# When Can We Work in Embedding Space? What Text Embeddings Preserve

Simon Freyaldenhoven

Federal Reserve Bank of Philadelphia*

August 26, 2026

# Online Appendix

*Email: simon.freyaldenhoven@phil.frb.org

# List of Tables



# List of Figures

# A Other Word-Embedding Algorithms under the Topic Model

Section 2.3 of the main paper states what other standard embedding algorithms target and what those targets recover. This appendix supplies the background. Section A.1 writes the four objectives in a common form. Sections A.2–A.5 derive the specific target of this algorithm by algorithm. Section A.6 then shows what the embedding inherits from the topic model, using only that every pair of words co-occurs.

## A.1 The algorithms as weighted low-rank fits

In the word embedding literature, $log(R)$ is called the pointwise mutual information $\mathrm{PMI}(v,u) = \log R_{vu}$ [Levy and Goldberg, 2014], and I follow this convention here. Set

$$R_{\min} := \min_{v,u} R_{vu}, \qquad R_{\max} := \max_{v,u} R_{vu}, \qquad \kappa_R := R_{\max}/R_{\min}. \tag{OA.1}$$

Each of the four algorithms scores the pair $(v,u)$ by an inner product $X_{vu} = w_v^\top \tilde{w}_u$ and fits that score to the data by minimizing a loss. Three ingredients pin the loss down: a *response* $t_{vu}$ read off the corpus, a convex generator $\phi$, and weights $\omega_{vu} > 0$ on the pairs. Response and score are different kinds of object. The response is data, fixed once the corpus is given; the score is the parameter, constrained to rank $r$. Lemma A.1 then shows that the weights drop out of the population optimum, so what the score converges to is determined by $t_{vu}$ and $\phi$ alone.

Take the generator first. Fix a strictly convex, differentiable $\phi$ on an interval $\mathcal{I}$ and set

$$D_\phi(t,y) \;:=\; \phi(t) - \phi(y) - \phi'(y)(t-y), \qquad t,y \in \mathcal{I}, \tag{OA.2}$$

the vertical gap at $t$ between $\phi$ and its tangent line at $y$ — the *Bregman divergence* generated by $\phi$ [Bregman, 1967] (Also see Collins et al. [2001], Udell et al. [2016]). Convexity makes $D_\phi \geq 0$ and strict convexity makes it vanish only at $t = y$, so $D_\phi(t,\cdot)$ measures discrepancy from $t$; it is not symmetric and is not a metric. Every loss below is a $D_\phi$ for one of three generators,

$$\phi(y) = \tfrac{1}{2}y^2, \qquad \phi(y) = \textstyle\sum_v y_v \log y_v, \qquad \phi(y) = y\log y + (1-y)\log(1-y),$$

which give squared error, the Kullback–Leibler divergence, and the Bernoulli log-loss respectively.

The previous paragraph gave us a discrepancy, the Bregman divergence. But there is a scale mismatch: the score $X_{vu}$ lives in $\mathbb{R}$. On the other hand, the domain of the response $t_{vu}$, $\mathcal{I}$, depends on the algorithm, but generally $\mathbb{R} \neq \mathcal{I}$ (cf. Table 1). The derivative $\phi'$ is the *link* that maps $\mathbb{R}$ onto the response scale: Strict convexity makes $\phi'$ strictly increasing and hence invertible, so $(\phi')^{-1}(X_{vu})$ is the score expressed on the response's scale — the fitted value for $t_{vu}$. A loss of the

form $D_\phi\big(t_{vu},\ (\phi')^{-1}(X_{vu})\big)$ therefore compares response with fitted value, and I call the matrix with entries $\phi'(t_{vu})$ the algorithm's *target*.

