Weighted difference-in-differences using text embeddings

Ricardo E. Miranda* Michael Pollmann†

November 6, 2026

PRELIMINARY AND INCOMPLETE – DO NOT CITE OR CIRCULATE

Abstract

Written text contains much otherwise latent information that may lead to more plausible assumptions required for causal interpretations of treatment effect estimates. We propose a method that combines textual information on units of observations with the difference-in-differences identification strategy. Our method finds weights that approximately balances *embeddings* of unit identifiers based on relevant text sources between treated and later- or never-treated units. This weight construction is motivated by an analogy property between embeddings observed in natural language processing (`uncle` $\approx$ `aunt - (woman - man)` in Euclidean space) that mirrors difference-in-differences estimation in semantic space. We demonstrate the empirical validity of the conjectured relationship between embeddings and parallel trends by reanalyzing a large number of studies published in economics. We find that embeddings-weighting reduces observed violations of parallel trends assumptions pre-treatment by 10-30% in the average study.

1 Introduction

The difference-in-differences strategy fundamentally relies on an untestable "parallel trends" assumption to identify causal effects. Different implementation choices, such

*Department of Economics, Duke University. Email: ricardo.e.miranda@duke.edu.
†Department of Economics, Duke University. Email: michael.pollmann@duke.edu.

as using never-treated, later-treated, or not-yet-treated units in the control group, require different parallel trends assumptions. Yet researchers rarely possess ex-ante knowledge justifying any one particular assumption, and tests using pre-treatment data have well-known limitations.

In this paper, we propose using unstructured text data to construct a control group that, ex-ante, more plausibly satisfies the parallel trends assumption. Comprehensive text data plausibly encodes not only standard covariates, but also latent information about the units that is not included in standard analyses. Hence, using text data follows in the tradition of viewing *conditional* parallel trends assumptions as more plausible ex-ante than *unconditional* ones: Units that are more alike in their observable characteristics plausibly also are more alike in their unobservable trends in potential outcomes.

Our approach uses "text embeddings" and an intriguing geometric property that emerges from some models generating embeddings. Modern computational models using text data typically internally represent words (or parts of words) by vectors, called embeddings, of real numbers. Training of these language models adjusts the numerical values constituting the embedding of each word to optimize an objective function, for instance built around correctly predicting the next word in a text document. Computer scientists have noted that the embeddings of some trained models, viewed as vectors in Euclidean space, have an intriguing property, with the frequent example: "$\texttt{king} + \texttt{woman} - \texttt{man} \approx \texttt{queen}$" using standard vector addition and subtraction and cosine similarity or the Euclidean distance to measure the quality of the approximation. Intuitively, the relative locations of embeddings in Euclidean space can reflect logical, or semantic, relationships between the words they represent.

We propose using a weighted parallel trends assumption, with weights constructed to balance text embeddings. Specifically, we find the weighted average of embeddings of untreated observations that minimizes the Euclidean distance to the treated unit's embedding. This construction reflects the observed geometric property that relative locations of embeddings in Euclidean space have semantic meaning. Concretely, suppose that units are U.S. states, there are two periods $t = 0, 1$, and California in period 1 is the only treated observation. Intuitively, if $\texttt{CA} \approx 0.2 \cdot \texttt{CO} + 0.2 \cdot \texttt{MT} + 0.25 \cdot \texttt{NV} + 0.35 \cdot \texttt{UT}$ in embeddings space, then plausibly also $\mathrm{E}(Y_{\texttt{CA},1}(\vec{0}) - Y_{\texttt{CA},0}(\vec{0})) = \mathrm{E}(\sum_s w_s(Y_{s,1}(\vec{0}) - Y_{s,0}(\vec{0})))$, where $Y_{s,t}(\vec{0})$ are untreated potential outcomes for state s at time t, and the weights are $w_{\texttt{CO}} = 0.2$, $w_{\texttt{MT}} = 0.2$, $w_{\texttt{NV}} = 0.25$, $w_{\texttt{UT}} = 0.35$, and

$w_s = 0$ otherwise.

We suggest conducting inference *conditional* on embeddings. The repeated sampling thought experiment asks how different the point estimate would have been across samples with the same embeddings but new draws of outcomes. This thought experiment reflects that we are interested in the precision of our estimate in counterfactual settings closely resembling the observed setting, in particular in how similar the treated and untreated units are to each other. In practice, because the embeddings are trained on large text corpora, we consider the uncertainty associated with random variation in text, holding fixed units' characteristics, as minor. Crucially, the conditional analysis holds fixed the weights which depend on analytically intractable ways on the embeddings of other units in the sample. We propose a practical, tuning-parameter free, approach to estimating the variance of our estimator building on ideas from the literature on matching estimators.

An important limitation of our approach is that it requires multiple appearances of each of the sample units in the text data. Many canonical examples satisfy the requirement: states, countries, large companies, and public figures are all frequently mentioned in news articles, Wikipedia pages, and other text sources. In contrast, the approach is less applicable when the units are "anonymous," such as individual survey respondents, or when the units are not mentioned in text data, such as small towns or villages. In some instances, embeddings for phrases, rather than individual words, may help resolve ambiguity in the text data for units with common names as we explore in an extension.

A key feature of our approach is that researchers can train embeddings on their own text data, rather than relying on pre-trained embeddings. This feature is important because studies will vary in which text data are most relevant as well as in the timing of treatment. The former is important for embeddings to yield benefits, the latter avoids biases when the text data contain post-treatment information. We demonstrate the training of the GloVe (Pennington, Socher, and Manning, 2014) model using four distinct sources of text data: Wikipedia, Congressional speeches, historical newspaper articles, and Form 10-K filings of publicly traded companies.

In extensive replications, the proposed approach reduces realized violations of parallel trends by 10-30% on average across studies. We re-analyze the data from 13 papers published in top general interest and field journals using difference-in-differences and related panel data methods. For each study, we construct event study estimates

for the main outcomes and treatments reported in the original study. Considering only the estimated differences in pre-trends, we find that embeddings-weighting yields substantial decreases in root mean squared error and median absolute error (relative to zero differences between treated and control). In the average application, we therefore expect modest improvements from the proposal. These analyses corroborate our claim that weighting based on text embeddings can render the parallel trends assumption more plausible ex-ante.

In addition to the vast empirical literature using difference-in-differences, our paper relates to multiple strands of the methodological literature. First, our proposal is related to work on conditional parallel trends assumptions and the use of covariates (Heckman, Ichimura, Smith, and Todd, 1998; Abadie, 2005; Sant'Anna and Zhao, 2020; Caetano and Callaway, 2024). One may view embeddings as pre-treatment characteristics, and our weighting approach uses a balancing estimator as an attractive alternative to the more common outcome modeling and inverse probability weighting approaches. Second, our weight construction resembles synthetic controls without using pre-treatment outcomes (Abadie, Diamond, and Hainmueller, 2010) and allowing for unit fixed effects as in synthetic difference-in-differences (Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, 2021). See also Liu (2025) for partial identification under a class of weighted parallel trends assumptions, including synthetic control weights. Our paper proposes an alternative path towards inference based on ideas from matching (Abadie and Imbens, 2006). This proposal sidesteps the challenges of inference in synthetic controls due to weight estimation by effectively conditioning on weights, which is unattractive when using (pre-treatment) outcomes for weight estimation. Because weights in our method are not directly based on pre-treatment outcomes, event study plots also retain some value for assessing the plausibility of the required parallel trends assumption. Third, we connect the literature on balancing weights (Zubizarreta, 2015; Hirshberg and Zubizarreta, 2017; Athey, Imbens, and Wager, 2018) to difference-in-differences; see also (Słoczyński, Uysal, and Wooldridge, 2023). Fourth, our paper contributes to the growing literature on text as data (Gentzkow, Kelly, and Taddy, 2019; Grimmer, Roberts, and Stewart, 2022) using, in particular, embeddings (Pennington et al., 2014; Han, Schulman, Grauman, and Ramakrishnan, 2021; Ash and Hansen, 2023; Kim, Jiang, and Thomadsen, 2023; Jiang and Kim, 2024) to adjust for unstructured text data.

The remainder of the paper proceeds as follows. In Section 2, we introduce the

key ideas behind embeddings and describe how to construct them from raw text data. In Section 3, we discuss the weighted parallel trends assumption, its relation to conditional parallel trends, and our approach to estimation and inference. In Section 4, we provide extensions to staggered adoption, phrase (rather than word) embeddings, and a comparison to estimators based on a conditional parallel trends assumption. In Section 5, we replicate empirical applications from published studies. The paper concludes in Section 6.

2 Text Embeddings

Unstructured text can include substantial latent information about the units in difference-in-differences analyses. Consider an empirical application comparing U.S. states that adopted some policy to states that have not adopted the policy yet. The states differ along many dimensions, such as demographics, industry composition, laws and regulations, and consumer preferences. When units differ in their pre-treatment characteristics, researchers often find assumptions and analyses that adjust for these characteristics more credible to reduce biases associated with differences in them (for instance Heckman et al., 1998; Abadie, 2005; Sant'Anna and Zhao, 2020; Caetano and Callaway, 2024). However, typical covariates only describe a small subset of the dimensions along which units differ. In contrast, written text contains much of humanity's knowledge (Gentzkow et al., 2019): for instance, academic articles, newspapers, encyclopedias, legal documents, and financial filings. The key challenge is *how* researchers can use the information contained in text.

To use text to learn latent information about units, the units must appear in the text corpus. Many empirical applications compare "named" units, such as states, countries, firms, or industries, with recognizable identifiers used in written text. If, in contrast, an application studies a treatment that varies at the level of individuals who are not public figures, or if the data are anonymized such that the names or identities of the units are not known to the researcher, our proposal is not applicable. If units are individuals, but the treatment varies at, say, the state level, we can assign individual-level weights that aggregate to our proposed state-level weights to apply our proposal. We discuss an approach to address ambiguous or non-unique names in Section 4.

