Matching Methods for High-Dimensional Data with Applications to Text*

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

January 20, 2016

*We thank the following for helpful comments and suggestions on this work: David Blei, James Fowler, Erin Hartman, Seth Hill, Kosuke Imai, Gary King, Adeline Lo, Will Lowe, Chris Lucas, David Mimno, Jennifer Pan, Caroline Tolbert, Walter Mebane and 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. We especially thank Dustin Tingley for numerous insightful conversations on the connections between STM and causal inference. Dan Maliniak, Ryan Powers, and Barbara Walter graciously supplied data and replication code for the gender and citations study. Crimson Hexagon provided data for the study of blogs in China. 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.

†Assistant Professor, Department of Political Science, University of California, San Diego, Social Sciences Building 301, 9500 Gilman Drive, #0521, La Jolla, CA 92093-0521, 360-921-3540, meroberts@ucsd.edu, MargaretRoberts.net

‡Assistant Professor, Department of Sociology, Princeton University, 145 Wallace Hall, Princeton, NJ 08544, 757-636-0956, bms4@princeton.edu, brandonstewart.org

§Assistant Professor, Department of Political Science, Massachusetts Institute for Technology, 77 Massachusetts Avenue, E53 Room 455, Cambridge, MA, 02139, 857-998-8039, rnielsen@mit.edu, http://www.mit.edu/ rnielsen/research.htm

Abstract

Matching is a technique for preprocessing observational data to facilitate causal inference and reduce model dependence by ensuring that treated and control units are balanced along pre-treatment covariates. We identify situations where matching with thousands of covariates is desirable, particularly when pre-treatment covariates are measured from text. However, traditional matching approaches were designed for relatively small numbers of covariates and are ineffective in high dimensions. We propose a conceptually simple solution: estimate and match on a low-dimensional summary of the covariates to improve balance in high dimensions. Under this framework, we develop Topical Inverse Regression Matching (TIRM), a method that balances a low-dimensional projection of text-derived covariates. We illustrate by estimating the effect of censorship on the writing of Chinese bloggers, the effects of perceptions of author gender on citation counts in academia, and the effect of Usama bin Laden's death on the popularity of his writings.

1 Introduction

Matching is a statistical technique for modifying observational data to improve causal inferences and reduce model dependence by ensuring that treated and control units are similar in terms of confounding pre-treatment covariates (Rubin, 2006; Ho et al., 2007; Morgan and Winship, 2014).[1] Vast new data sources are transforming social science (King, 2009; Lazer et al., 2009) and are quickly outpacing existing matching techniques designed for smaller data. Current matching methods, including propensity score matching (Rosenbaum and Rubin, 1983) and coarsened exact matching (Iacus, King and Porro, 2011), were developed for applications with fewer matching variables than observations in the data set. For example, Rubin and Thomas (1996, 249) note that in "typical examples" of matching, the number of covariates is "between 5 and 50," while the number of observations is much larger. Even in settings with very few treated units, the combined number of treated and control units is typically larger than the number of covariates (Abadie, Diamond and Hainmueller, 2010). This paper introduces matching methods for settings with many more covariates than observations.

Matching with many covariates poses three distinct challenges. First, as the number of covariates increases, it becomes difficult to find units that are adequately similar on all dimensions. Exact matching, which requires observations to match on all covariates, often fails to find any matches in high-dimensional settings (Rubin and Thomas, 1996, 250). Related methods that relax the requirement of exact matching (Iacus, King and Porro, 2011) typically produce inconveniently small matched samples or require that variables be dramatically coarsened. Second, when there are many more covariates than observations, standard regression techniques fail to provide useful predictive models of treatment that are essential to methods like propensity score matching (Rosenbaum and Rubin, 1983). Third, assessing balance in high-dimensional data is difficult. In low-dimensional settings, analysts can choose an appropriate balance metric to directly optimize (Hainmueller, 2011; Diamond and Sekhon, 2013; Imai and Ratkovic, 2014) while trading off between balance and sample size (King, Lucas and Nielsen, 2015). However, there are no agreed upon imbalance measures for high-dimensional data and it is not clear how analysts should trade off imbalance along thousands of dimensions when global balance on all dimensions is impossible.

[1] Weighting and regression adjustment are alternatives for conditioning on confounders but we focus on matching because it facilitates straightforward human evaluation of the matched units serving as counterfactuals.

We present a simple approach for addressing the challenges of high-dimensional matching: aim for balance on a low-dimensional projection of the data. We develop this idea with a focus on applications to text data. We extend two existing methods — propensity score matching (PSM) and coarsened exact matching (CEM) — but find these extensions inadequate. The first extension proposes Multinomial Inverse Regression (Taddy, 2013*b*) as an analog to PSM for settings with many more covariates than observations, but we find that the resulting matches are not necessarily qualitatively similar, making it difficult to evaluate whether the matching procedure has gone awry. Similarly, we extend the CEM approach to high-dimensional data in a procedure we call Topically Coarsened Exact Matching (TCEM), but this approach does not sufficiently prioritize variables that affect treatment. We solve these problems with a new algorithm called Topical Inverse Regression Matching (TIRM), which combines the core insights of PSM and CEM in a way that retains their desirable properties while off-setting their weaknesses.

We demonstrate the utility of our approach with three social science applications, two of which draw from other data sources, and one of which we developed solely for this article. First, using data from Roberts (2015), we estimate how censorship affects bloggers in China. It is largely unknown how the experience of being censored affects social media users. Do bloggers avoid sensitive topics after they experience censorship, or does censorship backfire, angering bloggers and inspiring them to write on more sensitive topics? We apply TIRM to identify bloggers who experience censorship and bloggers who write almost identical blog posts but are not censored. We find that bloggers react against censorship; those who were censored write on more sensitive topics than those who were not censored. Our second application estimates whether the gender of journal article authors affects article citation counts in the academic discipline of International Relations. We use TIRM to reanalyze data from Maliniak, Powers and Walter (2013) and find that journal articles by women receive fewer citations on average than very similar articles written by men or mixed groups of men and women. Finally, using data from Nielsen (2015), we estimate the effect of Usama Bin Laden's death on his popularity among readers of a prominent jihadist website. We use TIRM to match Bin Laden's writings to similar texts by other authors and find that Bin Laden's death increased the popularity of his writings for at least three months after his death.

The unifying feature of our three applications is that the natural pre-treatment confounder is represented with text, a form of high-dimensional data which makes the application of existing

methods essentially impossible. Our method for text matching is broadly applicable in all empirical sub-fields of political science and throughout the social sciences. Scholars of American politics could use TIRM to 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 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.

Moreover, our solution can be applied to non-text data in any setting where a generative model of the pre-treatment covariates can reliably estimated.[2] Research in leading political science journals frequently uses matching;[3] as high-dimensional data become increasingly available, innovations such as ours will be crucial for empirical social science.

The paper proceeds as follows. We first introduce some basic ideas and define notation. We then describe new extensions of coarsened exact matching (CEM) and propensity score matching (PSM) that accommodate high-dimensional data. These methods may be useful in their own right but they also motivate the intuition for the next section which introduces our proposed method, Topical Inverse Regression Matching (TIRM) and show its properties using simulation. Finally, we illustrate TIRM in the three applications: estimating the effect of being censored on the reactions of bloggers, the effect of author gender on academic article citations, and the effects of killing Usama Bin Laden on his subsequent popularity.

2 Variable Selection and Coarsening for High-Dimensional Matching

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 bloggers — as a running example. In this example, the goal is to estimate whether Chinese bloggers who have a blog post censored react by writing on more or less politically sensitive topics in subsequent

[2]Generative models related to the ones used here have been applied to computer vision, population genetics, and biological microarrays (Pritchard, Stephens and Donnelly, 2000; Wang and Grimson, 2008; Perina et al., 2010; Bicego et al., 2010). Our approach could extend to these areas, or to applications where the observed data are noisy measures of a latent confounder Kuroki and Pearl (2014).

[3]Since 2010, articles using matching have appeared at a rate of one per year in the *American Political Science Review*, 10 per year in the *Journal of Politics*, and 15 per year in the *American Journal of Political Science*. Matching is also increasing in prominence in sociology and related disciplines (Morgan and Winship, 2014).

blog posts. In broad strokes, our matching strategy is to identify plausible counterfactuals by searching through blog posts to identify pairs of almost-identical posts by different authors where one post was censored and the other was not due to an error by the censors.

The general structure of the problem is as follows. 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 (Imbens and Rubin, 2015), 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 bloggers, the treatment T_i is censorship, and the outcome Y_i is the subsequent reaction of the blogger, measured as the predicted sensitivity of their writings.

In observational settings, T_i is not necessarily randomly assigned so treated and control groups may not be comparable. For example, bloggers who write on sensitive topics may be more likely to experience censorship *and* more likely to blog about sensitive topics in the future. The goal of matching is to condition on confounders in order to approximate random assignment of the treatment. We match on k pre-treatment covariates $\boldsymbol{X} = (X_1, X_2, \ldots X_k)$ to improve balance in those covariates between treatment and control. When the set of $\boldsymbol{X}$ includes all covariates which affect both T_i and Y_i (an assumption called 'selection on observables'), achieving balance implies that the treatment is independent of the potential outcomes in the matched data $(T_i \perp\!\!\!\perp Y_i(0), Y_i(1)|\boldsymbol{X})$, which allows unbiased estimation of the population average treatment effect on the treated, conditional on $\boldsymbol{X}$. Achieving balance on a broader set of covariates which affect only Y can help to reduce the variance of our estimate.

In most matching applications, $\boldsymbol{X}$ is low-dimensional, with $k << n$. However, in the cases we consider here, k is at best large and at worst undefined. In the case of Chinese blog posts, censorship could be assigned on the basis of particular words, particular combinations of words, all words, hierarchies of words (such as titles and section headings), and so on. For simplicity, we assume that every individual word in a blog post may influence censorship, but do not consider word order or combinations. These assumptions can be weakened to include word order, hierarchy, or other features of the documents, by modifying $\boldsymbol{X}$ to include these features. As is common in the text analysis literature (Grimmer and Stewart, 2013), we represent each document (blog post) in a sparse count matrix $\boldsymbol{X}$ whose typical element X_{ij}, contains the number of times the jth word appears within the text associated with unit i. The $\boldsymbol{X}$ matrix has dimension n by $k = V + r$, where V is the number of unique words in the corpus and r are

any other covariates to be matched on in addition to the text. The data are high-dimensional because the number of unique words, V, can be very large relative to the number of units, n.

For credible causal inference, we must balance $\boldsymbol{X}$ such that $T_i \perp\!\!\!\perp Y_i(1), Y_i(1)|\boldsymbol{X}$. In some settings, the k variables that make up $\boldsymbol{X}$ are known to the researcher because the treatment assignment mechanism is transparent. This is not the case for our applications. In the case of Chinese bloggers, we have a rough intuition about which features of a blog post might trigger censorship, but we do not know. We suspect that most of the V words in $\boldsymbol{X}$ are irrelevant to treatment assignment, but we are not sure which ones. We cannot be agnostic about which words in $\boldsymbol{X}$ are relevant. If we consider them all to be equally relevant, than we will not find any matches except in the unlikely event that two blog posts have *identical* word frequencies, a possibility that becomes vanishingly small as V grows large. Some type of dimension reduction is necessary, while maintaining $T_i \perp\!\!\!\perp Y_i(0), Y_i(1)|\boldsymbol{X}$.

