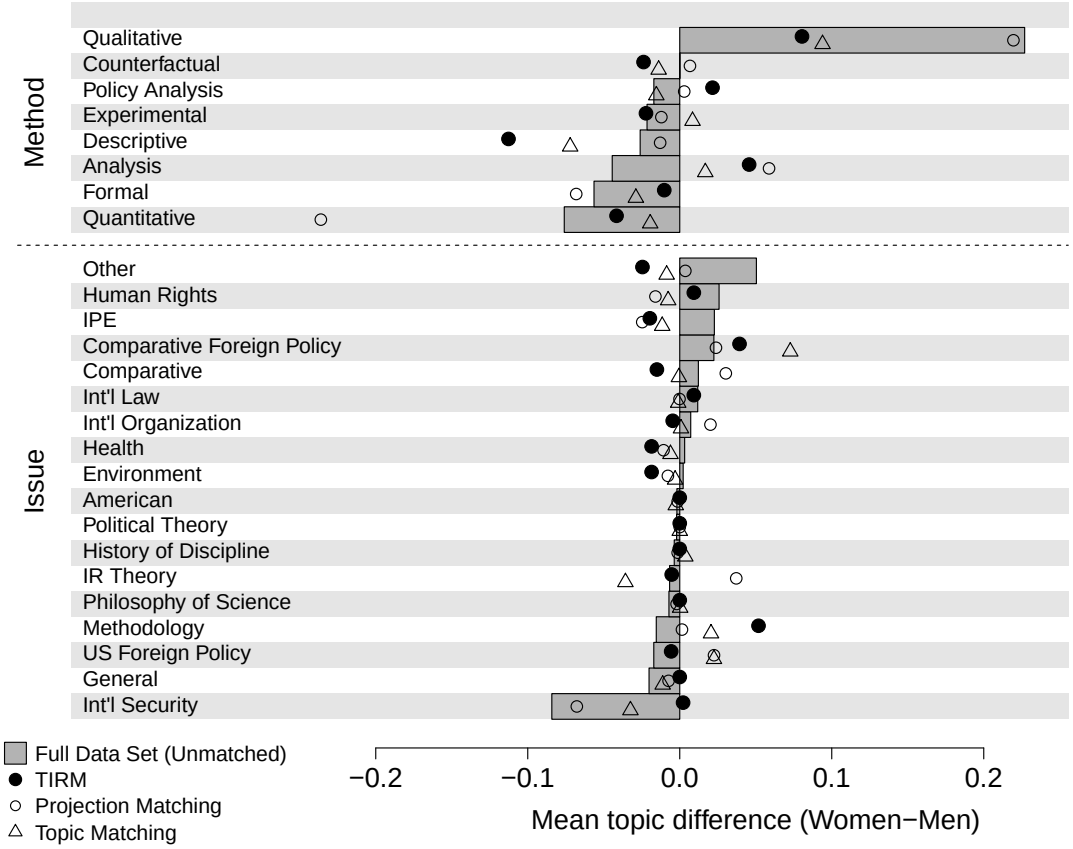

Figure 6: Automated Matching Comparison and Human Categories

Balance Checking: Comparing matched pairs

Treated Document	Matched Control Document
"Democratic Synergy and Victory in War, 1816-1992." Ajin Choi. 2004. *International Studies Quarterly.* **Abstract:** "This study investigates the question of why democracies are more likely to win wars than non-democracies. I argue that due to the transparency of the polities, and the stability of their preferences, once determined, democracies are better able to cooperate with their partners in the conduct of wars, and thereby are more likely to win wars. In support of my argument, the main findings in this study show that, other things being equal, the larger the number of democratic partners a state has, the more likely it is to win; moreover, democratic states are more likely to have democratic partners during wars. These results are in contrast with those in current literature about the high likelihood of prevailing by democracies in wars, which emphasize, on the one hand, the superior capacity of democratic states to strengthen military capabilities and, on the other hand, to select wars in which they have a high chance of winning."	"Third-Party Interventions and the Duration of Intrastate Conflicts." Patrick M. Regan. 2002. *Journal of Conflict Resolution.* **Abstract:** "Recent research has begun to focus on the role of outside interventions in the duration of civil conflicts. Assuming that interventions are a form of conflict management, ex ante expectations would be that they would reduce a conflict's expected duration. Hypotheses relating the type and timing of outside interventions to the duration of civil conflicts are tested. The data incorporate 150 conflicts during the period from 1945 to 1999, 101 of which had outside interventions. Using a hazard analysis, the results suggest that third-party interventions tend to extend expected durations rather than shorten them. The only aspect of the strategy for intervening that reduces the likelihood that a conflict will end in the next month is that it be biased in favor of either the opposition or the government. In effect, neutral interventions are less effective than biased ones."
"The Present As Prologue: Europe and Theater Nuclear Modernization." Catherine McArdle Kelleher. 1981. *International Security.* **Introduction:** "More than a year after formal Alliance decision, controversy still surrounds NATO's plan for new long-range theater nuclear force (LRTNF) deployments. The controversy focuses both on the substance and the process involved. Proponents, in Washington as elsewhere, see the decision as an Alliance success, and the process as a model for future decision-making. The Alliance has now demonstrated it can meet new Soviet challenges; the exhaustive consultation procedures did lead to a genuinely informed NATO consensus despite the inherent political risks. Although not designed to match Soviet LRTNF capabilities, ground-launched cruise missiles (GLCMs) and Pershing IIs do provide a new element in the overall East-West military balance. And there has been due, measured common attention to the problems both of reinforcing American strategic linkage, and of pursuing opportunities for East–West limitations on LRTNF deployments. "	"Counterforce: Illusion of a Panacea." Henry A. Trofimenko. 