Embeddings encode information about a unit (or, more generally, a word) based on

the *context* in which it appears in text. For instance, the GloVe model (Pennington et al., 2014), which we use for this paper, defines the context of word i as the 10 words to the left and the 10 words to the right of word i for each occurrence of i in the text. Let $X_{i,c}$ be the number of times word c appears in the context of word i. Then $P_{i,c} := X_{i,c} / \sum_{c'} X_{i,c'}$ is the proportion of i's context words that is c. The GloVe model approximates the log of context probabilities using a linear factor model, $\log(P_{i,c}) \approx w_i' \tilde{w}_c + b_i + \tilde{b}_c$.[1] See Pennington et al. (2014) for implementation details for estimation based on this approximation. The embedding for word i is then $E_i := (w_i + \tilde{w}_i)/2$. As most analyses and estimators in economics require numerical data, researchers have readily adopted embeddings to encode information contained in unstructured data (for instance Atalay, Phongthiengtham, Sotelo, and Tannenbaum, 2020; Magnolfi, McClure, and Sorensen, 2025; Ash, Chen, and Naidu, 2025).

Embedding models leverage the empirical regularity that conceptual relationships between words tend to shift context probabilities systematically. In the example "$\texttt{queen} - \texttt{woman} \approx \texttt{king} - \texttt{man}$," one such concept is royalty. This concept is the semantic difference between queen and woman as well as between king and man. Empirically, $\log(P_{\texttt{queen},c}/P_{\texttt{woman},c}) \approx \log(P_{\texttt{king},c}/P_{\texttt{man},c})$ for many contexts c. Hence, the concept of royalty shifts context word probabilities consistently across words.[2] The original GloVe model embeddings performed well in benchmark analogy tasks, such as `capital:country` ("Rome is to Italy as Paris is to ?") and `currency:country` ("Dollar is to United States as Yen is to ?") (Pennington et al., 2014). In Appendix A.2, we find that the embeddings we generate in this paper similarly perform well.

For our approach, we conjecture that, similarly, systematic relationships in embeddings (and hence context probabilities) tend to imply conceptual relationships. This conjecture reverses the direction relative to analogy tasks: Analogy tasks test whether conceptual relationships tend to imply systematic relationships in embeddings. We provide two sets of evidence in support of this conjecture.

First and foremost, Section 5 shows that relationships in embeddings tend to imply that trends in untreated potential outcomes are similar across empirical applications and settings. This analysis directly assesses the aspect of our conjecture that is the

[1] The similar Word2Vec model implicitly uses a similar approximation when estimated with the negative-sampling objective (Mikolov, Chen, Corrado, and Dean, 2013; Goldberg and Levy, 2014).

[2] Another such concept is gender, the difference between queen and king as well as between woman and man: Mechanically, also $\log(P_{\texttt{queen},c}/P_{\texttt{king},c}) \approx \log(P_{\texttt{woman},c}/P_{\texttt{man},c})$ by the properties of the logarithm.

most relevant for our proposal of weighted difference-in-differences. We find that, using embeddings to weight units, we can balance trends in untreated potential outcomes across treated and control units better than using unweighted methods.

Second, in analogy tasks, relationships in Euclidean space between embeddings of words tend to imply the corresponding semantic analogies. Using the same analogy datasets we test whether relationships in embeddings reconstruct analogies. For instance, consider the task of reconstructing the embedding for "Paris" using the embeddings for "France" as well as several `capital:country` pairs. If closeness in embeddings implies semantic similarity, we expect to find that the weights within any `capital:country` pair are approximately equal in magnitude but opposite in sign. Additionally, we expect the weight on "France" to be approximately equal in magnitude but opposite in sign to the weight on all other countries.

Beyond the analogy property, we use the GloVe model because it is computationally feasible to train with custom text corpora of moderate size. We provide a step-by-step guide to our process for generating embeddings in Appendix A.1. Different text corpora may contain relevant information about the units in different studies: Financial filings, such as Form 10-K filings, are informative about publicly traded firms, while local news stories are informative about small towns. We also recommend restricting the text to exclude any text written after the start of the sample time or the date of the treatment. Inclusion of post-treatment text risks systematic bias in treatment effect estimates (Ludwig, Mullainathan, and Rambachan, 2024; Sarkar and Vafa, 2024). For instance, it is possible that element $E_{i,k}$ of the embedding for unit i captures information about the realized outcome of unit i at time $t \geq t_i^*$. Then "balancing" embeddings as described in Section 3 also balances post-treatment realized outcomes, rather than counterfactual untreated potential outcomes. In the extreme case, treatment effect estimates are mechanically zero regardless of true effects.

In this paper, we use four distinct, publicly available, text corpora to illustrate the breadth of information available in text. Wikipedia, an online encyclopedia, contains vast amounts of information on diverse topics. We can reconstruct snapshots of Wikipedia going back to 2001 based on its revision history, with convenient historic snapshots available going back to 2008. Hence, Wikipedia cannot provide pre-treatment text for empirical applications with treatment dates prior to 2001. Congressional speech (Gentzkow, Shapiro, and Taddy, 2018, 2019) contains all text spoken in the U.S. House of Representatives and Senate. While the data contain speeches going back

to 1873, the context that the units we study appear in is likely tilted towards political, legal, or regulatory issues. The American Stories Project (Dell, Carlson, Bryan, Silcock, Arora, Shen, D'Amico-Wong, Le, Querubin, and Heldring, 2023) includes local and national U.S. newspaper articles from 1780 to 1963. It provides extensive historical text for applications with early treatment dates, as well as local entities such as minor historical figures, towns, and counties. Form 10-K filings by publicly-traded firms in the U.S. illustrate the use of regulatory and financial documents with specific information relevant to economic outcomes for firms. These filings provide information on firms' business operations, competitive environment, risk factors, and strategic outlook. For our analysis, we obtain filings from 1993 to 2020, offering pre-treatment text for many applications.

The corpora we use in this paper are English language sources and U.S.-centric. However, we stress that neither our proposal nor the GloVe model are specific to these corpora or English language text. In practice, we recommend researchers identify the language and text corpus that is most relevant to their application.

3 Theory: Assumption, Estimation, and Inference

We first present the notation. Subsection 3.1 states the weighted parallel trends assumption formally and contrasts it with the conditional parallel trends assumption. Subsection 3.2 defines the estimator and subsection 3.3 shows how to construct standard errors.

observations We consider settings with a balanced panel of observations. We observe units $i \in \mathcal{I} := \{1, \ldots, n\}$ at times $t \in \mathcal{T} := \{1, \ldots, T\}$. Our approach relies on comparisons between specific subsets of units based on their embeddings. An extension to repeated cross-sections or unbalanced panels with random missingness, while feasible in principle, therefore appears conceptually less attractive to us. If embeddings are only available for groups that aggregate multiple units (such as states when units are individuals), we can apply our approach to the aggregated units as long as the treatment varies only at the aggregate level. In that case, we find weights, as described below, at the aggregate level and distribute the weights evenly to the underlying units.

treatment We study the effect of a binary treatment, $X_{i,t} \in \{0,1\}$. At time $t = 1$, no unit is treated, $X_{i,1} = 0$. Units enter but never exit treatment, $X_{i,s} \geq X_{i,t}$ whenever $s \geq t$. Define the treatment path of unit i as the vector $X_i := (X_{i,t})_{t \in \mathcal{T}} \in \{0,1\}^T$. We use the shorthand $X_i = \vec{1}_t$ if $X_{i,s} = 0$ for all $s < t$ and $X_{i,s} = 1$ for all $s \geq t$. Similarly, we write $\vec{0}$ for the vector of no treatment, with elements $\vec{0}_s = 0$ for $s \in \mathcal{T}$. For ease of exposition, we first focus on settings with a single treated cohort, with the remaining observations never treated. Let $t^* \in \mathcal{T} \setminus \{1\}$ denote the time of treatment for the treated cohort. Then $X_i = \vec{1}_{t^*}$ if unit i is in the treated cohort, and $X_i = \vec{0}$ otherwise. Let $\boldsymbol{X} \in \{0,1\}^{n \times T}$ be the $n \times T$ matrix of treatment paths with elements $X_{i,t}$. All results extend readily to the general case of staggered adoption, as we show in Section 4.1.

outcomes Units are endowed with potential outcomes, which we write as $Y_{i,t}(x)$ for the (possibly counterfactual) outcome of unit i at time t if it was exposed to treatment path x. This notation imposes the assumption that there are no spillovers of the treatment of a unit $j \neq i$ affecting the outcome of unit i. We refer to $Y_{i,t}(\vec{0})$ as untreated potential outcomes. We also assume that the treatment causes no anticipatory effects on outcomes. Under these assumptions, the observed outcomes are

$$Y_{i,t} = \begin{cases} Y_{i,t}(\vec{1}_{t^*}) & \text{if } i: \ X_i = \vec{1}_{t^*} \text{ and } t \geq t^* \\ Y_{i,t}(\vec{0}) & \text{otherwise} \end{cases}. \tag{1}$$

weights For each unit $i \in \mathcal{I}$, we observe a K-dimensional vector of embeddings (typically $K \gg n$), $E_i \in \mathbb{R}^K$, which we use to construct an $n \times n$ matrix of weights, W. Let $\boldsymbol{E} \in \mathbb{R}^{n \times K}$ be the $n \times K$ matrix of embeddings with rows E_i. For each treated unit $i: \ X_i = \vec{1}_{t^*}$, let

$$\begin{aligned} (w_{i,j})_{j: \, X_j = \vec{0}} := \arg \min_{(w_j)_{j: \, X_j = \vec{0}}} & \sum_{k=1}^{K} \Big(E_{i,k} - \sum_{j: \, X_j = \vec{0}} w_j E_{j,k} \Big)^2 \\ \text{s.t.} \quad & \sum_{j: \, X_j = \vec{0}} w_j = 1 \quad \text{and} \quad w_j \geq 0 \text{ for all } j \end{aligned} \tag{2}$$

and $w_{i,j} = 0$ if $X_i = \vec{0}$ or $X_j \neq \vec{0}$. In the rare case where the solution is not unique, we suggest choosing the solution that minimizes the squared norm of the weights. We minimize (the square of) the Euclidean distance to balance differences in embeddings

and associated latent characteristics.[3] Other distance metrics are also conceivable, but the (squared) Euclidean distance has the advantage of yielding a quadratic program in $(w_j)_{j:\, X_j=\vec{0}}$, here with linear (in-)equality constraints, which is computationally fast to solve. We view convex weights as conceptually attractive for typical difference-in-differences settings. However, the example "$\texttt{king} + \texttt{woman} - \texttt{man} \approx \texttt{queen}$" uses non-convex weights; specifically, the weight on `man` is negative.