In practice, most matching methods involve simplifying $\boldsymbol{X}$ to make matching tractable because exact matches are rare even when $k < n$. Simplifying $\boldsymbol{X}$ typically takes one of two forms: (1) *variable selection*, in which some variables in $\boldsymbol{X}$ are weighted more than others, which is the core element of approaches like PSM; (2) *coarsening*, in which variables in $\boldsymbol{X}$ are replaced by coarsened versions that treat non-equivalent values as equivalent, which is the core insight of CEM. We consider both variable selection and coarsening as possible approaches for simplifying a high-dimensional $\boldsymbol{X}$ and describe why one, absent of the other, leads to poor behavior in high-dimensional matching.

2.1 Variable Selection

There are many approaches to variable selection, but one of the most prominent among matching methods is PSM, which weights the elements of $\boldsymbol{X}$ by their value for predicting T. The PSM procedure summarizes this information in a minimally sufficient statistic for the part of $\boldsymbol{X}$ that is correlated with T, called the *propensity score* (Rubin, 2006, 178, 264, 283). The propensity score is the the probability of treatment conditional on $\boldsymbol{X}$, or:

$$\pi_i = p(T_i = 1|X_i) \tag{1}$$

with $\hat{\pi}_i$ typically estimated via logistic regression of T on $\boldsymbol{X}$. All variables in $\boldsymbol{X}$ are generally included as predictors "unless there is consensus that it is unrelated to the outcome variables

or not a proper covariate" (Rubin, 2006, 269). The estimated probabilities $\hat{\pi}_i$ from this regression, or the linear predictor, are used to match. When successful, propensity score matching approximates conditions of a fully randomized experiment (King and Nielsen, 2015).

When applied to text, the variable selection approach leads us to consider ways to identify words that predict treatment status and match only on those. Directly applying standard implementations of PSM to high-dimensional data is impossible because estimation of the conditional distribution $p(T_i|X_i)$ will not be tractable or efficient unless the number of observations, n, scales well with the dimension of the pre-treatment covariates k.

We turn to inverse regression to estimate propensity scores when $k \gg n$ (Cook and Ni, 2005; Cook, 2007). Faced with the intractable problem of $p(T|\boldsymbol{X})$, inverse regression posits a parametric model for the inverse problem, $p(\boldsymbol{X}|T)$, which produces a sufficient reduction of the information in $\boldsymbol{X}$ about the conditional distribution $p(T|\boldsymbol{X})$.

In our applications, $\boldsymbol{X}$ is a matrix of word counts, so a natural approach is to assume that word counts arise from a multinomial distribution. We therefore rely on the Multinomial Inverse Regression (MNIR) framework developed in Taddy (2013*b*), which leads to the following model for a given document:

$$X_i \sim \text{Multinomial}(\vec{q}_i, m_i) \tag{2}$$

$$q_{i,v} = \frac{\exp(\alpha_v + \psi_v' T_i)}{\sum_{v=1}^{V} \exp(\alpha_v + \psi_v' T_i)} \tag{3}$$

where T_i is an ℓ-length containing a categorical encoding of the treatment variable (in ℓ categories) for document i. The coefficients ψ are often given a sparsity-inducing regularizing prior, a point which we return to below. Mechanically this amounts to estimating a multinomial logistic regression with the words as outcomes and the treatment as the predictor. After estimating the model we can calculate a sufficient reduction score:

$$z_i = \psi'(x_i/m_i) \tag{4}$$

which implies $T_i \amalg x_i, m_i|z_i$, as shown in Propositions 3.1 and 3.2 of Taddy (2013*b*) which establish the classical sufficiency properties of the projection.[4] This implies that given the

[4]Word counts are divided by the number of words in the document. Thus, we optimize balance with respect to length-normalized word counts, allowing documents with very different lengths to be matched. In applications where this is undesirable, analysts can include document length as a matching variable.

generative model in Equation 2 we can condition on z_i and discard the higher dimensional data x_i.

MNIR results in efficiency gains when estimating propensity scores with high-dimensional $\boldsymbol{X}$. This efficiency comes from making fairly strong assumptions about the generative process of the predictor $\boldsymbol{X}$. In addition to the usual assumptions for propensity scores, we also introduce assumptions about the suitability of the generative model.[5] Under the standard propensity score model the variance in the MLE of the coefficients for the propensity score model decreases in the number of documents. However with MNIR the variance decreases with the number of total words (Taddy, 2013*c*, See Proposition 1.1). We defer to Taddy (2013*b*) and Taddy (2013*c*) for a more complete description of the technical properties of MNIR and Cook (2007) for inverse regression more generally.

Multinomial inverse regression requires estimating many coefficients because the coefficient matrix ψ has one column per word in the vocabulary. We accomplish variable selection by estimating the coefficients with a regularizing prior (Taddy, 2013*b*, see details in the Supplemental Information). Due to the use of the sparsity-promoting prior, we are effectively only considering a subset of words which the model estimates have substantially different rates of use in the treated group in comparison to the control group. Following estimation of the MNIR, we can estimate the propensity score using the forward regression $\hat{\pi}_i = \frac{1}{1+\exp(-z_i\hat{\beta})}$ if propensity score itself is of interest. However, the forward regression provides no new information for matching so we skip it and match directly on the sufficient reduction.[6]

The primary strength of this approach is that it simplifies $\boldsymbol{X}$ in an efficient way while maintaining $T_i \perp\!\!\!\perp Y_i(0), Y_i(1)|\boldsymbol{X}$. This means that the well-developed literature on propensity scores can be directly applied here both for matching and alternative approaches such as inverse propensity score weighting. However, we find that this approach has serious a practical weakness when applied to the problem of matching text. Matching on the MNIR-generated propensity score may not result in matched texts that seem similar to human readers because propensity score matching only provides balance in distribution and does not necessarily recover (nearly)

[5] In this model, ψ estimates the population-average effect of the treatment on the word count vector and does not include, for example, the topic-specific types of generative models that we consider next.

[6] An alternative to inverse regression would be to estimate the original "forward" model using regularization. However, under the parametric assumptions of the inverse model, it is substantially more efficient (Taddy, 2013*b*, see). Furthermore, regularizing with the forward regression using a sparsity inducing penalty such as the LASSO (Friedman, Hastie and Tibshirani, 2010) will tend to remove all but one of a group of correlated words, while the inverse regression approach allows them all to enter the projection.

exactly matching pairs (King and Nielsen, 2015). For example, in the application to Chinese bloggers, we find that blog posts about protest and pornography have similarly high probabilities of triggering censorship and are often matched. If the model is correct, using a matched data set with these pairs will result in unbiased causal estimates. However, it dramatically complicates our efforts to validate the procedure in any given case through reading-based assessments of balance.

2.2 Coarsening

Coarsening offers an alternative to variable selection for reducing the dimensions of $\boldsymbol{X}$. Coarsening is often associated with CEM, which applies exact matching to a modified version of $\boldsymbol{X}$ in which variables have been replaced by less-granular summaries of those variables. In most applications, these summaries are created by the analyst, who coarsens each variable into "substantively indistinguishable" bins and then performs exact matching within strata defined by these bins. For example, an analyst matching survey respondents based on years of education might coarsen the many-valued variable *years of education* into substantively meaningful bins: *no high-school degree*, *high-school degree*, *college degree* and *post-graduate degree*. Units that are identical according to these coarsened categories are in the same strata, and thus are matched. When no matches are available for a unit it is dropped from the data, changing the estimand from the Average Treatment Effect on the Treated (ATT) to a treatment effect specific to the treated units which remain.

CEM has the desirable monotonic imbalance bounding (MIB) property, meaning that the researcher bounds the differences between treated and control to the extrema of the strata (Iacus, King and Porro, 2011). Thus CEM approximates a blocked randomized experimental design, which is more efficient than the fully randomized design approximated by PSM (Imai, King and Stuart, 2008; King and Nielsen, 2015). In PSM, matches can be very far apart on particular covariates but still be matched because they have similar probabilities of treatment. In CEM, the multivariate difference between matched units is bounded.

CEM is generally only feasible when the set of matching variables is small relative to the number of observations. To see why, consider an application with a single matching variable coarsened into four categories. This results in four strata that should ideally be populated with treated and control units. If a second matching variable is added and also coarsened into four categories, these results in $4 \times 4 = 16$ strata that should be populated. Now consider a

very small text corpus with only 100 unique words. Even if we apply the broadest possible coarsening — replacing the term frequencies with a binary variable indicating whether or not the document contains the word — this results in 2^{100} strata. Unless combinations of words are almost perfectly correlated, the number of strata is so incredibly large that there will generally be almost no matches. If a document contains even a single unique word, then no coarsened exact matches are possible.

To develop a coarsening approach for dimension reduction in high-dimensional matching problems, we extend the logic of CEM. Rather than grouping units with similar values on individual variables into the same strata, we group the variables themselves. To illustrate, we might assume that when the words "censor", "censoring", and "censored," occur in a set of blog posts, they all have approximately the same referent and can be grouped together into one concept ("censor") and represented by a single indicator (which now takes on the value of 1 if any of the words "censor", "censoring", and "censored" are present and is 0 otherwise). This procedure is called stemming and is already widely used for dimension reduction in text analysis. Crucially, CEM retains the MIB property with stemmed text data because semantic distance between texts remains bounded: any matched documents must have equal counts of the stem "censor," although counts of the unstemmed words "censor", "censoring", and "censored" may not be equal.

However, traditional stemming does not normally reduce dimensionality enough to facilitate matching with high-dimensional text data.[7] Instead, we propose a procedure called *Topically Coarsened Exact Matching* (TCEM) in which we estimate a topic model (Blei, Ng and Jordan, 2003; Roberts et al., 2014; Roberts, Stewart and Airoldi, Forthcoming) and then apply coarsened exact matching to the estimated topics. Under the topic model, each word in the corpus has a latent topic assignment and two words with the same topic assignment are stochastically equivalent. TCEM results in matched documents that have comparable amounts of stochastically equivalent words, although the specific observed words may differ. In our running example, we could use TCEM to find documents that have the same amount of a "censorship" topic, while not conditioning directly on precisely which censorship words are used.

TCEM maintains the monotonic imbalance bounding property in the topical space, a fact which we interpret in two ways. First, it may be the case that treatment assignment is, in fact,

[7]Stemming is also underdeveloped for languages like Chinese (where verbs are not conjugated) or Arabic (where words are modified by infixing) (see Lucas et al., 2015).

based on the topics of texts, rather than on the exact wording. If so, then the MIB property of TCEM is ideal. Alternatively, we may believe that treatment assignment is based on features that are not perfectly summarized by latent topics. Even so, matching on the density estimate of the topic proportions is a way of reducing variance at the risk of introducing a small amount of bias. From this perspective, TCEM is appropriate if two observations with a common density estimate are stochastically equivalent and deviations between them are essentially random noise.

Topic models require that the analyst choose the number of topics. This choice can be fraught for cases where semantic interpretation of topics is the primary concern (Grimmer and Stewart, 2013), but it matters less for matching. The number of topics should be sufficient for matched documents to be good counterfactuals for each other, as determined by the demands of the research design. In general, more topics will result in closer matches. Redundant topics will decrease efficiency but will not cause bias, so the risks of choosing too few topics are much greater than choosing too many.

TCEM ensures that matched texts will be topically similar, a feature which facilitates balance-checking via manual comparison of matched documents. However, TCEM is weak precisely where the MNIR approach explored above is strong: the topics are estimated without information about treatment status so TCEM can fail to detect sets of words that predict treatment assignment. This happens because topic models must assign all words in the corpus to a topic, which generally means that topics will capture the subject matter of the document rather than, say, the sentiment, even though sentiment may be a strong predictor of treatment assignment. Thus, TCEM could be greatly improved if information about treatment assignment could be included.