1981. *International Security.* **Introduction:** "In recent years, American military strength has been moving in a vicious cycle. It has been unable to get out of the impasse created by Washington's desire to outstrip the Soviet Union in strategic arms and by the practical impossibility of achieving this aim. The Soviet Union is fully resolved not to fall behind the United States, nor to permit such U.S. preponderance. Recent evidence of this strategic merry-go-round was provided by the Carter Administration's Directive No. 59, which returned U.S. strategy to the concepts of counterforce nuclear targeting. This policy concludes the long campaign in the U.S. media and academic press that was intended to frighten the Americans with a purported Soviet counterforce and "war-winning" threat. "
"Locating "Authority" in the Global Political Economy." A. Claire Cutler. 1999. *International Studies Quarterly.* **Abstract:** "This article addresses the problematic nature of "authority" in the global political economy. Focusing on the rules governing international commercial relations, which today form part of the juridical conditions of global capitalism, the location and structure of political authority are argued to be historically specific. They have changed with the emergence of different historic blocs and as a result of consequent alterations in state-society relations. The article emphasizes the significance of private corporate power in the construction of the global political economy and hegemonic authority relations. However, the significance of private authority is obscure and little understood by students of international relations. This gives rise to analytical and normative grounds for adopting a historical materialist approach to the analysis of global authority that incorporates national, subnational, and transnational influences."	"South Korean and Taiwanese Development and the New Institutional Economics." David C. Kang. 1995. *International Organization.* **Introduction:** "The publication of books by both Alice Amsden and Robert Wade provide an opportune moment to reflect on the study of East Asian development. After an inital surge of interest beginning in the 1970s, the field has reached a plateau, and scholars recently have cast a wide net searching for ways to extend the field. In assessing the "state of the art" regarding economic development of the East Asian newly industrialized countries (NICs), this review will treat three themes. First, I will argue that the focus on states versus markets is becoming stale and much scholarly interest lies in the politics behind the economics. Second, I argue that political scientists have underexplored the historical origins of Korean and Taiwanese capitalism and that such attention promises to strengthen both theories and explanations of development. Third, I argue that the international system has been more important in promoting development in East Asia than accounts in the "first wave" have recognized."