3.1 Assumption: Weighted Parallel Trends and Its Relation to Conditional Parallel Trends

In this section, we state our key identifying assumption that trends in expected untreated potential outcomes are the same for the treated units $i:\, X_i = \vec{1}_{t^*}$ and the weighted average of untreated units. We then contrast it with the assumption that trends are parallel conditional on covariates.

Assumption 1 (Weighted Parallel Trends)**.** The expected trends in untreated potential outcomes of treated units are equal to the expected trend of the weighted average of untreated units, for any time $t \geq t^*$:

$$\begin{aligned} &\mathrm{E}\Big(Y_{i,t}(\vec{0}) - Y_{i,t^*-1}(\vec{0}) \mid X_i = \vec{1}_{t^*}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{E} = \boldsymbol{e}\Big) \\ =\, &\mathrm{E}\Big(\sum_{j:\, X_j=\vec{0}} w_{i,j}\Big(Y_{j,t}(\vec{0}) - Y_{j,t^*-1}(\vec{0})\Big) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{E} = \boldsymbol{e}\Big) \end{aligned} \tag{3}$$

where expectations are taken over randomness in potential outcomes conditional on embeddings (and hence weights), holding fixed treatment paths of all units.

Assumption 1 allows us to impute the "missing" untreated potential outcomes of treated units $i:\, X_i = \vec{1}_{t^*}$ at times after treatment $t \geq t^*$. We define the corresponding weighted difference-in-differences estimator explicitly in Section 3.2.

The weighted parallel trends Assumption 1 resembles, in some aspects, the assumption of parallel trends conditional on covariates. Consider, specifically, the special

[3]Prior research in computer science that notes how linear combinations of embeddings reflect semantic meaning measure it using cosine similarity (see Appendix A.2). When embedding vectors are of equal norm, Euclidean distance is a monotonic transformation of cosine similarity. When the norm differs across units, the norm itself may contain information about units that is valuable to balance.

case where we use the embeddings as covariates. The conditional parallel trends assumption states that

$$\begin{aligned} &\mathrm{E}\Big(Y_{i,t}(\vec{0}) - Y_{i,t^*-1}(\vec{0}) \mid X_i = \vec{1}_{t^*}, E_i = e\Big) \\ =&\mathrm{E}\Big(Y_{j,t}(\vec{0}) - Y_{j,t^*-1}(\vec{0}) \mid X_j = \vec{0}, E_j = e\Big). \end{aligned} \tag{4}$$

Equations 3 and 4 both equate conditional means of the trends of treated ($X_i = \vec{1}_{t^*}$) and untreated ($X_j = \vec{0}$, through the summation in the case of equation 3) units. In both equations, the expectations condition on (some of) the embeddings ($\boldsymbol{E}$ and E, respectively) and treatment paths ($\boldsymbol{X}$ and X, respectively).

The assumptions differ substantively, however, in the across-unit comparisons they justify, unless there are exact matches in embeddings space. The weighted parallel trends assumption justifies comparing treated units to particular weighted averages of untreated units. The weights, determined through the optimization in equation 2, interpolate between untreated units to construct a counterfactual trend for each treated unit. The conditional parallel trends assumption, in contrast, only justifies comparisons between treated and untreated units with identical embeddings. With exact matches, both assumptions coincide (up to non-uniqueness of the weights) because placing weight only on exact matches optimizes the criterion in equation 2.

3.2 Estimation

Define the average treatment effect on the treated (ATT) at each post-treatment time $t \geq t^*$

$$\tau_t := \mathrm{E}\Big(Y_{i,t}(\vec{1}_{t^*}) - Y_{i,t}(\vec{0}) \mid X_i = \vec{1}_{t^*}\Big), \tag{5}$$

For the conditional inference exercise below, we also define the corresponding conditional ATT among the realized treated units

$$\tau_t^c := \frac{1}{n_{t^*}} \sum_{i:\, X_i = \vec{1}_{t^*}} \mathrm{E}\Big(Y_{i,t}(\vec{1}_{t^*}) - Y_{i,t}(\vec{0}) \mid \boldsymbol{X}, \boldsymbol{E}\Big). \tag{6}$$

We use the weighted difference-in-differences estimator

$$\hat{\tau}_t := \frac{1}{n_{t^*}} \sum_{i:\, X_i = \vec{1}_{t^*}} \left(\Big(Y_{i,t} - Y_{i,t^*-1}\Big) - \sum_{j:\, X_j = \vec{0}} w_{i,j} \Big(Y_{j,t} - Y_{j,t^*-1}\Big) \right) \tag{7}$$

where $n_{t^*} := \sum_{i \in \mathcal{I}} \mathbb{1}\{X_i = \vec{1}_{t^*}\}$ is the number of treated units.

Theorem 1. *If Assumption 1 holds, there is at least one treated unit ($n_{t^*} > 0$), and there are no anticipatory effects and spillovers, then the weighted difference-in-differences estimator is conditionally unbiased for the conditional ATT at each post-treatment time $t \geq t^*$:*

$$\mathrm{E}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E}) = \tau_t^c. \tag{8}$$

Proof: *Follows standard arguments, see Appendix B.1.*

Theorem 1 shows unbiasedness under our weighted parallel trends assumption. Our estimator also yields an approximately unbiased estimate if the conditional mean function $\mathrm{E}(Y_{i,t}(\vec{0}) - Y_{i,t-1}(\vec{0}) \mid E_i)$ is linear in embeddings for each i and t. Specifically, the worst-case bias is a function of the discrepancy in embeddings between treated and the weighted average of untreated units. This result follows from Zubizarreta (2015, Proposition 4.1).

3.3 Inference

We propose inference conditional on embeddings $\boldsymbol{E}$ and treatment paths $\boldsymbol{X}$. The repeated sampling thought experiment here compares samples of counterfactual worlds that strongly resemble the real world: Embeddings (derived from detailed text data) and treatment choices are identical, but do not uniquely determine the outcomes, which vary across counterfactual worlds. Empirical researchers may find a thought experiment that most closely resembles the sample at hand intrinsically appealing. In practice, it often leads to smaller standard errors for both pre-trends and effect estimates. This conditional exercise targets τ_t^c and omits variation in $\mathrm{E}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E})$ across repeated samples with different embeddings or treatment paths. Inference for the unconditional target τ_t would require this additional variance component.

Conditional inference resolves a fundamental challenge in weighted difference-in-differences by fixing the weights. The standard difference-in-differences estimator uses weights that depend only on the number of treated and untreated units. Those weights are therefore analytically tractable under various repeated sampling thought experiments. In contrast, the weighted difference-in-differences estimator uses weights that depend on the identities (or at least the embeddings) of other untreated units in the sample. Hence, in repeated sampling thought experiments that resample units

from a population of units with different embeddings, it is challenging to analytically control how weights change from sample to sample.

Assumption 2. The dependence of outcomes across units is limited, conditional on treatment paths and embeddings:

(a) $\mathrm{cov}(Y_{i,t} - Y_{i,t^*-1}, Y_{i',t} - Y_{i',t^*-1} \mid \boldsymbol{X}, \boldsymbol{E}) = 0$ for $i \neq i'$; and

(b) $\mathrm{E}(Y_{i,t} - Y_{i,t^*-1} \mid \boldsymbol{X}, \boldsymbol{E}) = \mathrm{E}(Y_{i,t} - Y_{i,t^*-1} \mid X_i, E_i)$ and
$\mathrm{var}(Y_{i,t} - Y_{i,t^*-1} \mid \boldsymbol{X}, \boldsymbol{E}) = \mathrm{var}(Y_{i,t} - Y_{i,t^*-1} \mid X_i, E_i)$ for each i.

To motivate our estimator for the variance of $\hat{\tau}_t$, we write

$$\hat{\tau}_t = \sum_{i: X_i = \vec{1}_{t^*}} \frac{1}{n_{t^*}} (Y_{i,t} - Y_{i,t^*-1}) - \sum_{j: X_j = \vec{0}} \bar{w}_j (Y_{j,t} - Y_{j,t^*-1}), \qquad \bar{w}_j := \frac{1}{n_{t^*}} \sum_{i:\, X_i = \vec{1}_{t^*}} w_{i,j}. \tag{9}$$

Under Assumption 2, we can decompose the conditional variance of $\hat{\tau}_t$ into a weighted sum of variances of terms $(Y_{i,t} - Y_{i,t^*-1})$ for $i \in \mathcal{I}$.