3 Topical Inverse Regression Matching

In this section, we propose a matching algorithm called topical inverse regression matching (TIRM) that combines the variable selection properties of the MNIR approach with the coarsening properties of TCEM for high-dimensional matching applications. Conceptually, TIRM combines MNIR with TCEM to inherit the properties of both. For estimation we leverage a newly developed topic model, the Structural Topic Model (STM) (Roberts, Stewart and Airoldi, Forthcoming), which facilitates joint estimation of both topics (for coarsening) and document-level propensity scores (for variable selection). We find that matching on both of these estimated quantities achieves the benefits of both.

The Structural Topic Model extends the popular latent Dirichlet allocation model (Blei, Ng and Jordan, 2003; Blei, 2012) for use with covariates that model *topic prevalence* – the frequency with which a topic is discussed – and *topical content* – the words used to discuss a topic (Roberts et al., 2014; Roberts, Stewart and Airoldi, Forthcoming). We show that if analysts use T_i as the *topical content covariate* in an STM model, this injects information about treatment assignment into a topic model which effectively combines the MNIR and TCEM frameworks developed above.

The data generating process of STM can be given as:

$$\vec{\gamma}_k \sim \text{Normal}_P(0, \sigma_k^2 I_P), \qquad \text{for } k = 1 \dots K-1, \tag{5}$$

$$\vec{\theta}_d \sim \text{LogisticNormal}_{K-1}(\mathbf{\Gamma}'\vec{x}_d', \mathbf{\Sigma}), \tag{6}$$

$$\vec{z}_{d,n} \sim \text{Multinomial}_K(\vec{\theta}_d), \qquad \text{for } n = 1 \dots N_d, \tag{7}$$

$$\vec{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B}\,\vec{z}_{d,n}), \qquad \text{for } n = 1 \dots N_d, \tag{8}$$

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa^{(\text{t})}_{k,v} + \kappa^{(\text{c})}_{y_d,v} + \kappa^{(\text{i})}_{y_d,k,v})}{\sum_v \exp(m_v + \kappa^{(\text{t})}_{k,v} + \kappa^{(\text{c})}_{y_d,v} + \kappa^{(\text{i})}_{y_d,k,v})}, \qquad \text{for } v = 1 \dots V \text{ and } k = 1 \dots K, \tag{9}$$

Note that Equation 9 mirrors the MNIR model in Equation 2, but includes topic-specific effects and (optionally) topic-covariate interactions. This is because STM can be interpreted as embedding a multinomial inverse regression into a topic model, or embedding topic-specific random effects inside the MNIR model.[8]

In the MNIR framework above, we developed an analog to the propensity score by estimating a projection from the model that was a sufficient reduction of the information in the word counts about the probability of treatment. This is also possible in the STM model. Following Taddy (2013*b*), we derive a sufficient reduction of the information contained in the word counts about treatment, but because we include topics, the projection now represents the information about the treatment *not* carried in the topics. If we were to omit topic-covariate interactions, the sufficient reduction would take the simple form $(\kappa^{(c)})'(x_i/m_i)$.[9] However, topic-covariate interactions are desirable for injecting sufficient information about T into the estimated topics. With interactions present in the content covariate, the projection becomes $(\kappa^{(c)})'(x_i/m_i)$ +

[8] Taddy (2013*b*) uses the Gamma-Laplace scale mixture prior for κ while we use the more basic Laplace prior. This does not matter in practice.

[9] This projection was explored in work by Rabinovich and Blei (2014) where it is used as a sufficient reduction to improve prediction.

$\frac{1}{m_i} \sum_v x_{i,v} \left(\left(\kappa_v^{(\text{int})} \right)' \theta_i \right)$. This projection is analogous to the unnormalized propensity score.

The other estimate we extract from the STM model is the topical content of the documents. A slight complication arises because when topic-treatment interactions are present: the same topic can have different estimated distributions of words under treatment and control. We ensure that topics are comparable irrespective of treatment/control differences by adding a final estimation step to the STM in which we re-estimate the topic proportions of all control as though they were treated. This choice is consistent with an estimand that is a (local) average treatment effect on the *treated*. We now 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-topic probabilities of treatment. We recommend using CEM for this matching step if pruning treated units is acceptable. In Section 4.3 we give an example where dropping treated units is not acceptable and consequently we use fixed-ratio Euclidean distance matching.

3.1 Balance Checking

Balance checking — confirming that matched units are in fact similar — is important for assessing whether matching is successful, but can be difficult with high-dimensional confounders measured from text. This is because there is no agreed upon metric for assessing text similarity, and there is not likely to be a universally best metric because the aspects of text that confound causal inference differ from application to application. 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. Next, we suggest checking similarity of match texts on a metric that is *not* directly minimized by TIRM: string kernels, which measure similarities in sequences of characters (Lodhi et al., 2002; 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 omitting word order is not overly consequential. Finally, we recommend close reading of matched documents. Reading is subjective, but it is often most revealing about whether matching has succeeded or failed to identify adequately similar texts.

3.2 Strengths and Limitations

The core strength of TIRM is that it incorporates both variable selection and coarsening within a single framework. TIRM estimates both document topic proportions and within-topic propensities for treatment and as a result increased weight is given to words that predict treatment (as in MNIR) while the resulting matches are topically similar (as in TCEM). This allows TIRM to prioritize variables which are related to treatment assignment while approximating a blocked design on the full set of confounders.

The primary limitation of TIRM is that matches rely on a parametric model of a complex data generating process. Thus the matches will only be useful if the topic model is an adequately accurate summary of the documents. Analysts can 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.

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 (Bowers, Fredrickson and Panagopoulos, 2013; Aronow and Samii, 2013). 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 on a case-by-case basis. 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 also changing. Particularly when dropping treated units it is necessary to carefully characterize the group to which the estimated effect applies (King, Lucas and Nielsen, 2015; Rubin, 2006, 221-230).

3.3 Simulation

Before proceeding to applications, we show that TIRM can recover causal effects in simulated data. It is challenging to design a simulation which accurately reflects the complexities of real world texts and so we use observed data as much as possible. We summarize the basic strategy here, leaving details to the Supplemental Information B. Starting with real text data,

we generate a simulated treatment variable as a function of unobserved topics and a small, unknown selection of individual words. A simulated outcome variable is generated from a Poisson regression model including the unobserved topics, the words that affected treatment, and a constant treatment effect that we will attempt to recover. This strategy generates imbalance between treatment and control documents which TIRM seeks to minimize. Importantly as with the real applications, the variable causing imbalance are either not directly observed (topics) or are mixed in with other unimportant variables (individual words). The challenge for TIRM is to correctly recover the topics, identify the small subset of important words, and balance these adequately to recover the simulated treatment effect.

Figure 1 summarizes the estimated treatment effect over 200 simulations. The left panel uses a Poisson regression to condition on variables extracted from TIRM – the estimated topics and the topic-specific probabilities of treatment. The figure shows three different conditioning sets: nothing (the naïve estimator), topics only and topics with projection (capturing the individual words). As expected we see that the naïve estimator is badly biased, the topics only estimator performs better and the topics with projection estimator performs the best. The right panel shows the result using coarsened exact matching on the topics and projections for two different sized coarsening bins. After matching the estimand is estimated using a Poisson regression on the matched data, controlling for remaining imbalance using the estimated covariate. Both levels of coarsening correctly recover the estimate of interest.

3.4 Related Work

Before moving to applications of TIRM to real data, we briefly mention how it relates with other approaches to similar problems. There has been very little work on general matching strategies for high-dimensional data.[10] The majority of the related work involves methods for estimating propensity scores in high dimensions, which falls squarely in the PSM framework we have found to be inadequate on its own (Schneeweiss et al., 2009; Westreich, Lessler and Funk, 2010; Hill, Weiss and Zhai, 2011; Belloni, Chernozhukov and Hansen, 2014). We are not aware of any work which matches on a density estimate although loosely related approaches have appeared in

[10]There has been excellent work in weighting strategies which are applicable to high-dimensional data (Hainmueller, 2011; Hazlett, 2015).

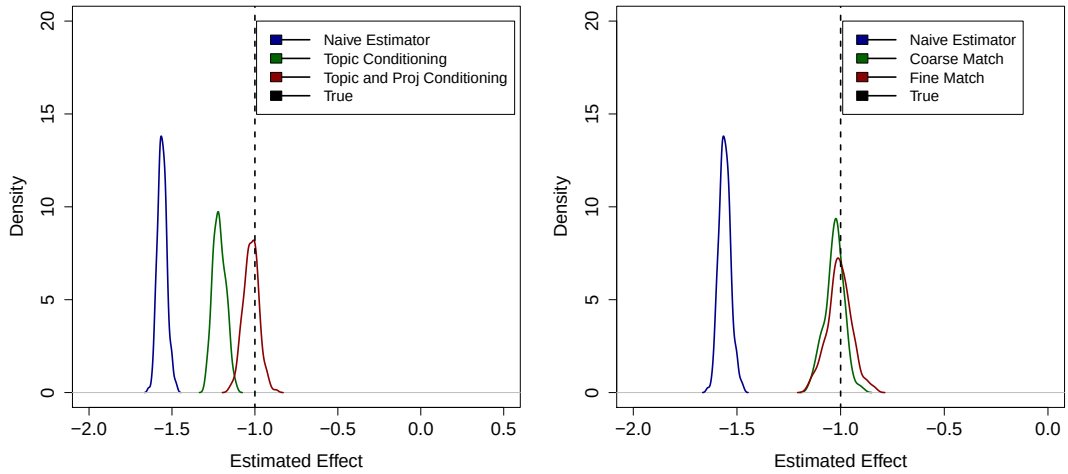


Figure 1: Simulation Results. 200 simulations. Left Panel: Poisson regression estimate of treatment effect, without conditioning, conditioning only on topics, and conditioning on topics and projection. Right panel: Coarsened exact matching on topics and projection compared to the naive estimator

statistical genetics[11] and optimal design for manual coding.[12] Both of these interpretations rely on the intuition that if two words commonly co-occur, then they are essentially interchangeable for the purposes of identifying appropriate counterfactual documents. Finally the problem of causal inference with high-dimensional data has started to be explored using the framework of regression adjustment for both experimental (Bloniarz et al., 2015) and observational (Sales, Hansen and Rowan, 2015) data. Both of these approaches leverage models of the outcome to reduce variance in the estimate (see also, Rubin and Thomas, 2000; Hansen, 2008). By contrast our approach is focused on the analysis of observational data and falls in the matching framework which does not rely on a separate regression model of the outcome data.

4 Applications

Having shown that TIRM can recover effects in simulated data we turn to three applications to observational studies. Each application is both substantively important in its own right

[11]Price et al. (2006) suggests analyzing genotype data with an eigen decomposition followed by a regression-based adjustment. Although the procedure is regression-based, the paper invokes the idea that this creates "a virtual set of matched cases and controls" (904). A topic model could be considered a form of model-based discrete principal components analysis (Buntine and Jakulin, 2004).

[12]Taddy (2013*a*) considers 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.

and affords us the opportunity to show a unique facet of our method. For comparison, we also demonstrate the performance of our newly developed analogs for propensity scores and CEM through multinomial inverse regression (MNIR) and topically coarsened exact matching (TCEM) respectively.

In Section 4.1 we estimate the effect of government censorship on subsequent writing by Chinese bloggers. In this application, we find that TIRM and TCEM recover virtually identical matches, while MNIR performs poorly. Section 4.2 builds on previous work by Maliniak, Powers and Walter (2013) to estimate the effect of author gender on the citation counts of academic journal articles in the discipline of International Relations. In addition to continuing to unpack the differences in our proposed methods, this application allows us to compare matching based on TIRM to the vastly more expensive human coding of journal articles used in the original project. Finally, we estimate the effect of Usama Bin Laden's death on the popularity of his writings among users of a prominent jihadist website. This example demonstrates how TIRM can be modified to accommodate different quantities of interest, and to include additional non-text covariates. Space constraints prevent us from describing some of the particulars of each application, so we refer readers seeking more detail to Roberts (2015) and Nielsen (2015), where these applications are given article-length treatment.