**Lemma A.1** (Weights do not move the target)**.** *Let $\omega_{vu} > 0$ and $t_{vu} \in \operatorname{int}\mathcal{I}$ for all $v, u$. Then, the unconstrained minimizer of $\sum_{v,u}\omega_{vu}D_\phi\big(t_{vu},\ (\phi')^{-1}(X_{vu})\big)$ over $X \in \mathbb{R}^{V\times V}$ is $X_{vu} = \phi'(t_{vu})$, entrywise and independently of the weights.*

*Proof.* Each term is nonnegative and vanishes iff $(\phi')^{-1}(X_{vu}) = t_{vu}$; $\phi'$ is strictly increasing, hence invertible, so this holds iff $X_{vu} = \phi'(t_{vu})$. The weights are positive, so minimizing the sum minimizes each term. □

Lemma A.1 states that the score equals the target whenever the rank constraint is non-binding.

## A.2 Full-softmax skip-gram

The population full-softmax skip-gram objective is

$$\mathcal{L}_{\text{full}}(W, \tilde{W}) \ := \sum_{v,u\in[V]} M_{vu}\ \log\frac{\exp(w_v^\top \tilde{w}_u)}{\sum_{v'\in[V]}\exp(w_{v'}^\top \tilde{w}_u)}, \tag{OA.3}$$

where $w_v$ is the $v$-th row of $W$ and $\tilde{w}_u$ is the $u$-th row of $\tilde{W}$, so that $X := W\tilde{W}^\top$ collects the scores $X_{vu} = w_v^\top \tilde{w}_u$. Equation (OA.3) is the population counterpart of the skip-gram model of Mikolov et al. [2013c, eqs. 1–2], whose objective averages$\log P(w \mid c)$ over corpus positions within a context window; here that average is replaced by the population co-occurrence $M$.

**Lemma A.2** (Full-softmax skip-gram target)**.** *Under Assumptions 1 and 2, the unconstrained maximizers of* (OA.3) *satisfy $X^* = \log R + \log q\,\mathbf{1}^\top - \mathbf{1}g^\top$ for some $g \in \mathbb{R}^V$.*

*Proof.* Lemma A.1 does not apply directly: the softmax in (OA.3) normalizes over the whole column $X_{\bullet u}$, so the loss is not separable across $v$ within a column, and we need to consider entire columns instead. Writing $M_{vu} = q_u P(w{=}v \mid c{=}u)$ and $s_{vu} := \operatorname{softmax}(X_{\bullet u})_v$,

$$\mathcal{L}_{\text{full}} \ = \ -\sum_u q_u\,\mathrm{KL}\big(P(w \mid c{=}u)\,\big\|\,s_{\bullet u}\big) \ - \ \sum_u q_u H\big(P(w \mid c{=}u)\big), \tag{OA.4}$$

whose second term does not involve $X$. So the objective is again a weighted Bregman fit, with response $t_{vu} = P(w{=}v \mid c{=}u)$, the entropy generator, and weight $\omega_{vu} = q_u$; what differs from Lemma A.1 is only that the divergence is taken one column at a time rather than one entry at a time. It is maximized at $s_{\bullet u} = P(w \mid c{=}u)$ for every $u$, that is at $X_{vu} = \log P(w{=}v \mid c{=}u) + g_u$, the column gauge $g$ being free because $\operatorname{softmax}$ is invariant to adding a constant to a column. Finally $\log P(w{=}v \mid c{=}u) = \log R_{vu} + \log q_v$ by definition of $R$. □

## A.3 SGNS

Computing the population full-softmax skip-gram objective is computationally difficult. Mikolov et al. [2013c] therefore introduce negative sampling as an approximation used in practice, which I consider next.

Summing the pair-specific objective of Levy and Goldberg [2014, eq. 5] over $(v, u)$ and normalizing counts by corpus size gives the population SGNS objective

$$\mathcal{L}_{\text{pop}}(W, \tilde{W}) \;:=\; \sum_{v,u\in[V]} \Big\{ M_{vu}\, \log \sigma(w_v^\top \tilde{w}_u) \;+\; \nu\, q_v q_u\, \log \sigma(-w_v^\top \tilde{w}_u) \Big\}, \tag{OA.5}$$

with $\sigma(x) := 1/(1+e^{-x})$, $\nu \geq 1$ the negative-sampling rate, and $w_v, \tilde{w}_u$ as above.[1]