Let $\mathcal{I}_+$ denote the units with positive squared weight in this decomposition:

$$\mathcal{I}_+ := \{i :\, X_i = \vec{1}_{t^*}\} \cup \{j :\, X_j = \vec{0},\, \bar{w}_j > 0\}. \tag{10}$$

We estimate the variance of i's trend using i's closest within-cohort match. Assume that every treatment-path cohort represented in $\mathcal{I}_+$ contains at least two units. If several units solve the nearest-neighbor problem, choose one using any fixed tie-breaking rule. For each unit $i \in \mathcal{I}_+$, find the closest match in embeddings, restricting to units with identical treatment paths:

$$m(i) := \arg \min_{i' :\, X_i = X_{i'},\, i' \neq i} \sum_{k=1}^{K} (E_{i,k} - E_{i',k})^2. \tag{11}$$

Then we estimate the conditional variance for unit i by

$$\hat{\sigma}_i^2 := \frac{1}{2} \Big(\big(Y_{i,t} - Y_{i,t^*-1} \big) - \big(Y_{m(i),t} - Y_{m(i),t^*-1} \big) \Big)^2. \tag{12}$$

Subtracting the trend of $m(i)$ (approximately) cancels the mean of i's trend in expectation. Squaring estimates the variance of both i and $m(i)$ in expectation. The covariance is zero if the random components of outcomes are uncorrelated across units.

To estimate only i's variance, we divide the result by 2, using that both variances are approximately equal if $m(i)$ is a good match. This construction is frequently used to estimate the variance of matching estimators of treatment effects (Abadie and Imbens, 2006).

Our estimate of the conditional variance of $\hat{\tau}_t$ sums over $\hat{\sigma}_i^2$ using the square of the unit-level weights in $\hat{\tau}_t$:

$$\hat{\sigma}^2 := \sum_{i:\, X_i = \vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \hat{\sigma}_i^2 + \sum_{j:\, X_j = \vec{0},\, \bar{w}_j > 0} \bar{w}_j^2 \hat{\sigma}_j^2 \tag{13}$$

We make an assumption that any match discrepancies do not systematically lead to large under-estimates of the variance. To state the assumption compactly, define

$$\begin{aligned} \mu(x,e) &:= \mathrm{E}(Y_{i,t} - Y_{i,t^*-1} \mid X_i = x, E_i = e), \qquad \mu_i := \mu(X_i, E_i) \\ \sigma^2(x,e) &:= \mathrm{var}(Y_{i,t} - Y_{i,t^*-1} \mid X_i = x, E_i = e), \quad \sigma_i^2 := \sigma^2(X_i, E_i) \end{aligned} \tag{14}$$

Assumption 3. On average, the conditional variance of $m(i)$'s outcome does not understate the conditional variance of i's outcome by more than the squared difference in conditional means between $m(i)$ and i:

$$\sum_{i \in \mathcal{I}_+} \omega_i (\sigma_i^2 - \sigma_{m(i)}^2) \leq \sum_{i \in \mathcal{I}_+} \omega_i (\mu_i - \mu_{m(i)})^2 \tag{15}$$

where $\omega_i := (1/n_{t^*})^2$ if $X_i = \vec{1}_{t^*}$ and $\omega_i := \bar{w}_i^2$ if $X_i = \vec{0}$ and $\bar{w}_i > 0$.

Assumption 3 is a high-level sufficient condition. It summarizes the second-moment restrictions needed for the nearest-neighbor variance estimate to be conservative, rather than deriving them from primitive smoothness or increasing-sample matching conditions. There are two sources of bias for $\hat{\sigma}^2$ relative to the true conditional variance, both caused by imperfect matches. First, imperfect matches fail to exactly cancel the conditional mean of the trend of i before squaring. This source necessarily moves us towards over-estimating the true variance. Second, imperfect matches mean that the conditional variance of the trend of $m(i)$ may differ from that of i. This source can move us towards over- or under-estimating the true variance if the direction of bias happens to be systematic in either direction.

Assumption 3 states that the second source, even if it is systematic towards under-estimation of the variance does not overwhelm the first source of bias. Differencing of

outcomes over time removes the time-invariant unobserved heterogeneity. There is typically little reason to expect heteroskedasticity in the remaining trend to be large and systematically biased towards under-estimating the variance. For the conditional mean, in contrast, we expect discrepancies in embeddings to cause discrepancies in expected trends: This feature is the primary motivation for the weighted difference-in-differences analysis, and hence moves us towards over-estimating the variance. If, nevertheless, the validity of Assumption 3 appears questionable, doubling the variance estimate (multiplying standard errors by $\sqrt{2} \approx 1.4$) gives a conservative variance estimate in expectation without this assumption.

Theorem 2. *If Assumptions 2 and 3 hold and there are at least two units in each cohort represented in $\mathcal{I}_+$, then*

$$\mathrm{E}(\hat{\sigma}^2 \mid \boldsymbol{X}, \boldsymbol{E}) \geq \mathrm{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E}) \tag{16}$$

with equality whenever Assumption 3 holds with equality.
If only Assumption 2 holds and there are at least two units in each cohort represented in $\mathcal{I}_+$, then

$$\mathrm{E}(2\hat{\sigma}^2 \mid \boldsymbol{X}, \boldsymbol{E}) \geq \mathrm{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E}). \tag{17}$$

Proof: *Uses ideas from matching estimators of treatment effects, see Appendix B.2.*

Theorems 1 and 2 show that $\hat{\tau}_t$ is conditionally unbiased for τ_t^c, and that $\hat{\sigma}^2$ is conservative in expectation for the conditional variance of $\hat{\tau}_t$. To obtain asymptotically valid confidence intervals based on a normal approximation, one could specify additional large-sample conditions. For example, one can consider a triangular array of units, strengthen Assumption 2(a) to conditional independence, impose a Lindeberg condition, assume no unit's squared aggregate weight times conditional variance dominates the total conditional variance, and require the nearest-neighbor variance estimates to aggregate to their conditional expectations or conservative limits. While treated units receive equal weight $1/n_{t^*}$, the aggregate weights $\bar{w}_j$ could, in principle, concentrate on a small number of untreated units. In practice, researchers can inspect a histogram of weights $\bar{w}_j$ to decide whether the approximation due to asymptotic normality based on averaging many small units appears appropriate.

4 Extensions

4.1 Staggered Adoption

The theoretical results of Section 3 generalize to settings with staggered adoption with slight modifications to the assumptions and estimators. Here, we focus on using the not-yet-treated, rather than the never-treated, as the control group because this choice requires the most modifications. We also use only the time immediately preceding treatment as the baseline outcome. Variations of both of these aspects are straightforward.

notation Denote the time of unit i's treatment by $t_i^* \in \{2, \ldots, T\} \cup \{\infty\}$, with $t_i^* = \infty$ for never-treated units. Define $t_{\max}^* = \max_{i=1,\ldots,n} t_i^*$ as the time in which the last cohort receives treatment ($t_{\max}^* = \infty$ if there are never-treated units).

We present results for event-study estimates that aggregate cohort-time specific effects by time relative to treatment $h \geq 0$. Similar results hold for $h < 0$ and other averages over treatment effects. Let $\mathcal{I}_h := \left\{ i \in \mathcal{I} : \; t_i^* + h \leq \min\{T, t_{\max}^* - 1\} \right\}$ be the set of units for whom we observe h periods post treatment ($t_i^* + h \leq T$) with untreated units remaining at that time ($t_i^* + h \leq t_{\max}^* - 1$).

weights We index weights by time t to only place weight on outcomes of units that are untreated at this time, for each i and t such that $t_i^* \leq t \leq \min\{T, t_{\max}^* - 1\}$:

$$\begin{aligned}(w_{i,j,t})_{j:\, t_j^* > t} := \arg \min_{(w_j)_{j:\, t_j^* > t}} & \sum_{k=1}^{K} \Big(E_{i,k} - \sum_{j:\, t_j^* > t} w_j E_{j,k} \Big)^2 \\ \text{s.t.} \quad & \sum_{j:\, t_j^* > t} w_j = 1 \quad \text{and} \quad w_j \geq 0 \text{ for all } j.\end{aligned} \tag{18}$$

By convention, let $w_{i,j,t} = 0$ for all other (i, j, t).

parameter of interest We are interested in the average treatment effect h periods after treatment for the treated units:

$$\tau_h := \frac{1}{n_{\vec{1}}} \sum_{i \in \mathcal{I}_h} \mathrm{E}\Big(Y_{i,t_i^*+h}(\vec{1}_{t_i^*}) - Y_{i,t_i^*+h}(\vec{0}) \mid \boldsymbol{X}, \boldsymbol{E} \Big). \tag{19}$$

assumption We assume that trends in untreated outcomes from one period to the next are similar between the treated and the weighted average of not-yet-treated units.

Assumption 4. For each unit i with $t_i^* = t^* \in \mathcal{T}$ and time t such that $t^* \leq t < t^*_{\max}$,

$$\begin{aligned} &\mathrm{E}\Big(Y_{i,t}(\vec{0}) - Y_{i,t-1}(\vec{0}) \mid X_i = \vec{1}_{t^*}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{E} = \boldsymbol{e}\Big) \\ =&\mathrm{E}\Big(\sum_{j:\, t_j^* > t} w_{i,j,t}\Big(Y_{j,t}(\vec{0}) - Y_{j,t-1}(\vec{0})\Big) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{E} = \boldsymbol{e}\Big) \end{aligned} \tag{20}$$

where expectations are taken over randomness in potential outcomes conditional on embeddings (and hence weights), holding fixed treatment paths of all units.