4.1 Government Censorship Emboldens Chinese Bloggers

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. By Chinese law, it is illegal to write about anything that "harms the interest of the nation", "spreads rumors or disturbs social order", "insults or defames third parties", or "jeopardizes the nation's unity."[13] Bloggers who violate any of these ambiguous laws may have their post censored, lose their web account, or be jailed.

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 how having a post censored changes the online behavior of bloggers in China. This issue is particularly important because the number of bloggers vastly outnumbers the enforcers leading to the inconsistent application of

[13] `http://china.org.cn/government/whitepaper/2010-06/08/content_20207978.htm`

laws. This has led scholars of Chinese politics to debate the degree to which the government relies on self-censorship, effectively using threats of punishment to deter bloggers from writing about sensitive topics (Wacker, 2003; Kalathil and Boas, 2010).

The quantity of interest we are interested in estimating is the effect of experiencing censorship on bloggers who conceivably would be censored. An ideal experiment to estimate this effect would involve taking the set of active bloggers who are candidates to be censored on a given day and then randomly assigning a subset to actually be censored. We could then observe the writing of these blogs into the future confident that our random assignment addressed any confounding. This experimental ideal is both unethical and impossible for us to implement, but we approximate the experiment by identifying censors' mistakes. Chinese censorship is quite thorough, but occasionally misses sensitive blog posts. We exploit these mistakes by using TIRM to identify pairs of nearly identical blog posts where one is censored and the other is not. We can then observe the subsequent blogging activity of both bloggers to estimate the causal effect of censorship on the treated units who remain in our sample. We note that this gives us information about self-censorship for active bloggers writing on sensitive topics during our time-frame, but cannot speak to self-censorship effects that may cause a blogger to either never write about sensitive topics or simply not blog at all.

We use a dataset of 593 bloggers observed over six months, with posts collected by Roberts (2015) in real time and then revisited later to observe whether they were censored. Our goal is to identify posts that have both nearly identical text and nearly identical ex-ante probability of being censored. We compare the three procedures we have introduced: MNIR, TCEM, and their combination TIRM. After processing the text to account for linguistic features of Chinese, we convert the text of blog posts into a matrix of document word counts and estimate each of these models, as described above. We specify 200 topics for the topic models. We match closely on these 200 topics because we hope to retrieve close to identical pairs of blogposts.

We find that TIRM and TCEM are both effective at identifying essentially identical pairs of blog posts with different censorship statuses. In contrast, MNIR does not identify these pairs. We suspect that this is because blog posts on very different topics can face similarly high probability of censorship. To see the quality of matches from each method, consider Figure 2, in which we compare the average difference in the 10 topics estimated to be most and least associated with censorship in the unmatched dataset. We find that while TIRM and TCEM

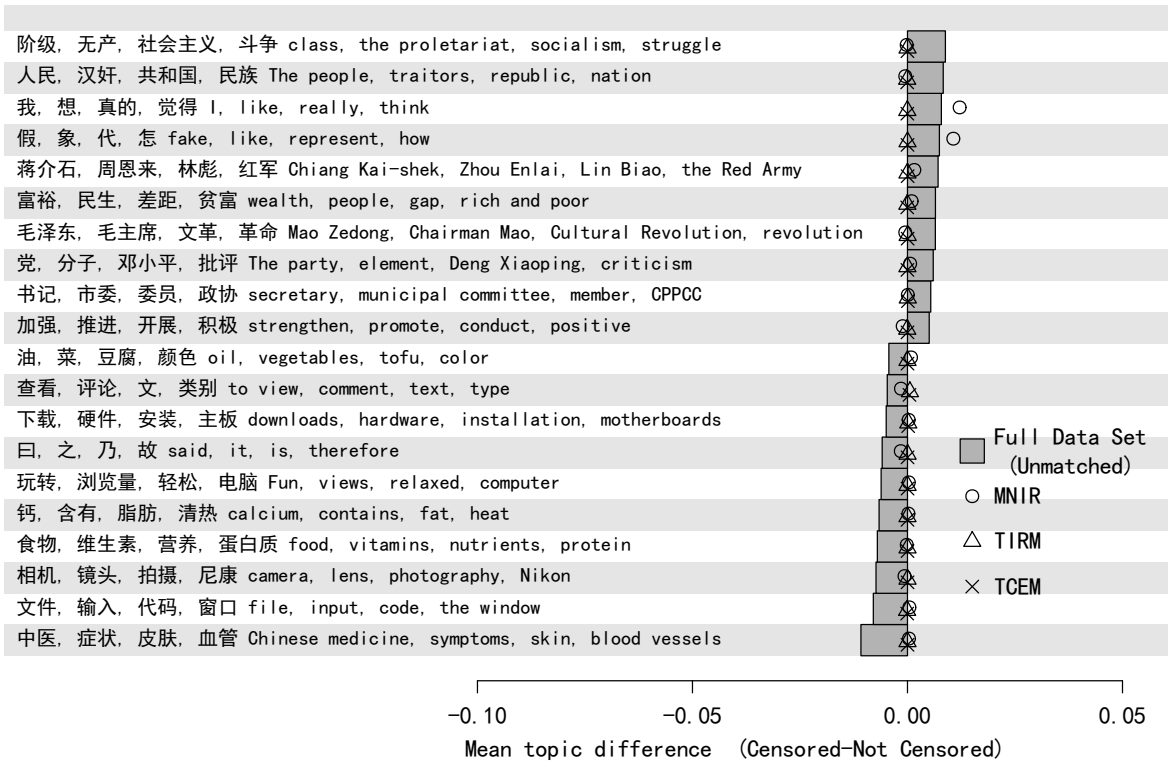


Figure 2: Topic balance comparison between unmatched, MNIR matched, TCEM matched, and TIRM matched.

reduces topical differences to essentially zero in every case, MNIR makes balance worse for a few topics. Next, we compare the string kernel similarity of documents in the matched data sets produced by each method and the unmatched data set and again find that TIRM and TCEM outperforms other methods by this metric (details in the Supplemental Information C).

In the text matching case we greatly prefer that matched units have not only a similar propensity to be censored but are also similar in their content. Matching on the MNIR propensity scores allows for balance to be achieved in expectation but it is difficult for humans to evaluate the balance by examining the matches. Thus there is a significant risk that the model could fail without our knowledge. By contrast we read through all the matched documents produced by TIRM and were able to quickly establish that in the matched posts 40% contained the exact same text, 60% were mostly identical with only minor differences, and only one pair where the differences changed the meaning of the post.

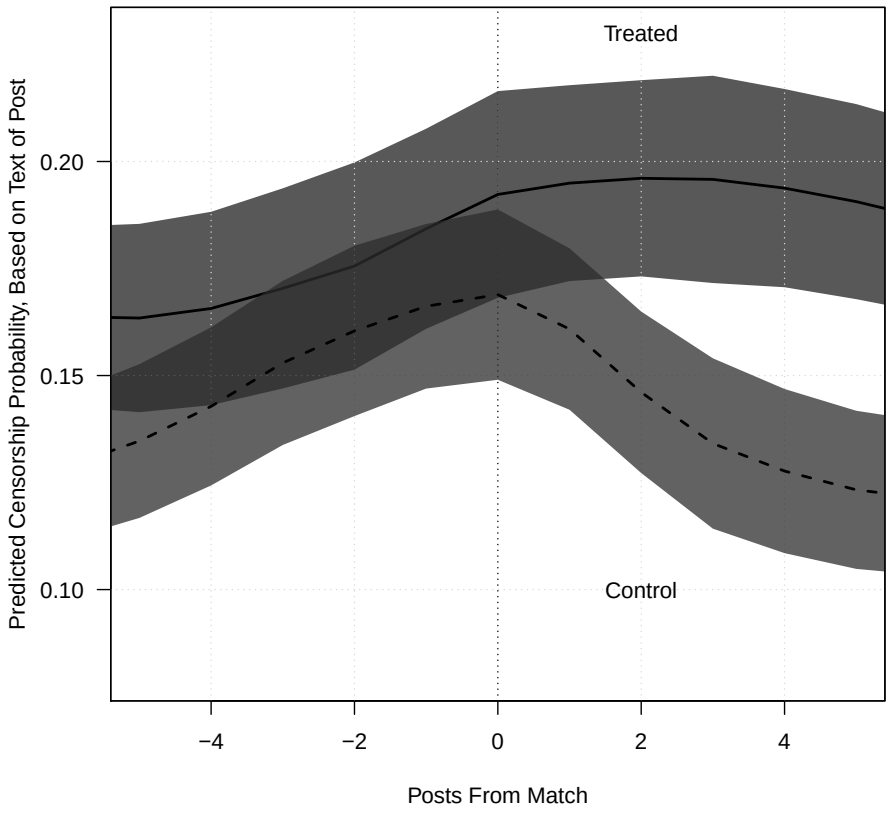


Figure 3: The Effect of Censorship on the Sensitivity of Writing

After determining that the propensity score and topic proportions estimated by TIRM produce the best text matches, we add additional covariates — previous censorship experience, previous sensitivity of writing as estimated by TIRM, date, and previous post rate — and match using CEM to yield a final matched sample of 46 posts. To measure the sensitivity of bloggers' writing after the matched posts, we use the TIRM model to calculate the probability of censorship for each subsequent post which we refer to as "sensitivity". Figure 3 plots the sensitivity of blog posts before and after the post of one blogger is censored. Prior to censoring, the blog posts of soon-to-be censored and uncensored bloggers have statistically similar sensitivity. After censoring occurs, however, subsequent blog posts by censored bloggers are more sensitive, meaning that censorship seems to inspire bloggers to write more about topics deemed off-limits.

4.2 Female IR Scholars are Cited Less Often

Our next application re-evaluates the evidence of Maliniak, Powers and Walter (2013) showing gendered citation bias in the discipline of International Relations (IR). If an 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?[14] Maliniak, Powers and Walter (2013) say yes. 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 using regression to account for article age, journal, academic rank of author, the broad issue area of the article, and the methodology used in the article. Based on citation network analysis, they attribute bias to two factors: men tend to cite other men, and men tend to self-cite more than women.

Evidence of citation bias against female scholars in IR would be strongest if women wrote identical articles to men with identical scholarly credentials but were then cited less often. This is not the case. Men and women tend to write on different topics within IR, use different methods, and have different epistemological commitments. Because these factors might affect citation counts, it possible the lower citation counts of women reflect bias against certain topics and approaches, rather than against women themselves. Maliniak, Powers and Walter (2013) address this challenge using information from the TRIP Journal Article Database (Peterson and Tierney, 2009) to control for the broad sub-field of each article, the issue areas covered,

[14] 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.

the general methodology, paradigm,[15] and epistemology. However, these broad measures of article characteristics may be insufficient. Women could conceivably write in the same subfield and issue areas as men, using the same broad methods, paradigms, and epistemology, and still produce articles that are textually different in ways that result in fewer citations. We use TIRM to address this problem by matching female-authored articles to textually similar male-authored or male and female co-authored articles. This allows us to demonstrate that: (1) TIRM correlates with the manual content coding used by Maliniak, Powers and Walter (2013), and (2) matching via TIRM produces a matched sample of even more similar articles than the set analyzed by Maliniak, Powers and Walter (2013).