Table 4: Matched pairs of documents (female authorial teams on left).

Results

We estimate negative binomial regression models using similar specifications to those in Table 2 of Maliniak, Powers and Walter (2013). Our results are in Table 5. Column 1 of Table 5 shows the results of a specification with no conditioning variables other than the TIRM matching. Columns 2 and 3 add progressively more control variables, but only includes those that could be coded easily, without human reading. Column 4 adds the human-coded variables for research orientations, paradigms, and methodology. Column 4 corresponds closely to the "kitchen sink" model in Maliniak, Powers and Walter (2013), but omits a handful of variables that do not vary in our matched sample.

These results support the conclusion of Maliniak, Powers and Walter (2013) that there is a detectable gender citation gap in the IR literature. Maliniak, Powers and Walter (2013, 906) estimate the gap to be about 4.7 citations, while our equivalent model in column 4 estimates this gap to be 6.5 citations. Our estimate may not be directly comparable to theirs because we use coarsened exact matching which discards treated units and changes the quantity of interest from the SATT to FSATT (King, Lucas and Nielsen, 2017). Our result show that the gender citation gap persists in the subset of women's and men's articles that are textually similar according to TIRM. We cannot extrapolate our estimate to the broader population of IR articles without additional assumptions, and we don't undertake that extrapolation exercise here. This means our findings are not definitive evidence of the magnitude of the gender citation gap in the overall population of IR articles. However, they do show that evidence of a gap persists even when we condition on the text of the articles in ways Maliniak, Powers and Walter (2013) could not.

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 pro-

Table 5: Maliniak, Powers, and Walter 2013 with Text Matching

	Dependent variable:			
	Citation Count (source: SSCI)			
	(1)	(2)	(3)	(4)
Female author(s)	−1.02***	−0.91***	−0.92***	−0.59***
	(0.25)	(0.30)	(0.28)	(0.22)
Mixed gender authors		−0.49	0.36	0.46
		(0.69)	(0.69)	(0.53)
Article age		0.08	0.05	0.07
		(0.08)	(0.08)	(0.06)
Article age^2		−0.003	−0.003	−0.003**
		(0.002)	(0.002)	(0.002)
Tenured		1.28***	0.96***	0.33
		(0.27)	(0.27)	(0.23)
Tenured female		−0.55	−0.29	0.34
		(0.50)	(0.47)	(0.38)
Coauthored		−0.12	−0.84**	−0.56*
		(0.37)	(0.40)	(0.31)
R1		0.38*	0.45**	0.09
		(0.22)	(0.22)	(0.18)
AJPS			0.70	1.14*
			(0.81)	(0.65)
APSR			1.56**	1.39***
			(0.68)	(0.54)
IO			1.08***	0.40*
			(0.28)	(0.23)
IS			−0.48	−0.26
			(0.33)	(0.28)
ISQ			−0.49	−0.11
			(0.40)	(0.34)
JCR			0.62*	0.76**
			(0.37)	(0.36)
Positivist				0.60*
				(0.31)
Materialist				0.86
				(0.65)
Ideational				−0.22
				(0.19)
Paradigm: Atheoretical				−0.69*
				(0.40)
Paradigm: Constructivist				1.53***
				(0.56)
Paradigm: Liberal				−0.52**
				(0.21)
Paradigm: Realist				0.07
				(0.41)
Method: Qualitative				1.27***
				(0.37)
Method: Quantitative				0.82**
				(0.37)
Method: Formal theory				0.67*
				(0.36)
Method: Analytical				1.42***
				(0.49)
Method: Description				0.82*
				(0.45)
Method: Counterfactual				1.47***
				(0.57)
Constant	3.60***	2.79***	2.88***	0.37
	(0.13)	(0.72)	(0.72)	(0.98)
Observations	181	181	181	181
Log Likelihood	−749.84	−738.62	−723.23	−665.45

Note: Negative binomial generalized linear models. *p<0.1; **p<0.05; ***p<0.01

cedure 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 Γ that corresponds to the magnitude of the potential unobserved confounder. When $\Gamma = 1$, there is no confounding. As Γ increases from 1, the odds of units being treated increase. We do not know the true Γ, 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 Γ is higher are considered less sensitive to potential confounding.