In the literature, our Assumption 4 most closely resembles Assumption 3 of Bellégo, Benatia, and Dortet-Bernadet (2025) albeit with weighting and not-yet-treated units forming the control group. Alternative assumptions using long-differences or never-treated units as the control group are straightforward to state and yield analogous estimators and results. The chained short-differences of the not-yet-treated units presented here represents the most complex formulas. With alternative assumptions, the analogous formulas below are generally simpler. However, chaining of single period trends (Bellégo et al., 2025) of the control group avoids introducing large changes in estimates across event times due to compositional changes in the control group. Hence, we recommend this approach in practice.

estimation Assumption 4 suggests that we chain together weighted short-difference trends of the not-yet-treated units to impute counterfactual outcomes for treated units. Define the estimators

$$\begin{aligned} \hat{\tau}_{i,h} &:= \Big(Y_{i,t_i^*+h} - Y_{i,t_i^*-1}\Big) - \sum_{s=0}^{h} \sum_{j:\, t_j^* > t_i^*+s} w_{i,j,t_i^*+s}\Big(Y_{j,t_i^*+s} - Y_{j,t_i^*+s-1}\Big) \\ \hat{\tau}_h &:= \frac{1}{n_{\vec{1}}} \sum_{i \in \mathcal{I}_h} \hat{\tau}_{i,h} \end{aligned} \tag{21}$$

with $n_{\vec{1}} := |\mathcal{I}_h|$ the number of units for whom we can estimate the effect h periods post treatment.

variance We allow for arbitrary within-unit dependence of outcomes over time by constructing a match for the unit-specific weighted sum over time. Rewrite $\hat{\tau}_h$ as a

weighted sum of outcomes over all units and time periods:

$$\begin{aligned}
\hat{\tau}_h &= \sum_{i=1}^{n}\sum_{t=1}^{T} \frac{w_{i,t}^{\vec{1}} - w_{i,t}^{\vec{0}}}{n_{\vec{1}}} Y_{i,t}; \\
w_{i,t}^{\vec{1}} &:= \mathbb{1}\{t = t_i^* + h\} - \mathbb{1}\{t = t_i^* - 1\} \text{ if } i \in \mathcal{I}_h, \quad \text{and 0 otherwise;} \\
w_{i,t}^{\vec{0}} &:= \sum_{j \in \mathcal{I}_h} \Big(w_{j,i,t} \sum_{s=0}^{h} \mathbb{1}\{t = t_j^* + s\} - w_{j,i,t+1} \sum_{s=0}^{h} \mathbb{1}\{t = t_j^* + s - 1\} \Big)
\end{aligned} \tag{22}$$

The weight $w_{i,t}^{\vec{1}}$ is due to i's appearance as a treated unit. The weight $w_{i,t}^{\vec{0}}$ is due to i's appearance as a control unit for other treated units.

Then we estimate the conditional variance of $\hat{\tau}_h$ by

$$\hat{\sigma}_h^2 := \sum_{i=1}^{n} \frac{1}{2} \Big(\sum_{t=1}^{T} \frac{w_{i,t}^{\vec{1}} - w_{i,t}^{\vec{0}}}{n_{\vec{1}}} (Y_{i,t} - Y_{m(i),t}) \Big)^2 \tag{23}$$

using the same approach to find within-cohort matches $m(i)$ as in Section 3.

With staggered adoption, there may be few units within any given cohort, leading to relatively large match discrepancies. In principle, we may find a match $m(i)$ for unit i from a different cohort than t_i^*. Intuitively, a match from a different cohort yields smaller match discrepancies due to distance in embeddings space, but larger match discrepancies due to mismatch of the treatment state. Suppose $\tilde{m}(i)$ is the match for i from a different treatment state. The match discrepancies in mean outcomes for unit i using $m(i)$ and $\tilde{m}(i)$ are

$$\Big(\sum_{t=1}^{T} \frac{w_{i,t}^{\vec{1}} - w_{i,t}^{\vec{0}}}{n_{\vec{1}}} (\mu_{i,t} - \mu_{m(i),t}) \Big)^2 \qquad \text{and} \qquad \Big(\sum_{t=1}^{T} \frac{w_{i,t}^{\vec{1}} - w_{i,t}^{\vec{0}}}{n_{\vec{1}}} (\mu_{i,t} - \mu_{\tilde{m}(i),t}) \Big)^2 \tag{24}$$

where $\mu_{i,t} := \mathrm{E}(Y_{i,t} \mid X_i, E_i)$. For $m(i)$, $X_i = X_{m(i)}$, so the only discrepancy is due to $E_i - E_{m(i)} \neq 0$. For $\tilde{m}(i)$, this first discrepancy is smaller, but $X_i \neq X_{\tilde{m}(i)}$. Hence, $\mu_{i,t}$ and $\mu_{\tilde{m}(i),t}$ may be expected potential outcomes under different treatment states. The conditional variance behaves analogously, but discrepancies can yield an over- or understatement of the variance as discussed around Assumption 3. In practice, we recommend using within-cohort matches whenever possible, and adjacent cohorts when necessary.

results As in Section 3, the estimator $\hat{\tau}_h$ is unbiased for τ_h if Assumption 4 holds, and $\hat{\sigma}_h^2$ is conservative for $\text{var}(\hat{\tau}_h \mid \boldsymbol{X}, \boldsymbol{E})$ if generalizations of Assumptions 2 and 3 to multiple time periods hold. We omit formal statements of the assumptions and results for brevity.

4.2 Embeddings for Phrases

In some applications, researchers may prefer embeddings for short phrases rather than individual words. Here, we discuss two potential advantages of using phrase embeddings, as well as two disadvantages.

The first advantage of phrase embeddings is that it may help to disambiguate the identity of units. When training our word embeddings using the GloVe model, we use text replacement as a preprocessing step to replace multi-word unit names with a single word. For instance, we can replace "New York City" with "NewYorkCity." However, such replacements cannot, at scale, address ambiguities such as "New York" referring to either "New York City" or "New York State," without keywords in the context deterministically distinguishing between meanings. Similarly, "Apple" may refer to the fruit or the company. A model of phrase embeddings does not require such preprocessing and provides an embedding for "New York City" or "Apple Inc." not only based on appearances of the entire phrase, but also based on appearances of the individual words in the phrase.

The second advantage of phrase embeddings is that they may be more outcome-specific than word embeddings. For instance, suppose the outcome variable is the unemployment rate in a state. Weights based on averaging embeddings of "unemployment in `untreated state`" to replicate the embedding of "unemployment in `treated state`" may be more predictive of the counterfactual unemployment rate than weights based on averaging embeddings of "`untreated state`" to replicate the embedding of "`treated state`."

The first disadvantage is that phrase embeddings are more difficult to train than word embeddings. While pre-trained models are readily available, they pose the risk of leaking post-treatment information into the embeddings. The Sent2vec model of Pagliardini, Gupta, and Jaggi (2018) is one of the first dedicated phrase embedding models, trained with text up to 2017. Most more recent models are trained on more recent text. For our illustration of the ideas in this section, we trained the Sent2vec

model on our largest text corpus.

The second disadvantage is that phrase embeddings create additional researcher degrees of freedom in the choice of phrases to embed. In unweighted difference-in-differences, the only weight-choice for the researcher is in the inclusion of units. Word embeddings offer an interesting trade-off: In theory, researchers can augment the input and tuning parameters at the training stage. In practice, however, the computational cost of intentional, directional manipulation is typically large. The in- or exclusion of a specific unit may, in contrast, often have less impact on the estimate as the simplex constraint on weights sets many weights to 0. Phrase embeddings, however, allow the researcher to manipulate the exact phrase to be embedded for each unit. This manipulation occurs after the training stage at minimal computational cost. Hence, we recommend researchers use phrase embeddings only when there are compelling reasons to do so.

4.3 Relation to Covariates under Conditional Parallel Trends

When there are no exact matches in embeddings or covariates, a variety of strategies based on the conditional parallel trends assumption exists. These strategies differ in their properties relative to our strategy based on weighted parallel trends.

First, one may resort to nearest neighbor comparisons based on continuity arguments and small discrepancies. This strategy then extrapolates, rather than interpolates. See Kellogg, Mogstad, Pouliot, and Torgovitsky (2021) for a detailed discussion of interpolation and extrapolation in a closely related context.

Second, one may use inverse probability weighting, exploiting that equation 4 implies that

$$\begin{aligned} &\mathrm{E}\Big(Y_{i,t}(\vec{0}) - Y_{i,t^*-1}(\vec{0}) \mid X_i = \vec{1}_{t^*}\Big) \\ &= \mathrm{E}\Big(\tilde{w}_j\Big(Y_{j,t}(\vec{0}) - Y_{j,t^*-1}(\vec{0})\Big) \mid X_j = \vec{0}\Big) \end{aligned} \tag{25}$$

where $\tilde{w}_j := \frac{p(E_j)}{1-p(E_j)} \frac{\Pr(X_i=\vec{0})}{\Pr(X_i=\vec{1}_{t^*})}$ using the propensity score $p(e) := \Pr(X_i = \vec{1}_{t^*} \mid E_i = e)$. The weights $\tilde{w}_j$, in general, are distinct from the (properly normalized) aggregate weights $(1/n_{t^*}) \sum_{i:\, X_i=\vec{1}_{t^*}} w_{i,j}$ underlying our weighted parallel trends strategy. However, Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) provide a frequently cited rule of thumb that one needs at least 10 events per covariate for accurate estimation of the propensity score function. Hence, with only a small number of units per cohort,

approaches based on the propensity score, including "double robust" augmented inverse probability weighting, can reliably use only one or two covariates in the typical applications for our proposal.