With the help of JSTOR's Data For Research Program, we obtain the full text of 3,201 articles in the IR literature since 1980, 333 of which are authored solely by women.[16] We apply TIRM to the data, specifying 15 topics in the STM portion of the algorithm to recover broad topics, roughly on the same scale as the issue area designations coded by hand. We note that although the text is typically written after the author's actual gender is assigned, the text is written prior to treatment as we define it: the moment at which a which a reader perceives the author's gender, typically from reading the byline. We also create matched samples using MNIR and TCEM, and conduct exact matching based on the human-coded data to see how TIRM compares.

We find that 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 are more likely to use words associated with statistical methods ("model", "estimate", and "data"). We test whether TIRM has adequately reduced these differences in the matched sample by identifying the words that have high *mutual information* with author gender in the raw data set — those words that contain the most statistical information about gender — and calculating the difference in frequency by gender for each word in the TIRM matched data. We evaluate TCEM, MNIR, and exact matching on the qualitatively coded issue-area information for comparison.

The top-left panel of Figure 4 shows the relationship between the difference in word occur-

[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.

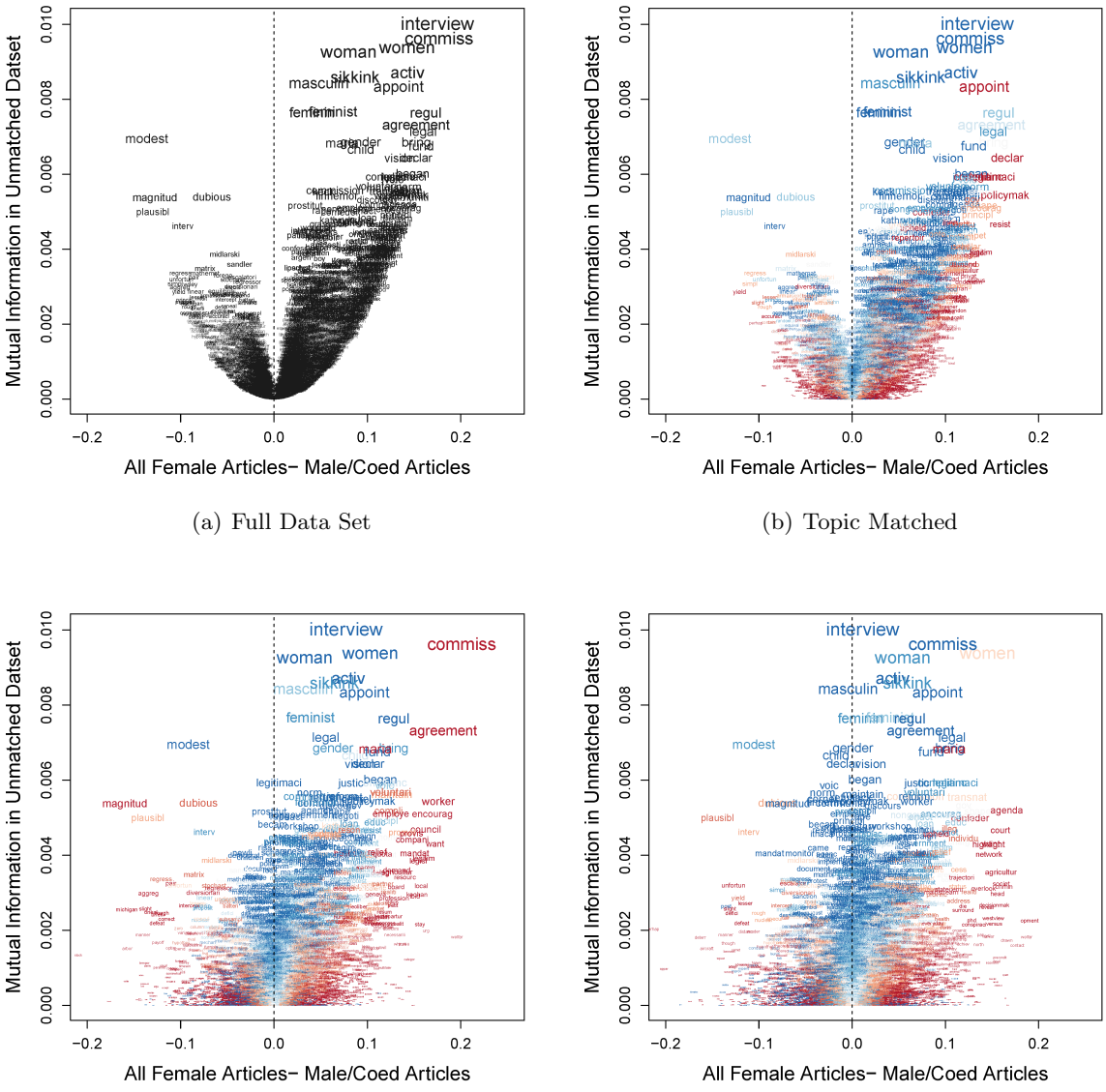
rence by gender and the mutual information of each word in the raw data set.[17] The distinctive fan shape occurs because of the mathematical relationship between these two quantities. In the other three panels, we re-plot the words according to the same mutual information calculation as before (so the same word always appears at the same height in all four panels), but after re-calculating word use differences by gender using one of the matched samples (the x-axis position of words changes). To make the figure interpretable, we color words darker blue if the absolute difference in use of a word is reduced, and darker red if it increases. 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. For example, the single word with the most mutual information about gender in the raw data is "interview," presumably because women use interviews more than men. TIRM balances usage of this word almost exactly. More generally, many words 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 reduce bias. Exact matching on human coding (bottom-left) performs similarly to TIRM, but is not quite as successful at reducing the imbalance of high mutual information words to zero. TCEM (top-right) and MNIR (not shown) are less effective at balancing these words.

We also compare balance on the 15 estimated topics and find that TCEM performs the best on this metric, followed closely by TIRM. Exact matching on human coding does reasonably well, with the exception of a single topic — foreign policy — where it makes balance worse. MNIR *increases* imbalance for most topics. We also check string kernel similarity of the matched data sets and find that TIRM and human coding perform equally well and both offer improvements over the raw data. More details are in the Supplemental Information D.

We can 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 because the coding system was not designed to facilitate our precise causal inference, but it is worth considering

[17] Difference in word occurrence, the percentage of all female documents in which a word appears minus the percentage of all male/co-ed documents in which the word appears, is not what is being balanced by TIRM, but provides good alternative balance metric between treated and control groups.



(a) Full Data Set

(b) Topic Matched

(c) Matched on human codes

(d) TIRM

Figure 4: Relationship between mutual information and difference in word occurrence (all female - male/coed) a) Full Data Set b) Topic Matched c) Matched on human codes d) TIRM. In panels b, c, and d, words for which matching decreased the absolute value of the difference in word appearance are in blue and words for which matching increased the absolute value of the difference in word appearance are in red.

whether TIRM balances categories derived from extensive reading.

Figure 5 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. TCEM most reduces imbalance in the human-coded categories, suggesting that the topic model in TCEM 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, MNIR makes balance worse on many human-coded categories, meaning that the MNIR matches are quite different than the matches a human would select.

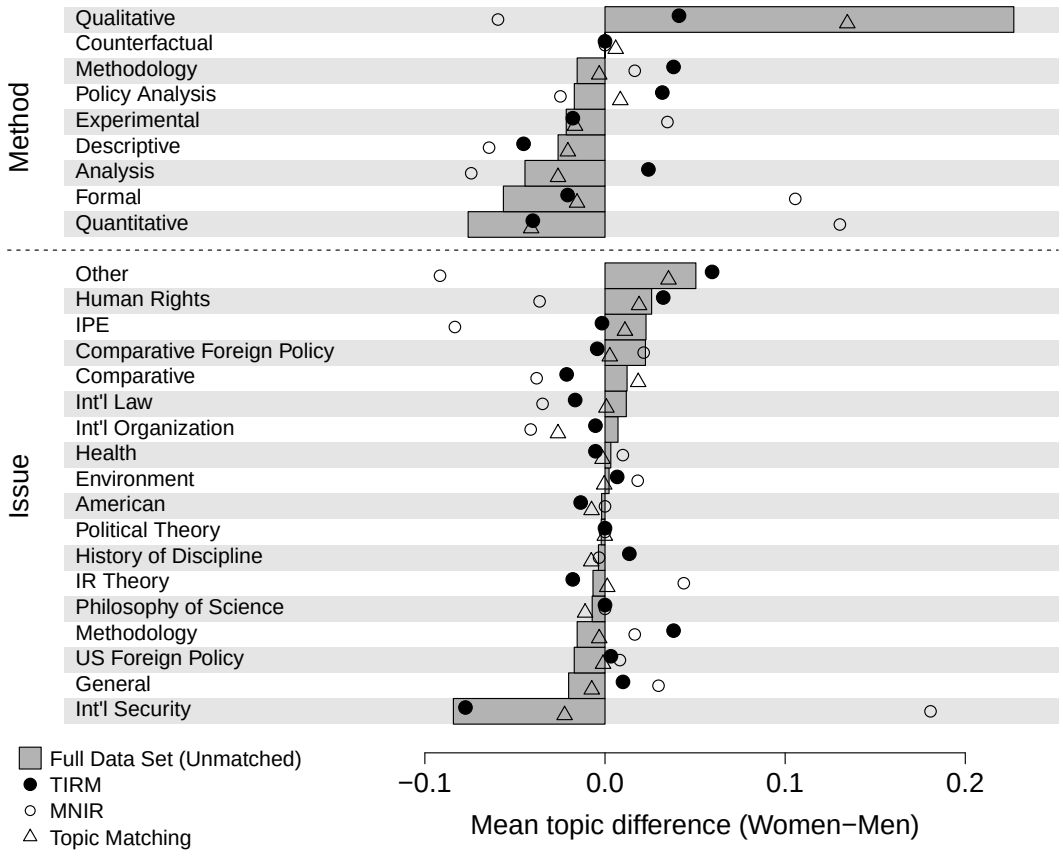


Figure 5: Automated Matching Comparison and Human Categories

We re-estimate the models of Maliniak, Powers and Walter (2013) using the TIRM matched sample and find that gender differences in citations are even more pronounced than those orig-

inally 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, suggesting that paradoxically, it is the most frequently cited female scholars who nevertheless are less cited than if they were men. Because this example is focused on the method, rather than the finding, we refer readers to the Supplemental Information D for further results.

4.3 The Effect of Bin Laden's Death on the Popularity of His Writings

Our final application estimates whether the targeted killing of Usama Bin Laden in May 2011 made his writings more popular with other jihadists. Sunni Muslim Jihadists have been targeted in a variety of ways during the "War on Terror," but there is little systematic evidence about the impact of this targeting on the influence of the leader. Scholars disagree about whether targeting the leaders of terrorist organizations helps or hurts counterterrorism efforts (Jordan, 2009; Johnston, 2012), and the death of Bin Laden raised concerns that his killing would lionize him and make his ideas more popular (Cronin, 2006, 40).

We use TIRM to estimate whether Bin Laden's death re-popularized his writings. Nielsen (2015) has collected data from a large Jihadist online library containing 6,115 documents authored by 492 individuals, most of whom are Arabic-speaking jihadists writing on some aspect of jihadist ideology. The web library counts page views accumulated by each document in real time, which Nielsen (2015) collected between February 2011 and September 2014. We estimate whether the death of Bin Laden made his ideas more or less popular in the short- to medium-term by comparing the post-targeting popularity of Bin Laden's 33 documents to the popularity of texts by non-targeted authors.