In the matched sample for the Maliniak data, we find that the result we report would be overturned with $\Gamma > 1.9$. This means that an unobserved confounder associated with female authorship by a factor of 1.9 would overturn our result. This is a middling value: it is possible to imagine an unobserved factor that might be this strongly correlated with the gender of authors and citation counts but the effect of this unobserved factor would have to be somewhat large. By comparison, the sensitivity analysis of the unmatched data indicates that the difference in citations between treated and control documents would be overturned with $\Gamma > 1.24$, indicating that the unmatched result is 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 that 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 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 7).

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 3.

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

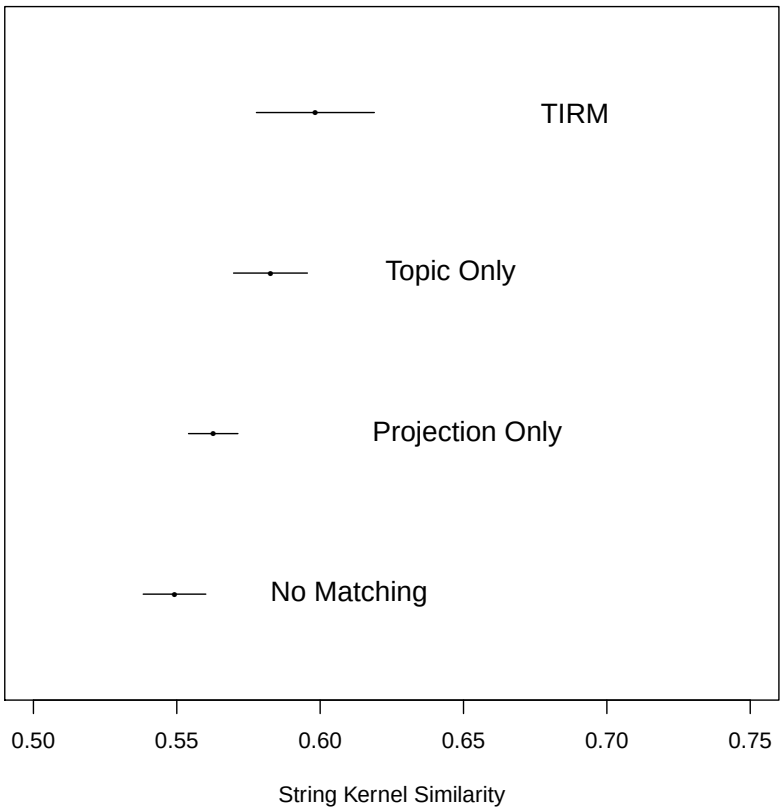

Figure 7: Mean String Kernel Similarity for Matched Posts Randomly Sampled Within Each matched Dataset

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

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

are provided in Table 7.

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

	Mean of Treated	Mean of Control	Difference	P-Value
Previous Censorship Proportion	0.003	0.003	<0.0001	0.223
Previous Not Exist Proportion	0.220	0.220	<0.0001	0.984
Previous Number of Posts	661.987	649.560	12.427	0.773

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

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 Γ, 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 8, we compare

the results across many matches of topical bin size – from 2 bins for each topic to 15 bins for each topic.[16] 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.

[16]We allow CEM to automatically construct the bins.