Third, one may impose functional form assumptions on the conditional expectations in equation 4, commonly linearity. The results of Zubizarreta (2015) imply that, if the conditional mean function is linear in covariates, estimators like the one we propose in this paper also yield unbiased estimates of treatment effects whenever "exact balance" in covariates is possible. Here, exact balance refers to no discrepancy in embeddings between treated and the weighted average of untreated units. If we drop the convexity constraints in the optimization problem defining the weights, exact balance is possible whenever the OLS estimator of the linear mean function exists. See Chattopadhyay and Zubizarreta (2023); Ben-Michael, Feller, Hirshberg, and Zubizarreta (2021); Słoczyński et al. (2023); Bruns-Smith, Dukes, Feller, and Ogburn (2025) and references therein for a detailed discussion of the relationship between weighting and other estimation approaches.

In summary, our strategy based on weighted parallel trends is related to but also has distinct features from common strategies based on parallel trends conditional on embeddings. In Section 5, we explore if, in some applications, embeddings-based weights also resemble the implicit weights arising from strategies that condition on standard, not text- or embeddings-based, covariates.

5 Applications

We demonstrate that pre-trends based on embeddings-weighting tend to be smaller, on average, than pre-trends in standard event studies. We reanalyze data from 13 studies estimating difference-in-differences or related panel data specifications.

5.1 Aggregate Analysis

In recent studies estimating difference-in-differences specifications, event study plots typically provide important suggestive evidence in support of a parallel trends assumption. These event study plots show estimates of effects aggregated by time relative to the onset of treatment. The estimates typically aggregate over multiple units, treated at the same or different calendar times. If the parallel trends assumption holds also

prior to treatment, event study estimates for negative relative event times (preceding treatment) are zero in expectation. Researchers interpret the magnitude of realized point estimates preceding treatment: Point estimates close to zero are typically viewed as suggestive of the validity of the parallel trends assumption.

To provide summary statistics that are representative of the improvements in typical studies, we weight each study equally. Some studies report estimates for multiple different treatments or outcomes. For any such specification, we estimate event studies yielding point estimates by relative event time. We restrict the analysis to negative relative event times. We take the estimate for a given study, specification, and event time as the basic object of analysis. We normalize the estimate by the standard deviation of outcomes across observed untreated potential outcomes (pre-treatment outcomes of later-treated units, and all outcomes of never-treated units). This normalization allows comparisons of the magnitude of differential pre-trends across specifications and applications. For aggregate summaries, we assign equal weight to each (negative) relative event time within a specification. We then assign equal aggregate weight to each specification within a study. Finally, we assign equal aggregate weight to each study.

Table 1 show that embedding-weighting reduces pre-trends by 10-30% in the average study.[4] Each row shows results for a different text corpus. Summary statistics show performance relative to standard event studies. The first column (MAE) shows the median absolute error of estimated differential pre-trends (relative to zero differential pre-trends) when using embeddings-weighting relative to the identically-computed median absolute error for standard event studies. For each text corpus, the MAE is below 100% of the standard event studies, indicating that trends prior to treatment are closer to parallel. For three of four text corpora, the improvements are substantial, yielding median differential pre-trends 25-30% closer to zero. The second column (RMSE) similarly shows improvements of 10-15% in terms of root mean squared error. The final column indicates that these aggregate improvements are not solely driven by outliers: embedding-weighting improves (reduces) differential pre-trends for the majority of individual event study estimates.

Figure 1 shows that the *distributions* of (absolute) differential pre-trends due to the

[4]These estimates are work in progress. For some studies, the text includes post-treatment information. While this leakage is a concern for post-treatment effect estimates, it does not mechanically bias the pre-treatment estimates of "effects" that are the focus of this section. Estimates restricted to pre-treatment text, where available, are in progress.

Table 1: Summary statistics of pre-trends across applications.

	relative		embedding
embedding corpus	MAE	RMSE	improves
American Stories	71%	89%	64%
Congressional Speech	71%	86%	68%
Form 10-K filings	88%	89%	56%
Wikipedia	75%	87%	62%

Note: The MAE and RMSE columns show the (weighted) median absolute error and (weighted) root mean squared error of event study pre-trends for a given embedding corpus relative to the same aggregate statistic for standard event study analyses. The column "embedding better" shows the (weighted) fraction of event study estimates for which the embedding estimate is closer to zero than the standard estimate.

embedding-weighted estimators are preferable to the distribution due to the standard estimator. Each panel plots the CDF of absolute values of pre-trends due to a different embedding against the CDF due to the standard estimator, which is the same across panels. The distribution due to the standard estimator nearly first-order stochastically dominates each of the distributions due to embeddings-weighted estimators. That is, for nearly any (absolute) value of pre-treatment parallel trends violation, we observe more estimates exceeding this level of violation for the the standard estimator than for any of the embeddings-weighted estimators.

6 Conclusion

We propose a difference-in-differences estimator that uses weights that balance text embeddings and a weighted parallel trends assumption. As such, the estimator is applicable when units are not anonymous and unit identifiers appear in text. Prior research in natural language processing has demonstrated that the relative locations of embedding vectors in Euclidean space reflect certain semantic relationships. We motivate our estimator intuitively with the conjecture that weights that approximately balance embeddings therefore adjust for differences in otherwise latent concepts between units. Empirically, we demonstrate that the estimator yields event study estimates with smaller differential pre-trends than standard estimators. Our estimator offers relatively few researcher degrees of freedom: The only new "tuning parameters" are chosen prior to training text embeddings, which is computationally intensive and

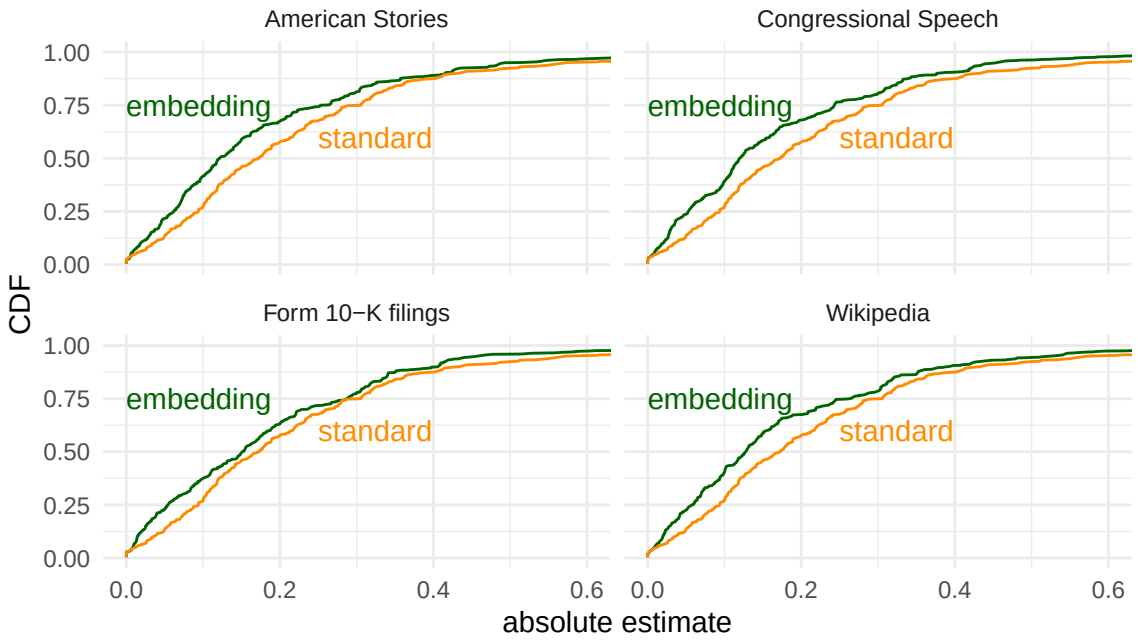


Figure 1: Cumulative distribution functions of (absolute) differential pre-trend estimates for embedding-weighted and standard event studies.

hence relatively resistant to target specification search. Importantly, our weighting approach is not directly based on pre-treatment outcomes. Hence, estimates of pre-trends retain their informativeness about the validity of the underlying parallel trends assumption. Our weight construction also enables a novel approach to variance estimation, circumventing familiar challenges due to estimated weights in the synthetic control literature.

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies 72*(1), 1–19.

Abadie, A., A. Diamond, and J. 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), 493–505.

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118.

Ash, E., D. L. Chen, and S. Naidu (2025). Ideas Have Consequences: The Impact of Law and Economics on American Justice. *The Quarterly Journal of Economics 141*(1), 845–887.

Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics 15*, 659–688.

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2020). The Evolution of Work in the United States. *American Economic Journal: Applied Economics 12*(2), 1–34.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(4), 597–623.

Bellégo, C., D. Benatia, and V. Dortet-Bernadet (2025). The chained difference-in-differences. *Journal of Econometrics 248*, 105783.

Ben-Michael, E., A. Feller, D. A. Hirshberg, and J. R. Zubizarreta (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*.

Bruns-Smith, D., O. Dukes, A. Feller, and E. L. Ogburn (2025). Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Caetano, C. and B. Callaway (2024). Difference-in-Differences when Parallel Trends Holds Conditional on Covariates. *arXiv preprint arXiv:2406.15288*.

Chattopadhyay, A. and J. R. Zubizarreta (2023). On the implied weights of linear regression for causal inference. *Biometrika 110*(3), 615–629.

Dell, M., J. Carlson, T. Bryan, E. Silcock, A. Arora, Z. Shen, L. D'Amico-Wong, Q. Le, P. Querubin, and L. Heldring (2023). American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers. *arXiv preprint arXiv:2308.12477*.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2018, January). Congressional Record for the 43rd–114th Congresses: Parsed Speeches and Phrase Counts. Dataset.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica 87*(4), 1307–1340.

Goldberg, Y. and O. Levy (2014). word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Grimmer, J., M. E. Roberts, and B. M. Stewart (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Han, S., E. H. Schulman, K. Grauman, and S. Ramakrishnan (2021). Shapes as product differentiation: Neural network embedding in the analysis of markets for fonts. *arXiv preprint arXiv:2107.02739*.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5), 1017.