If targeting were random, we might obtain unbiased estimates of the effect of Bin Laden's death by comparing the popularity of his writings to all other documents on the website. However, Bin Laden was not targeted randomly and the reasons for his targeting might also affect the popularity of his writing. We condition on two types of data to develop a credible counterfactual claim about what might have happened to Bin Laden's popularity had he not been targeted: (1) the content of documents, and (2) the prior page view trends. Our claim is that if we can identify documents that have similar themes and similar pre-targeting viewing patterns as Bin Ladin's 33 writings, these documents can serve as a comparison group.

We summarize the text using TIRM with 50 topics. After examining the topic model to

ensure that the topics are generally meaningful, we identify control documents that have similar topic proportions Bin Laden's texts. To match, we deviate from our standard TIRM algorithm by replacing the final matching step on the topic proportions and propensity score (where we typically use CEM) with a 1-to-25 nearest-neighbor algorithm that will not allow any of Bin Laden's documents to be discarded.[18] This prevents any treated units from being discarded, meaning that our estimand remains the sample average treatment effect on the treated (SATT) rather than the *feasible* sample average treatment effect on the treated (FSATT) (King, Lucas and Nielsen, 2015). We focus on this quantity of interest because we wish to estimate the effect of targeting on the popularity of Bin Laden's writings *as a whole*. We calculate nearest neighbors using a weighted combination of the STM results (topic proportions and the respective projections of the MNIR model) and the Euclidean distances in the trend of pre-treatment page views.

To evaluate the quality of the matching, we use Figure 6 to compare the topics in Bin Laden's writings to the topics in the unmatched and matched control samples. In the unmatched sample, we see that Bin Laden devotes more writing to topics of fighting, theaters of operations, and geopolitics. Matching on topics from TIRM reduces imbalance on the topics that are most associated with Bin Laden's writing, but this balance comes at the expense of balance on a few less predictive topics. This is acceptable because the topics to do not predict treatment cannot be confounders.

Manually validating the quality of matches in this corpus by comparing matched pairs of documents to randomly paired documents is infeasible because the documents are long and the material is technical (average word-count is 5,403, maximum word-count is 445,810). However, the structure of the jihadist web library provides a unique way to validate our matching procedure; the website is organized in a series of sub-pages and documents on the same sub-page are placed there because website administrators deem them similar in some way. We find that on average, documents matched by our method are 8 times more likely to co-occur on a sub-page than randomly selected pairings, meaning that our matches correspond to website administrators' judgments about document similarity.

The left panel of Figure 7 illustrates the results of the matching procedure for "Letter to

[18] We match 25 control texts to every treated unit, rather than more conventional 1-to-1 matching, because we need the additional matched controls for statistical precision. The results are similar, but less precise with 1-to-1 and 1-to-5 matching.

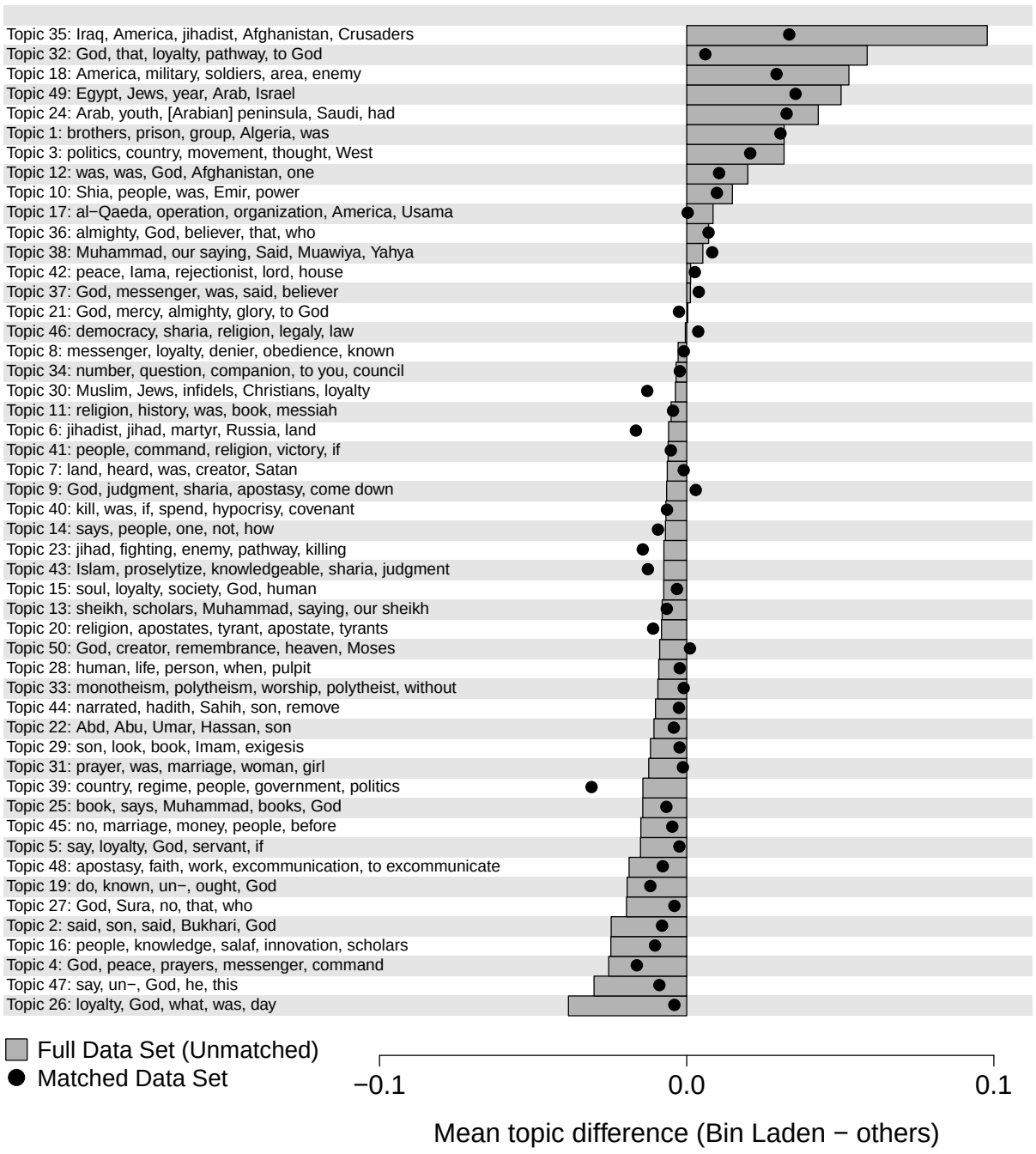


Figure 6: Matching Comparison for Topics in Jihadist Corpus

the American People," by Bin Laden. Cumulative page views of this document are plotted over time in black, with the page views of the 25 most similar documents in gray. All trends are normalized to zero on the day before Bin Laden's death and the matching procedure ensures that

these documents have similar page view trends in the 90 days before targeting. After targeting, "Letter to the American People" experiences a burst in popularity while the control documents have no such burst and slowly diverge over time. This suggests that the matching is successful in balancing the pre-treatment page views and hints that there may be an effect of targeting.

Our modified TIRM procedure produces a matched data set with 466 documents, 33 by Bin Laden and 433 control (some control documents are matched to more than one treated unit, so the data set is weighted accordingly). To estimate treatment effects, we regress post-treatment page views on the indicator for treatment status in the matched data set via weighted least squares. By estimating this regression for each day following treatment, we test whether the average effect of targeting was detectably different from zero on that day. The right panel of Figure 7 shows results of this analysis for Bin Laden, with the black line showing the average treatment effect on any given day and the gray shaded region showing the 95 percent confidence intervals.

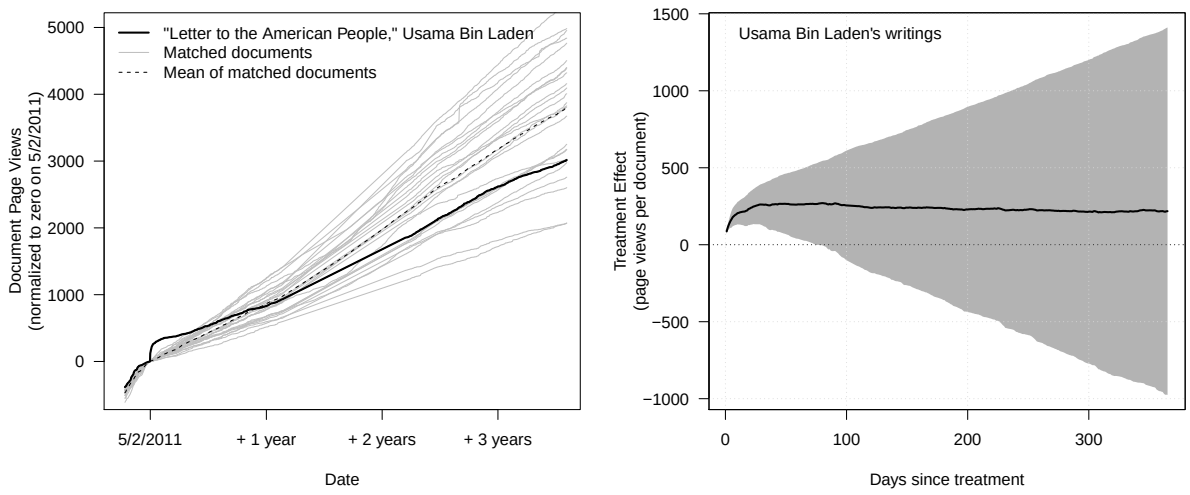


Figure 7: Left: The cumulative page views for a single document by Usama Bin Laden ("Letter to the American People") with the page view trends of 25 matched documents by other authors. Right: The estimated effects of targeting Bin Laden on the popularity of his writings.

The estimated effects in Figure 7 show that targeting Bin Laden increased the popularity of each of his writings by approximately 250 page views on average in the following month. This means that Bin Laden's death resulted in approximately 8,500 more page views for his work than would otherwise have occurred. After 30 days, there are no new page views attributable to targeting and by 90 days after targeting, the effect of targeting is not statistically distin-

guishable from zero. This is due to the widening confidence bands as time goes on (a natural result of the increasing variability in the control trends), rather than to dramatic changes in the treatment effect estimates. Overall, the results indicate that Bin Laden's documents were accessed substantially more in the month after his death, but we do not have sufficient evidence to conclude that his death produced sustained increases in the popularity of his ideas among other jihadists.

Conclusion

As sources of high-dimensional data become more readily available due to advances in measurement and computing, methods for causal inference in high-dimensional data will be increasingly important to social science. In this paper, we introduce the problem of matching for causal inference with high-dimensional covariates and propose several solutions. We focus on applications where causal inferences in observational studies can be improved by measuring confounders from text, but our models could be applied to other types of data.

Our review of existing approaches shows that the problem of reducing covariate dimensionality is not entirely new to the matching literature. Previously proposed matching algorithms reduce dimensionality either by variable selection or by variable coarsening, but these approaches do not work in very high-dimensional covariate spaces. We develop two extensions of existing methods to see whether they might be useful for high-dimensional matching. First, we extend the variable selection logic of propensity score matching (Rosenbaum and Rubin, 1983) to applications where the number of covariates is much larger than the number of cases by estimating balancing scores for texts using multinomial inverse regression (MNIR) (Taddy, 2013*b*). We also extend the coarsening logic of coarsened exact matching (Iacus, King and Porro, 2011) to high-dimensional applications by developing a procedure we call Topical Coarsened Exact Matching (TCEM). In the TCEM approach, we first construct a low-dimensional density estimate of the high-dimensional covariates using a topic model, which extends the logic of coarsened exact matching to incorporate coarsening *across* variables.