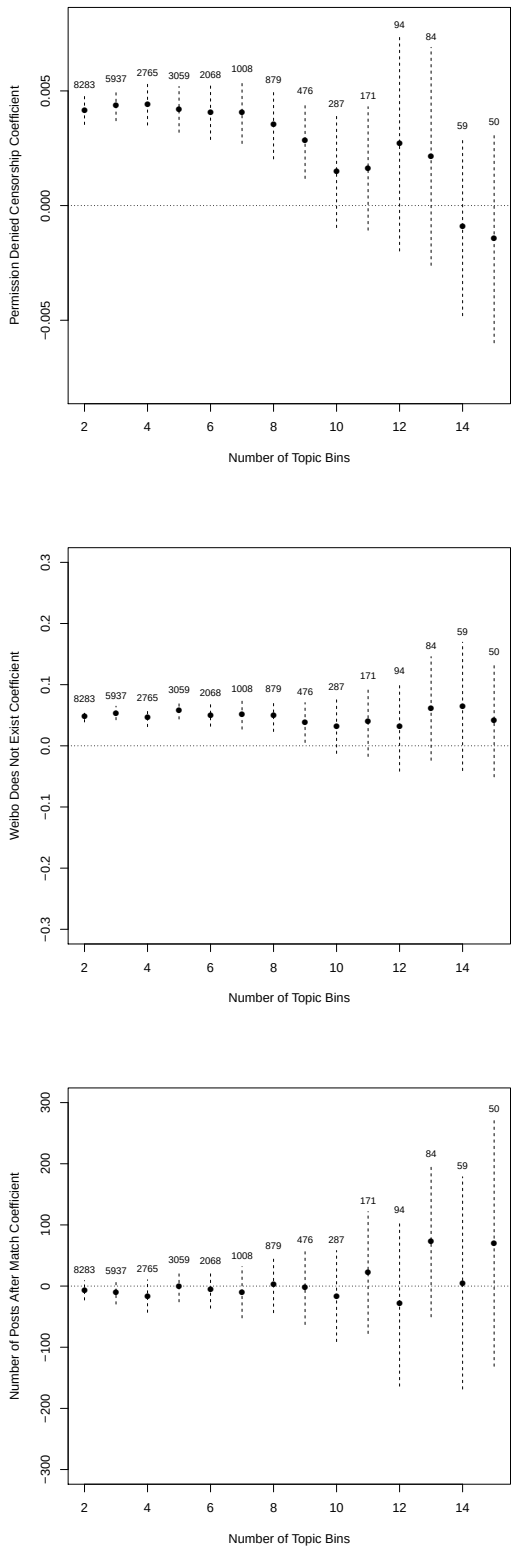

Figure 8: Sensitivity Analysis for Censorship Results, Varying Topical Bin Size. Numbers next to each confidence interval indicate sample size. Top Panel: Effect of Censorship on "Permission Denied" rate. Middle Panel: Effect of Censorship on 'Weibo Does Not Exist" rate. Bottom Panel: Effect of Censorship on Posting Rate.

References

Athey, Susan and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey, Susan, Guido W Imbens and Stefan Wager. 2016. "Approximate residual balancing: De-biased inference of average treatment effects in high dimensions." *arXiv preprint arXiv:1604.07125* .

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81(2):608–650.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon and Bin Yu. 2015. "Lasso adjustments of treatment effect estimates in randomized experiments." *arXiv preprint arXiv:1507.03652* .

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *arXiv preprint arXiv:1701.08687* .

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2017. "How to Make Causal Inferences Using Texts.".
URL: *https://scholar.princeton.edu/sites/default/files/bstewart/files/ais.pdf*

Eisenstein, Jacob, Amr Ahmed and Eric P. Xing. 2011. Sparse Additive Generative Models of Text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11 USA: Omnipress pp. 1041–1048.
URL: *http://dl.acm.org/citation.cfm?id=3104482.3104613*

Fu, King-wa, Chung-hong Chan and Michael Chau. 2013. "Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy." *IEEE Internet Computing* 17(3):42–50.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf and Alexander Smola. 2012. "A kernel two-sample test." *Journal of Machine Learning Research* 13(Mar):723–773.

Gretton, Arthur, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf and Alex J Smola. 2007. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*. pp. 513–520.