Hirshberg, D. A. and J. R. Zubizarreta (2017). On two approaches to weighting in causal inference. *Epidemiology 28*(6), 812–816.

Jiang, Z. and D. Kim (2024). The Effects of TV Advertising and Ad Content on Consumer Financial Decisions: Evidence from Mortgage Refinancing. *SSRN*.

Kellogg, M., M. Mogstad, G. A. Pouliot, and A. Torgovitsky (2021). Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association 116*(536), 1804–1816.

Kim, D., Z. Jiang, and R. Thomadsen (2023). TV Advertising Effectiveness with Racial Minority Representation: Evidence from the Mortgage Market. *SSRN*.

Liu, Y. (2025). Synthetic Parallel Trends. *arXiv preprint arXiv:2511.05870*.

Ludwig, J., S. Mullainathan, and A. Rambachan (2024). Large language models: An applied econometric framework. *Annual Review of Economics 18*, 283–316.

Magnolfi, L., J. McClure, and A. Sorensen (2025). Triplet Embeddings for Demand Estimation. *American Economic Journal: Microeconomics 17*(1), 282–307.

McDonald, B. (2022). 10x stage one parsing documentation. Software Repository for Accounting and Finance, University of Notre Dame. Updated January 2022; accessed July 6, 2026.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pagliardini, M., P. Gupta, and M. Jaggi (2018). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology 49*(12), 1373–1379.

Pennington, J., R. Socher, and C. D. Manning (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Sant'Anna, P. H. C. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–122.

Sarkar, S. and K. Vafa (2024). Lookahead Bias in Pretrained Language Models. *SSRN*.

Słoczyński, T., D. Uysal, and J. M. Wooldridge (2023). Covariate balancing and the equivalence of weighting and doubly robust estimators of average treatment effects.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.

A Details on Embeddings

A.1 Training

We train GloVe on four distinct, publicly available, corpora: i) American Stories, ii) Congressional Speech, iii) Wikipedia, and iv) Form 10-K filings by publicly-traded

firms in the U.S.

The American Stories dataset is a collection of nearly 20 million U.S. newspaper articles from the years 1770 to 1964 constructed by Dell et al. (2023). The dataset was constructed using deep learning automated reading on newspaper scans from the public domain Chronicling America collection maintained by the United States Library of Congress.

We obtained congressional speeches from the parsed speeches and phrase counts from the congressional record for the 43rd-114th Congresses dataset constructed by Gentzkow et al. (2018). This dataset contains processed text from the bound and daily editions of the United States Congressional Record as provided by HeinOnline (`https://heinonline.com/content/u-s-congressional-documents/`). The bound edition covers the 43rd to 111th Congresses, and the daily edition covers the 97th to 114th. Both versions include all text spoken on the floor of each chamber of Congress: the United States House of Representatives and the United States Senate during the covered Congresses.

Wikipedia is a free online encyclopedia written and maintained by volunteers through open collaboration and hosted by the American nonprofit Wikimedia Foundation. Wikipedia is the largest and most read reference work in history and one of the world's most visited websites.In the original GloVe paper, Pennington et al. (2014) trained the GloVe model on a combination of all Wikipedia articles written in the English language plus Gigaword (a comprehensive archive of newswire text data). Additionally, we train a second model using a collection of English-language Wikipedia articles.

The U.S. Securities and Exchange Commission requires all companies that are either publicly traded, or have more than 10 million $USD in assets to file annual (10-K) and quarterly (10-Q) reports on their financial performance. The filings are divided into parts in which companies provide information on their history, structure, properties, legal proceedings, operations, and financial performance among others. We obtained the full 10-K and 10-Q filings text from the years 1993 to 2024 from the Notre Dame Software Repository for Accounting and Finance (McDonald, 2022).

We train multiple GloVe models on each corpus, one version that uses all available text, and different versions tailored to the different empirical applications we use to illustrate the method. All versions were trained using the publicly available code included with the original GloVe paper (Pennington et al., 2014).

The GloVe model and training have tuning parameters that must be chosen by the researcher The most important tuning parameters are the dimension of the embeddings, and the size of the context window. Additionally, training requires specifying a maximum number of times the training algorithm will go over the entire co-occurrence matrix. We conducted a series of tests assessing the robustness of the method to these parameters and found that the bias and mean squared error of the difference-in-differences estimators in placebo analyses slightly improves with the the choice of dimension and is robust to the window size. All the results we present use embeddings of dimension 200 and a context window of 25 words. Finally, for all corpora we find that the value of the loss function stabilizes before 200 iterations, which is the maximum number of iterations we allow.

Before training the models, we transform all our text data to lower case, and tokenize the corpora. Tokenization consists on splitting a corpus into smaller units called *tokens*. Depending on the model, tokens can represent phrases, words, or even parts of words. In the case of GloVe, each token represents a unique word, number, symbol (such as an emoji), or special character. As recommended by the GloVe authors, we use the tokenizer from the Stanford Parsers. More specifically, from the version 4.2.0 of the Stanford Parser.[5]

All the empirical applications we consider have either U.S. states or countries as the unit of analysis, We ensure that we construct embeddings for all such units that appear in our corpora. One challenge is that some states or countries have names comprised by two words (for instance, *North Carolina*) or might appear under different names (for instance, *USA*, *United States of America*, and *United States*). For these cases, we manually transform all mutliple-words names into single tokens by finding all such appearances and removing empty spaces (for instance, *North Carolina* is transformed to *northcarolina*). For each entities with more than one name, we create a list of different strings, words, or phrases referring to the same entity. We then replace all occurrences of strings in the list for a given entity with a single unique token (for instance, *USA*, *United States*, and *United States of America* are transformed to *unitedstates*).

Finally, to reduce the computational cost of training the model, we eliminate all words that appear less than 35 times in the training corpus. This step guarantees

[5]Specifically, we use version 4.2.0. See `https://nlp.stanford.edu/software/lex-parser.shtml` for details.

that almost all recurrent words (including infrequent ones) are used for training and have an embedding or vector representation. At the same time, filtering eliminates typos and highly atypical tokens such as specific numbers or special characters. This filtering reduces the computational cost of training the model while retaining most of the information contained in the corpus. If a U.S. state or country appears less than 35 times in some corpus, we do not eliminate it from the training data. For entities that do not appear directly in a given corpus, we manually match it to the conceptually closest token in the corpus.

A.2 Analogy Tasks

GloVe is designed such that the locations of embedding vectors in Euclidean space reflect their semantic relationships. Specifically, the direction, in vector space, that results from differencing the embeddings of two words corresponds to the semantic relation that their juxtaposition implies. Thus, word couplings in which the relation between the first and second word is approximately the same, as in the `man - woman` and `king - queen` example, tend to point in a similar direction. This direction, implicitly, represents the common relation between the word pairs (male–female). We provide evidence that our own GloVe models trained on different corpora satisfy this property across different semantic relations.

To asses the quality of our embeddings trained on different corpora we use the evaluation data from Pennington et al. (2014). The evaluation data consist of 550 cases of word pairs constructed from 13 different semantic relations. The semantic relations from which the evaluation pairs are constructed are: 116 `capital::country`, 68 `city::state`, 41 `country::nationality`, 30 `country::currency`, 23 `male::female`, 32 `adjective::adverb`, 29 `adjective::antonym`, 37 `adjective::comparative`, 34 `adjective::superlative`, 33 `verb::present participle`, 40 `verb::past tense`, 37 `singular::plural`, and 30 `singular verb::plural verb` pairs.

Pennington et al. (2014) measure vector similarity using *cosine similarity*. This measure calculates the cosine of the angle θ that two vectors form in Euclidian space. Cosine similarity is equal to 1 if both vectors point in the same direction, equal to 0 if the vectors are orthogonal, and equal to -1 if the two vectors point in exact opposite

directions. Formally, cosine similarity is defined as

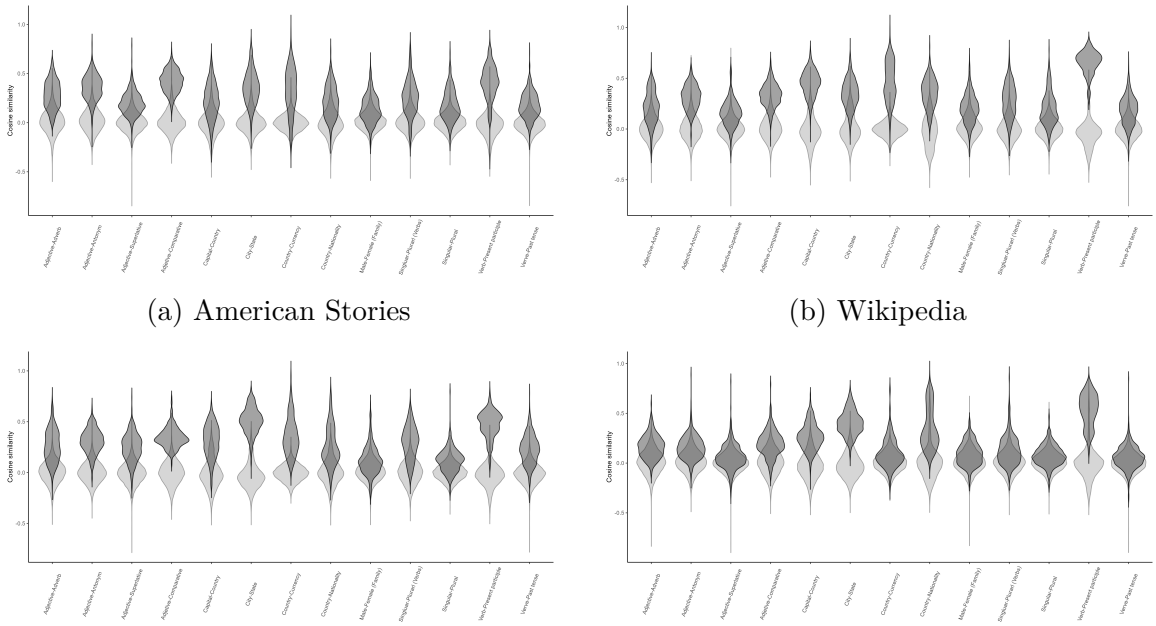
$$S_{\cos}(\boldsymbol{u}, \boldsymbol{v}) := \frac{\boldsymbol{u}'\boldsymbol{v}}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2} = \cos(\theta) \tag{26}$$

where, for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{u}'\boldsymbol{v} = \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2 \cos(\theta)$ is the inner product of $\boldsymbol{u}$ and $\boldsymbol{v}$, and θ is the angle between the two vectors.