While these two methods are useful in their own right, we find that each has weaknesses. MNIR does not produce matched pairs that are recognizable to analysts, making it difficult to evaluate the quality of matches. TCEM does not include information about which confounders are most predictive of treatment, meaning that it works hard to balance many dimensions that end up being irrelevant to causal inference. We propose a third solution that combines

variable selection and variable coarsening approaches into a single procedure called Topical Inverse Regression Matching (TIRM). This method harnesses the power of the newly developed Structural Topic Model to produce matches that are both substantively similar and have similar probabilities of treatment. We evaluate the performance of TIRM through simulation and across three applications and find that it is more useful than MNIR or TCEM in the practical settings we have encountered.

Our interest in matching high-dimensional data 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 problems reflect this diversity. The applications cover three different languages, corpus sizes ranging from 3,201 to 32,295 documents, and typical lengths as short as as a couple of sentences to over 30 pages. Collectively these applications demonstrate that our solution has the flexibility to address the tremendous variety characteristic of social science data.

Beyond this paper, we hope to supplement our simulations and applications by more fully exploring the theoretical properties of the matching methods we propose. As the vast and rapidly growing body of scholarship on matching shows, articulating all of the theoretical properties of a new matching method is beyond the scope of a single article. Nevertheless, we do not believe the remaining theoretical uncertainties should be an impediment to applied work by others in the short term. To assist applied researches wishing to make causal inferences with high-dimensional data, we will release software in the form of an `R` package that extends the existing `stm` package (Roberts, Stewart and Tingley, 2015) to implement the matching procedures described in this paper. Finally, while our research directly extends the matching literature by applying matching methods to text, we hope to apply these approaches to other types of high-dimensional data, including biological and genetic data, images, and high-volume flows.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105(490).

Aronow, Peter M and Cyrus Samii. 2013. "Estimating average causal effects under interference between units." *arXiv preprint arXiv:1305.6156* .

Bamman, David, Brendan O'Connor and Noah Smith. 2012. "Censorship and deletion practices in Chinese social media." *First Monday* 17(3).

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.

Bicego, Manuele, Pietro Lovato, Alberto Ferrarini and Massimo Delledonne. 2010. Biclustering of expression microarray data with topic models. In *International Conference on Pattern Recognition*.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4):77–84.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3:993–1022.

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* .

Bowers, Jake, Mark M Fredrickson and Costas Panagopoulos. 2013. "Reasoning about interference between units: A general framework." *Political Analysis* 21(1):97–124.

Buntine, Wray and Aleks Jakulin. 2004. Applying discrete PCA in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press pp. 59–66.

Cook, R Dennis. 2007. "Fisher lecture: Dimension reduction in regression." *Statistical Science* pp. 1–26.

Cook, R Dennis and Liqiang Ni. 2005. "Sufficient dimension reduction via inverse regression." *Journal of the American Statistical Association* 100(470).

Cronin, Audrey Kurth. 2006. "How al-Qaida Ends: The Decline and Demise of Terrorist Groups." *International Security* 31(1):7–48.

Diamond, Alexis and Jasjeet S Sekhon. 2013. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." *Review of Economics and Statistics* 95(3):932–945.

Esarey, Ashley and Xiao Qiang. 2008. "Political expression in the Chinese blogosphere: Below the radar.".

Friedman, Jerome, Trevor Hastie and Rob Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software* 33(1):1.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* p. mps028.

Hainmueller, Jens. 2011. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* p. mpr025.

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

Hazlett, Chad. 2015. "Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects." Unpublished manuscript.

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

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.

Iacus, Stefano M, Gary King and Giuseppe Porro. 2011. "Multivariate matching methods that are monotonic imbalance bounding." *Journal of the American Statistical Association* 106(493):345–361.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the royal statistical society: series A (statistics in society)* 171(2):481–502.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Imbens, Guido W and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Johnston, Patrick B. 2012. "Does Decapitation Work? Assessing the Effectiveness of Leadership Targeting in Counterinsurgency Campaigns." *International Security* 36(4):47–79.

Jordan, Jenna. 2009. "When heads roll: Assessing the effectiveness of leadership decapitation." *Security Studies* 18:719–755.

Kalathil, Shanthi and Taylor C Boas. 2010. *Open networks, closed regimes: The impact of the Internet on authoritarian rule*. Carnegie Endowment.

King, Gary. 2009. The Changing Evidence Base of Social Science Research. In *The Future of Political Science: 100 Perspectives*, ed. Gary King, Kay Schlozman and Norman Nie. New York: Routledge Press.

King, Gary, Christopher Lucas and Richard Nielsen. 2015. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference.".

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

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722.

King, Gary and Richard Nielsen. 2015. "Why Propensity Scores Should Not Be Used for Matching.".

Kuroki, Manabu and Judea Pearl. 2014. "Measurement bias and effect restoration in causal inference." *Biometrika* 101(2):423–437.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann et al. 2009. "Life in the network: the coming age of computational social science." *Science (New York, NY)* 323(5915):721.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini and Chris Watkins. 2002. "Text classification using string kernels." *The Journal of Machine Learning Research* 2:419–444.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer and Dustin Tingley. 2015. "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(3):254–277.

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

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

Marolt, Peter. 2011. Grassroots agency in a civil sphere? Rethinking internet control in China. In *Online society in China: Creating, celebrating, and instrumentalising the online carnival*, ed. David Herold and Peter Marolt. New York: Routledge pp. 53–68.

Morgan, Stephen L and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

Nielsen, Richard A. 2015. "Can Ideas be "Killed?" Evidence from Counterterror Targeting of Jihadi Ideologues." Unpublished manuscript.

Perina, Alessandro, Pietro Lovato, Vittorio Murino and Manuele Bicego. 2010. *Biologically-aware Latent Dirichlet Allocation (BaLDA) for the Classification of Expression Microarray*. Berlin: Springer-Verlag.

Peterson, Susan and Michael J Tierney. 2009. "Codebook and User's Guide for TRIP Journal Article Database." *Revised May* .

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick and David Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* 38(8):904–909.

Pritchard, Jonathan K, Matthew Stephens and Peter Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics* 155(2):945–959.

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

Roberts, Margaret E. 2015. "Experiencing Censorship Emboldens Internet Users and Decreases Government Support in China." Unpublished manuscript.

Roberts, Margaret E, Brandon M Stewart and Dustin Tingley. 2015. "stm: R package for structural topic models." *R package version 1.1* .

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart and Edoardo Airoldi. Forthcoming. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* .

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Rubin, Donald B. 1980. "Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test"." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 2006. *Matched sampling for causal effects*. New York: Cambridge University Press.

Rubin, Donald B. and Neal Thomas. 1996. "Matching using estimated propensity scores: Relating theory to practice." *Biometrics* 52:249–264.

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.

Spirling, Arthur. 2012. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

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.

Taddy, Matt. 2015*a*. "Distributed Multinomial Regression." *arXiv preprint arXiv:1311.6139* .

Taddy, Matt. 2015*b*. "One-step estimator paths for concave regularization." *arXiv preprint arXiv:1308.5623* .

Wacker, Gudrun. 2003. The Internet and censorship in China. In *China and the Internet: Politics of the digital leap forward*, ed. Christopher R Hughes and Gudrun Wacker. Routledge.

Wang, Xiaogang and Eric Grimson. 2008. Spatial Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 20*, ed. J.C. Platt, D. Koller, Y. Singer and S.T. Roweis. Curran Associates, Inc. pp. 1577–1584.
URL: *http://papers.nips.cc/paper/3278-spatial-latent-dirichlet-allocation.pdf*

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.

Supplemental Information – Online Only

A Inference for Multinomial Inverse Regression

The multinomial inverse regression model involves estimating a large number of coefficients. The coefficient matrix ψ has one row per level of the treatment (so typically 2 in this setting) and one column per word in the vocabulary. We don't however expect that there will be treatment effects on every word in the vocabulary. To simultaneously provide variable selection and estimation we follow Taddy (2013*b*) in estimating the coefficients with a regularizing prior.

Taddy (2013*b*) develops a particular penalization scheme called the Gamma-Lasso. This is a sparsity-inducing concave penalization method that has the attractive property that it is asymptotically unbiased for large coefficients. It is motivated as maximum a-posteriori (MAP) estimation under the Bayesian prior $\psi_v \sim \text{Laplace}(0, \tau_v)$, $\tau_v \sim \text{Gamma}(s, r)$ for some fixed hyperparameters s and r. This prior essentially zeroes out coefficient for words where the ratio of the use under treatment to use control is neither too large or too small.

A direct implementation of the above model would be prohibitively computationally expensive. However, because the sum of multinomial random variables is itself multinomial, we can collapse the word counts by treatment status. This radically simplifies computation. Leveraging later work in Taddy (2015*a*) we also distribute computation across words in the vocab using the connection between the multinomial and poisson. This estimation framework including techniques for computation are further developed in Taddy (2013*b*, 2015*b*,*a*).

B Simulation Details

In this appendix we provide the technical details of the simulation which we summarized in Section 3.3. We wanted to avoid simulating the actual documents, as simulated documents from the topic model data generating process do not look very much like actual documents. Therefore we use the data and model from the female IR scholarship application which was described in Section 4.2. Recall this uses a 15-topic STM on a corpus of 3,201 documents. The covariates for each observation in our simulation X_i that we used to predict both treatment status t_i and the outcome variable y_i (citation counts) were the estimated topic proportions for each document from the Structural Topic Model in Section 4.2 and counts of each of eight words that have high mutual information in the application: "interview", "commiss", "women", "woman", "modest",

"plausibl", "magnitud", and "dubious".

Our overall strategy is as follows:

1. simulate the treatment status from a logistic regression model conditional on the fifteen topics and eight words.

2. simulate the outcome from a poisson regression model conditional on the fifteen topics, eight words and the newly generated treatment.

3. use TIRM to estimate the topics and projection from the original documents and the new treatment indicator.

4. calculate estimates of the poisson regression coefficient representing the treatment effect.

We now describe each of these steps in more detail below.

Treatment Model To generate the treatment status for each document, first we estimated the regression coefficients $\hat{\gamma}$ from a logistic regression of the application's observed treatment status (an indicator variable for all female) on the topic proportions and the individual eight word counts for each document. We estimated the mean of the regression coefficients $\bar{\hat{\gamma}}$ and the variance of the regression coefficients var($\hat{\gamma}$). We then drew each coefficient for the simulation from a univariate Normal distribution based off of these coefficients as follows:

$$\phi_v \sim N(\bar{\hat{\gamma}} + .5, var(\hat{\gamma}))$$

To insure there would be strong confounding from the individual words, we subtracted 3 from the coefficient on Topic 5 and added 3 to the coefficient on the word "magnitude". Using the resulting coefficients ϕ, we generated a simulated treatment status for each unit t_i using a Bernoulli distribution with a logit link:

$$\begin{aligned} \pi_i &= \frac{1}{1+\exp\left(-X_i\phi\right)} \\ t_i &\sim Bernoulli(\pi_i) \end{aligned}$$

Outcome Model To generate the simulated outcome variable for each document, we drew coefficients $\beta_{v,topics}$ for topics from a Normal distribution with mean 0 and variance 1 and the

coefficients for the words $\beta_{v,words}$ from a normal distribution with mean 0 and variance 0.0004.

$$\begin{aligned} \beta_{v,topics} &\sim N(0,1) \\ \beta_{v,words} &\sim N(0,.0004) \end{aligned}$$

As before we ensure a strong confounding of the treatment effect by adding 4 to the coefficient on Topic 5 and subtracting .1 from the coefficient on the word "magnitude". We then generated the outcome variable from a Poisson distribution, setting the coefficient on treatment to -1.

$$y_i \sim Poisson(\exp(-1 * t_i + X_i\beta))$$

TIRM Estimation We generated 200 different sets of outcome variables to produce 200 different simulated datasets. For each of these 200 sets, we estimated the Structural Topic Model on the documents with treatment status as a prevalence and content covariate. We then estimated topic proportions and the projection for TIRM from each STM.