Hansen, Ben B. 2008. "The prognostic analogue of the propensity score." *Biometrika* 95(2):481–488.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown and Matt Taddy. 2016. "Counterfactual Prediction with Deep Instrumental Variables Networks." *arXiv preprint arXiv:1612.09596* .

Hartman, Erin and F Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4):1000–1013.

Hazlett, Chad. Forthcoming. "Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects." *Statistica Sinica* .

Hill, Jennifer, Christopher Weiss and Fuhua Zhai. 2011. "Challenges with propensity score strategies in a high-dimensional setting and a potential alternative." *Multivariate Behavioral Research* 46(3):477–513.

Keele, Luke. 2010. "An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data." *White Paper. Columbus, OH* pp. 1–15.
URL: *https://pdfs.semanticscholar.org/c0b1/823186cf5e0869ee35657d6809f803c8e35c.pdf*

King, Gary, Christopher Lucas and Richard A Nielsen. 2017. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference." *American Journal of Political Science* 61(2):473–489.

Maliniak, Daniel, Ryan Powers and Barbara F Walter. 2013. "The gender citation gap in international relations." *International Organization* 67(04):889–922.

Rabinovich, Maxim and David Blei. 2014. The inverse regression topic model. In *Proceedings of The 31st International Conference on Machine Learning*. pp. 199–207.

Ratkovic, Marc and Dustin Tingley. 2017. "Causal Inference through the Method of Direct Estimation." *arXiv preprint arXiv:1703.05849* .

Rosenbaum, Paul R. 2002. *Observational Studies*. Springer.

Rubin, Donald B. and Neal Thomas. 2000. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association* 95(450):573–585.

Sales, Adam C, Ben B Hansen and Brian Rowan. 2015. "Rebar: Reinforcing a Matching Estimator with Predictions from High-Dimensional Covariates." *arXiv preprint arXiv:1505.04697* .

Schneeweiss, Sebastian, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun and M Alan Brookhart. 2009. "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." *Epidemiology (Cambridge, Mass.)* 20(4):512.

Sejdinovic, Dino, Bharath Sriperumbudur, Arthur Gretton and Kenji Fukumizu. 2013. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing." *The Annals of Statistics* pp. 2263–2291.

Sridhar, Dhanya and Lise Getoor. 2019. "Estimating Causal Effects of Tone in Online Debates." *arXiv preprint arXiv:1906.04177* .

Szabó, Zoltán, Arthur Gretton, Barnabás Póczos and Bharath Sriperumbudur. 2015. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*. pp. 948–957.

Taddy, Matt. 2013*a*. "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression." *Technometrics* 55(4):415–425.

Taddy, Matt. 2013*b*. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association* 108(503):755–770.

Taddy, Matt. 2013*c*. "Rejoinder: Efficiency and Structure in MNIR." *Journal of the American Statistical Association* 108(503):772–774.

Tran, Dustin and David M Blei. 2017. "Implicit causal models for genome-wide association studies." *arXiv preprint arXiv:1710.10742* .

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Veitch, Victor, Yixin Wang and David M Blei. 2019. "Using embeddings to correct for unobserved confounding." *arXiv preprint arXiv:1902.04114* .

Wang, Yixin and David M Blei. 2019. "The blessings of multiple causes." *Journal of the American Statistical Association* (just-accepted):1–71.

Westreich, Daniel, Justin Lessler and Michele Jonsson Funk. 2010. "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression." *Journal of clinical epidemiology* 63(8):826–833.

Wood-Doughty, Zach, Ilya Shpitser and Mark Dredze. 2018. "Challenges of Using Text Classifiers for Causal Inference." *arXiv preprint arXiv:1810.00956* .

Yao, Liuyi, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao and Aidong Zhang. 2019. On the estimation of treatment effect with text covariates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press pp. 4106–4113.

Zhang, Qinyi, Sarah Filippi, Arthur Gretton and Dino Sejdinovic. N.d. "Large-scale kernel methods for independence testing." *Statistics and Computing*. Forthcoming.