**Lemma A.3** (SGNS target)**.** *Under Assumptions 1 and 2,* (OA.5) *is a weighted Bernoulli-likelihood fit with weights $\omega_{vu} = q_v q_u (R_{vu} + \nu)$ and responses $t_{vu} = R_{vu}/(R_{vu} + \nu)$, and its unconstrained maximizers satisfy $X^* = \log R - \log \nu\, \mathbf{1}_V \mathbf{1}_V^\top$.*

*Proof.* With $\omega_{vu} := M_{vu} + \nu q_v q_u = q_v q_u (R_{vu} + \nu)$ and $t_{vu} := M_{vu}/\omega_{vu} = R_{vu}/(R_{vu} + \nu)$,

$$\begin{aligned} \mathcal{L}_{\text{pop}}(W, \tilde{W}) \;&=\; \sum_{v,u} \omega_{vu} \Big\{ t_{vu} \log \sigma(X_{vu}) + (1 - t_{vu}) \log \sigma(-X_{vu}) \Big\} \\ &=\; -\sum_{v,u} \omega_{vu} D_\phi\big(t_{vu}, \sigma(X_{vu})\big) \;+\; \text{const}, \end{aligned}$$

where $X := W\tilde{W}^\top$ and $\phi(y) = y \log y + (1-y)\log(1-y)$, for which (OA.2) is the Bernoulli Kullback–Leibler divergence and $\phi' = \text{logit} = \sigma^{-1}$. Assumption 2 puts $t_{vu}$ in $(0, 1)$, so Lemma A.1 applies and gives $X^*_{vu} = \text{logit}(t_{vu}) = \log(R_{vu}/\nu)$. □

The target is Levy and Goldberg [2014, eqs. 6–7]: they let the inner products vary freely, which is the step Lemma A.1 formalizes, and solve for $\vec{w} \cdot \vec{c} = \text{PMI}(w, c) - \log \nu$. The weighted-Bernoulli form is what I add. Levy and Goldberg [2014, Section 3.2] observe that the loss on a pair depends on its co-occurrence count and its expected negative-sample count, and conclude that SGNS performs a weighted matrix factorization; they do not identify the weight $\omega_{vu}$ and response $t_{vu}$ that make it one, and those are what place SGNS in the same frame as the other three algorithms in Table 1.

[1] Implementations draw negatives from a smoothed unigram distribution rather than from $q$: `gensim` uses $P_{0.75}(u) \propto q_u^{3/4}$ [Mikolov et al., 2013c, Levy et al., 2015]. This replaces $q_v q_u$ by $q_v P_{0.75}(u)$ in (OA.5) and shifts the target to $\log R_{vu} + \log\big(q_u / P_{0.75}(u)\big) - \log \nu$. The added term is constant down each column, so it cancels in the row differences that Proposition 2 uses (cf. Remark 4); I therefore work with the unsmoothed objective.

## A.4 GloVe

The GloVe model [Pennington et al., 2014] minimizes the population objective

$$\mathcal{L}_{\mathrm{GloVe}}(W, \tilde{W}, b, \tilde{b}) \;:=\; \sum_{v,u\in[V]} f(M_{vu}) \Big( w_v^\top \tilde{w}_u \;+\; b_v \;+\; \tilde{b}_u \;-\; \log M_{vu} \Big)^2, \tag{OA.6}$$

where $f : [0,\infty) \to [0,\infty)$ is positive on the support of $M$ and $b_v, \tilde{b}_u \in \mathbb{R}$ are learned per-word biases.

**Lemma A.4** (GloVe target)**.** *Under Assumptions 1 and 2 and with $f$ positive on* $\mathrm{supp}(M)$*, the unconstrained minimizers of* (OA.6) *satisfy $X_{vu} + b_v + \tilde{b}_u = \log M_{vu}$ entrywise. With $b_v^* = -\log q_v$ and $\tilde{b}_u^* = -\log q_u$ the implied score matrix is $X^* = \log R$.*

*Proof.* Apply Lemma A.1 with $\phi(y) = y^2/2$, response $\log M_{vu}$, and weights $f(M_{vu})$. Decomposing $\log M_{vu} = \log q_v + \log q_u + \log R_{vu}$ and absorbing the marginals into the biases leaves $X_{vu} = \log R_{vu}$. □

The biases do here what the gauge $g$ does in Lemma A.2: both absorb the marginal offsets that separate $\log M$ from $\log R$.