Calculation of Estimates of the Poisson Coefficient Finally, as described in Section 3.3, we calculated the Poisson regression coefficient on treatment for Poisson regression conditioning and not conditioning on topics and projections. We also calculated the coefficient on treatment in the Poisson regression on the weighted dataset produced by CEM matching on the TIRM quantities of interest.

C String kernel results for Chinese Bloggers

In the analysis of Chinese bloggers, we use string kernel similarities between matched blog posts to compare each matched dataset and the unmatched dataset. String kernels examine consecutive sequences of text, in this case groups of five consecutive words, and estimate the percentage of these kernels that are shared between two documents.

Figure 8 shows the density of string kernel similarities between matched blog posts in the MNIR, TCEM, and TIRM matched datasets. For the unmatched dataset, we compare a random sample of treatment and control pairs. We see that MNIR does slightly better than the unmatched dataset in finding similar matches. TCEM and TIRM drastically reduce textual differences between pairs, with TIRM outperforming TCEM.

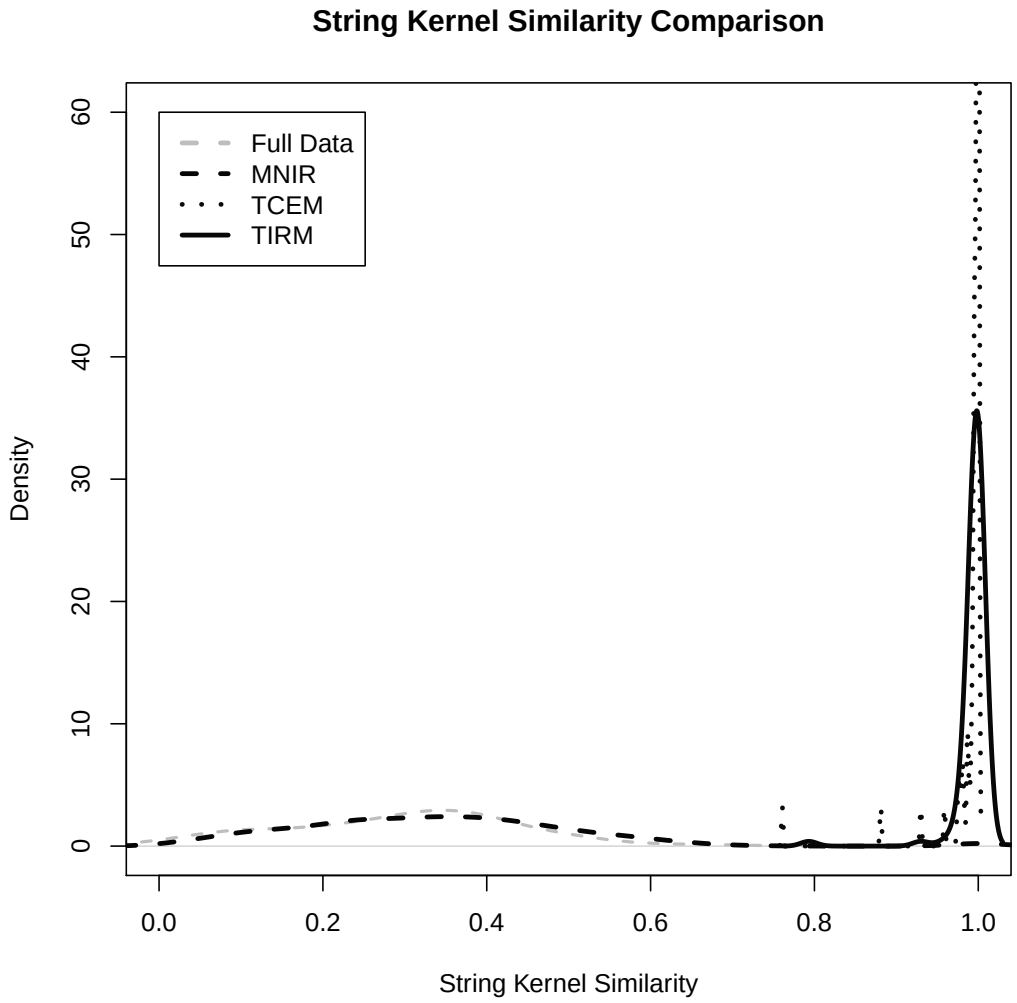


Figure 8: String kernel similarity densities: unmatched, MNIR matched, TCEM matched and TIRM matched.

D Gender Citations Results

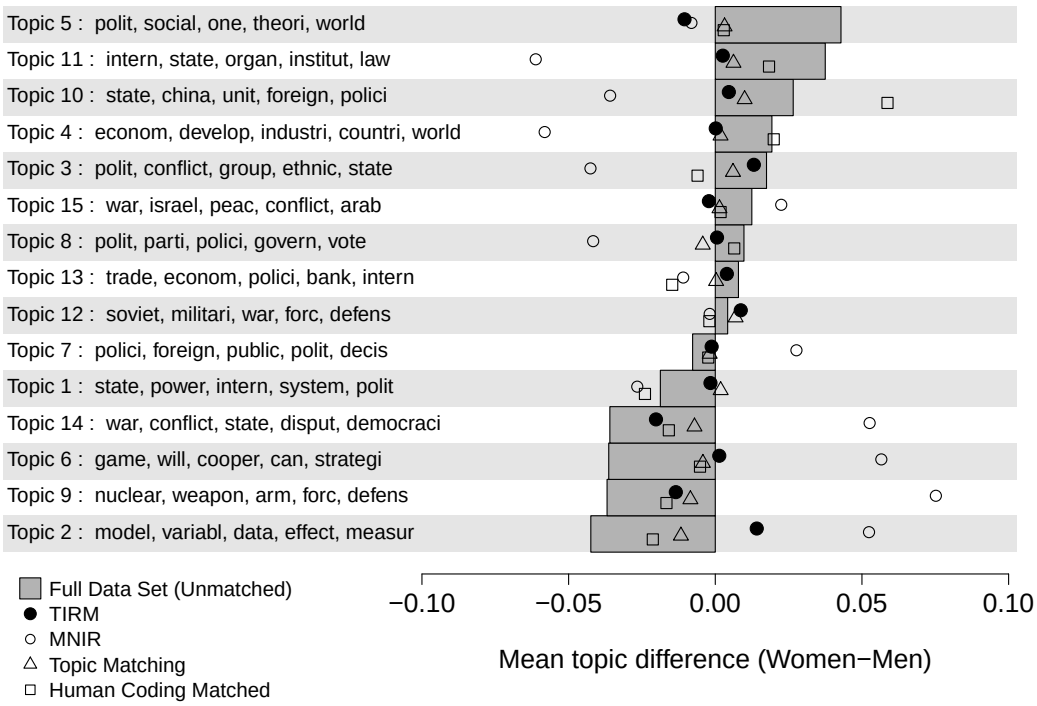


Figure 9: Matching Comparison for Topics

We compare the matching based on human coding to TIRM using a string kernel similarity metric. Figure 10 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.

Here we reproduce the models in Maliniak, Powers and Walter (2013) using the matched data set from automated text matching. 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.

Table 1

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
all_female	-0.694^{***}	-0.931^{***}	-0.968^{***}	-0.660^{***}
	(0.169)	(0.201)	(0.200)	(0.173)
article_age			0.107^{**}	-0.001
			(0.045)	(0.041)
article_age_sq			-0.003^{***}	0.0003
			(0.001)	(0.001)
tenured			-0.492^{***}	-0.548^{***}
			(0.186)	(0.168)
tenured_female		0.544^{*}	1.105^{***}	1.332^{***}
		(0.307)	(0.347)	(0.302)
gender_compAll Female				
gender_compCoed		-0.654^{**}	-0.410	0.023
		(0.327)	(0.327)	(0.288)
coauthored		0.043	0.007	0.037
		(0.187)	(0.186)	(0.169)
R1		-0.163	-0.026	
		(0.154)	(0.150)	
issue_american				
issue_cfp				-0.644
				(0.565)
issue_comparative				-2.752^{***}
				(0.858)
issue_env				1.241
				(0.805)
issue_general				0.318
				(0.961)
issue_health				1.996^{*}
				(1.096)
issue_hist_disc				1.190
				(1.254)
issue_hr				0.833
				(0.644)

Table 1 cont

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
issue_ir				0.940*
				(0.561)
issue_io				0.173
				(0.544)
issue_ipe				0.245
				(0.557)
issue_is				0.445
				(0.545)
issue_meth				0.871
				(0.623)
issue_other				−0.141
				(0.580)
issue_pos				
issue_political_theory				
issue_usfp				−0.366
				(0.609)
meth_qual				−0.151
				(0.254)
meth_quant				−0.050
				(0.255)
meth_exp				−1.074**
				(0.528)
meth_formal				0.256
				(0.229)
meth_anal				0.109
				(0.330)
meth_policy				−0.700
				(0.464)
meth_desc				−0.392
				(0.376)
meth_count				
positivist				1.197***
				(0.256)
material				0.792*
				(0.442)
idea				0.565***
				(0.150)

Table 1 cont

	Dependent variable:			
	Citation Count			
	(1)	(2)	(3)	(4)
AJPS				0.811**
				(0.347)
APSR				0.945**
				(0.421)
BJPS				0.733**
				(0.358)
EJIR				
IO				1.486***
				(0.214)
IS				0.654**
				(0.312)
ISQ				1.023***
				(0.241)
JCR				1.397***
				(0.247)
JOP				1.344**
				(0.522)
SS				
RIPE				
WP				1.522***
				(0.286)
Constant	3.475***	3.611***	2.895***	−0.245
	(0.084)	(0.144)	(0.423)	(0.905)
Observations	289	289	289	289
Note:				*p<0.1; **p<0.05; ***p<0.01

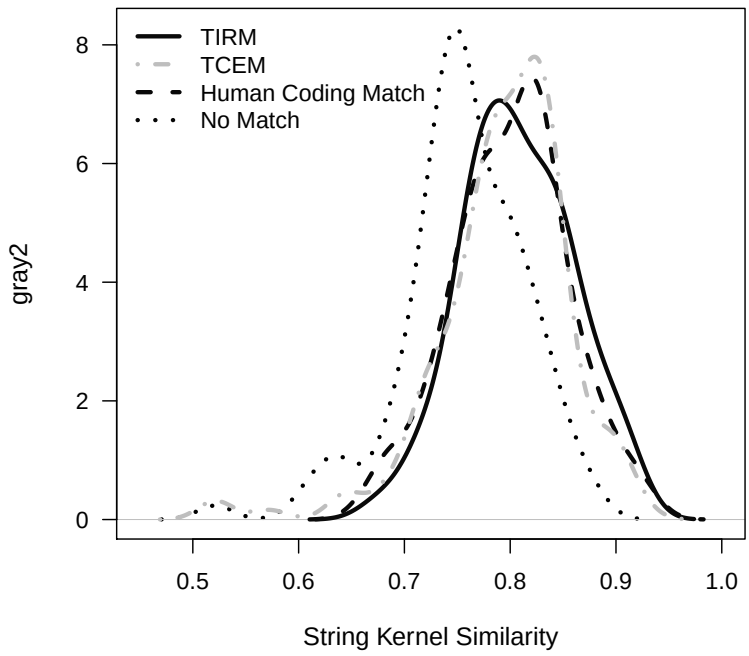


Figure 10: String Kernel Similarity Comparison