For every training corpus, we identify all the word pairs from the evaluation data for which our implementation of GloVe produced an embedding and calculate the difference. For example, if the words `queen` and `king` appear in American Stories, we compute the element-wise difference between the vector representations associated with both words. The result is a list of word pairs and each within-pair difference in embedding space. When combined into all possible differences, the 550 word pairs from the training data result in 302,500 distinct combinations. Of these combinations, we were able to represent 236,196 when training the model using American Stories, 148,996 using Hein Daily, 257,049 using Hein Bound, 292,681 using Wikipedia, and 199,809 using form 10-K filings. After creating all possible combinations of word pairs, we compute the cosine similarity of their difference. If the model correctly captures a semantic relationship, then embedding differences associated with pairs constructing from that same semantic relation should have a high cosine similarity.

For interpretability, we contrast the cosine similarity within semantic relationship against cosine similarity across semantic relationships. In particular for high-dimensional vectors, it is challenging to interpret the level of cosine similarity. Hence, we asses the similarity of different directions in relative terms If the n-th semantic relation is accurately captured by the embeddings, then differences associated with pairs constructed from the same semantic relation should be higher than for pairs obtained from distinct semantic relations.

Figure 2 shows the distribution of cosine similarity across corpora and evaluation pairs. Each panel corresponds to a different corpus. Within each panel, we plot two times 13 violin plots, one per semantic relation. The violin plots show the distribution of the cosine similarity of the embeddings representations. Within each of the 13 pairs of distributions, the dark gray distribution shows cosine similarity when both difference-pairs correspond to the focal semantic relation. The light gray distribution shows cosine similarity when one element of the difference-pairs corresponds to the focal relation, while the other element corresponds to one of the 12 other semantic



(a) American Stories

(b) Wikipedia

(c) Congressional Speech

(d) Form 10-K filings

Figure 2: Distribution of cosine similarity within and across semantic relations.

relations.

In ongoing work, we compute these distributions for the pre-trained embeddings of Pennington et al. (2014), as well as the the summary statistics reported therein. We also compute summary statistics for the reverse analogy tasks described in the main text.

B Proofs

B.1 Proof of Theorem 1

B.2 Proof of Theorem 2

Proof. We first provide an expression for the conditional variance of the estimator $\hat{\tau}_t$. Then we derive the conditional expectation of the variance estimator $\hat{\sigma}^2$.

By Assumption 2(a), outcomes are conditionally uncorrelated across units. By Assumption 2(b), the conditional variance of i's trend depends only on i's own

treatment path and embeddings. Hence,

$$\begin{aligned}
\operatorname{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E}) &= \sum_{i:X_i=\vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \operatorname{var}(Y_{i,t} - Y_{i,t^*-1} \mid \boldsymbol{X}, \boldsymbol{E}) \\
&\quad + \sum_{j:X_j=\vec{0}} \bar{w}_j^2 \operatorname{var}(Y_{j,t} - Y_{j,t^*-1} \mid \boldsymbol{X}, \boldsymbol{E}) \\
&= \sum_{i:X_i=\vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \operatorname{var}(Y_{i,t} - Y_{i,t^*-1} \mid X_i, E_i) \\
&\quad + \sum_{j:X_j=\vec{0}} \bar{w}_j^2 \operatorname{var}(Y_{j,t} - Y_{j,t^*-1} \mid X_j, E_j) \\
&= \sum_{i:X_i=\vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \sigma_i^2 + \sum_{j:X_j=\vec{0}} \bar{w}_j^2 \sigma_j^2
\end{aligned} \tag{27}$$

Next, consider the variance estimators $\hat{\sigma}_i^2$. For ease of notation, define $\Delta_i := Y_{i,t} - Y_{i,t^*-1}$. Add and subtract $\mu_i - \mu_{m(i)}$, and expand the square to get

$$\begin{aligned}
\hat{\sigma}_i^2 &= \frac{1}{2}\Big((\Delta_i - \mu_i) - (\Delta_{m(i)} - \mu_{m(i)}) + (\mu_i - \mu_{m(i)})\Big)^2 \\
&= \frac{1}{2}\Big((\Delta_i - \mu_i)^2 + (\Delta_{m(i)} - \mu_{m(i)})^2 + (\mu_i - \mu_{m(i)})^2 \\
&\qquad - 2(\Delta_i - \mu_i)(\Delta_{m(i)} - \mu_{m(i)}) \\
&\qquad + 2(\Delta_i - \mu_i)(\mu_i - \mu_{m(i)}) - 2(\Delta_{m(i)} - \mu_{m(i)})(\mu_i - \mu_{m(i)})\Big)
\end{aligned} \tag{28}$$

By Assumption 2(b) and the definitions of μ_i and σ_i^2,

$$\mathrm{E}((\Delta_i - \mu_i)^2 \mid \boldsymbol{X}, \boldsymbol{E}) = \mathrm{E}((\Delta_i - \mu_i)^2 \mid X_i, E_i) = \sigma_i^2 \tag{29}$$

By Assumption 2(b), $\mu_i = \mathrm{E}(\Delta_i \mid \boldsymbol{X}, \boldsymbol{E})$. Hence, by Assumption 2(a),

$$\mathrm{E}\Big((\Delta_i - \mu_i)(\Delta_{m(i)} - \mu_{m(i)}) \mid \boldsymbol{X}, \boldsymbol{E}\Big) = \operatorname{cov}(\Delta_i, \Delta_{m(i)} \mid \boldsymbol{X}, \boldsymbol{E}) = 0 \tag{30}$$

Finally, by Assumption 2(b) and the definition of μ_i,

$$\begin{aligned}
\mathrm{E}\Big((\Delta_i - \mu_i)(\mu_i - \mu_{m(i)}) \mid \boldsymbol{X}, \boldsymbol{E}\Big) &= \mathrm{E}\Big((\Delta_i - \mu_i) \mid \boldsymbol{X}, \boldsymbol{E}\Big)(\mu_i - \mu_{m(i)}) \\
&= \mathrm{E}\Big((\Delta_i - \mu_i) \mid X_i, E_i\Big)(\mu_i - \mu_{m(i)}) \\
&= 0.
\end{aligned} \tag{31}$$

Applying and combining these results for i and $m(i)$,

$$\begin{aligned}\mathrm{E}(\hat{\sigma}_i^2 \mid \boldsymbol{X}, \boldsymbol{E}) &= \frac{1}{2}(\sigma_i^2 + \sigma_{m(i)}^2 + (\mu_i - \mu_{m(i)})^2) \\ &= \sigma_i^2 + \frac{1}{2}(\mu_i - \mu_{m(i)})^2 - \frac{1}{2}(\sigma_i^2 - \sigma_{m(i)}^2).\end{aligned} \tag{32}$$

Hence,

$$\begin{aligned}\mathrm{E}(\hat{\sigma}^2 \mid \boldsymbol{X}, \boldsymbol{E}) = &\sum_{i:\, X_i = \vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \sigma_i^2 + \sum_{j:\, X_j = \vec{0},\, \bar{w}_j > 0} \bar{w}_j^2 \sigma_j^2 \\ &+ \frac{1}{2} \sum_{i:\, X_i = \vec{1}_{t^*}} \Big(\frac{1}{n_{t^*}}\Big)^2 \Big((\mu_i - \mu_{m(i)})^2 - (\sigma_i^2 - \sigma_{m(i)}^2) \Big) \\ &+ \frac{1}{2} \sum_{j:\, X_j = \vec{0},\, \bar{w}_j > 0} \bar{w}_j^2 \Big((\mu_j - \mu_{m(j)})^2 - (\sigma_j^2 - \sigma_{m(j)}^2) \Big)\end{aligned} \tag{33}$$

The first line of the right-hand side equals $\mathrm{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E})$, because untreated units with $\bar{w}_j = 0$ do not contribute to the conditional variance. The sum of the second and third line is non-negative by Assumption 3. Hence, $\mathrm{E}(\hat{\sigma}^2 \mid \boldsymbol{X}, \boldsymbol{E}) \geq \mathrm{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E})$.

Finally, consider $2\hat{\sigma}^2$. Then, even without Assumption 3,

$$2\,\mathrm{E}(\hat{\sigma}_i^2 \mid \boldsymbol{X}, \boldsymbol{E}) = \sigma_i^2 + \sigma_{m(i)}^2 + (\mu_i - \mu_{m(i)})^2 \geq \sigma_i^2, \tag{34}$$

Aggregating this inequality over the units in $\mathcal{I}_+$ gives $\mathrm{E}(2\hat{\sigma}^2 \mid \boldsymbol{X}, \boldsymbol{E}) \geq \mathrm{var}(\hat{\tau}_t \mid \boldsymbol{X}, \boldsymbol{E})$ as claimed. □

C Details on Empirical Applications