## A.5 CBOW

CBOW [Mikolov et al., 2013a] reverses the conditioning relative to skip-gram: it predicts the target word from the average of its surrounding context embeddings. The population objective is

$$\mathcal{L}_{\mathrm{CBOW}}(W, \tilde{W}) \;:=\; \sum_{w,\mathbf{u}} P(w, \mathbf{u}) \, \log \frac{\exp\big(h_w^\top \bar{h}_{\mathbf{u}}\big)}{\sum_{w'\in[V]} \exp\big(h_{w'}^\top \bar{h}_{\mathbf{u}}\big)}, \tag{OA.7}$$

where $\mathbf{u} = (u_1, \ldots, u_{2J})$ is the context window and $\bar{h}_{\mathbf{u}} := \frac{1}{2J} \sum_{j=1}^{2J} h_{u_j}$ is the average context embedding. At the population optimum the softmax form yields $h_w^\top \bar{h}_{\mathbf{u}} = \log P(w \mid \mathbf{u}) + \kappa(\mathbf{u})$ with $\kappa(\mathbf{u})$ the per-context log-normalizer.

With a single-word context this is Lemma A.2.. The same relabeling applies under negative sampling: with a one-word context the CBOW negative-sampling objective is (OA.5) with $W$ and $\tilde{W}$ exchanged, and $M$ is symmetric, so its target is again $\log R - \log \nu \, \mathbf{1}_V \mathbf{1}_V^\top$ and Proposition 2 applies verbatim. With more than one context word it does not: $h_w^\top \bar{h}_{\mathbf{u}}$ is a function of the context bundle rather than of pairs $(w, c)$, so there is no $V \times V$ target matrix analogous to those in Table 1, and a PMI factorization for CBOW analogous to Lemma A.3 remains open in the literature.

## A.6 What the targets inherit from the topic model

The proof of Proposition 2 bounds distances between rows of $\log R$ and then carries them to the embedding. That proposition is stated for the SGNS target, whose only offset is a scalar. The following lemma isolates its first step and states it more generally — for an arbitrary entrywise transform and arbitrary offsets constant along a row or a column. The row offsets are what it takes to cover the full-softmax and GloVe targets: both carry an offset constant along a row, and a row offset does not cancel in a row difference (Remark 4).

**Lemma A.5** (Index sufficiency)**.** *Let Assumptions 1 and 2 hold, let $\psi$ be any function on $(0, \infty)$, let $a, b \in \mathbb{R}^V$, and set $T_{vu} := \psi(R_{vu}) + a_v + b_u$. Then, for all $v, v' \in [V]$, $\tilde{B}_{v\bullet} = \tilde{B}_{v'\bullet}$ implies $T_{v\bullet} - T_{v'\bullet} = (a_v - a_{v'})\mathbf{1}^\top$.*

*Proof.* $R = \tilde{B} G \tilde{B}^\top$ (8) gives $R_{vu} = \tilde{B}_{v\bullet}^\top G\, \tilde{B}_{u\bullet}$, so $\tilde{B}_{v\bullet} = \tilde{B}_{v'\bullet}$ gives $R_{v\bullet} = R_{v'\bullet}$ and hence $\psi(R_{v\bullet}) = \psi(R_{v'\bullet})$ entrywise. The column offsets $b_u$ cancel in the difference. □

The hypothesis is the one used throughout: since $\tilde{B} = D_q^{-1} B$, $\tilde{B}_{v\bullet} = \tilde{B}_{v'\bullet}$ holds exactly when $B_{v\bullet} = \lambda B_{v'\bullet}$ for some $\lambda > 0$, in which case $q_v = \lambda q_{v'}$. Words with proportional loadings therefore have target rows that differ by the constant $a_v - a_{v'}$, in every one of the targets and whatever transformation produced them. Whether that constant is zero is what separates the algorithms: for SGNS $a = 0$, so the rows coincide and Proposition 2 carries this to the embedding; for full-softmax and GloVe $a \neq 0$, and Remark 4 works out what survives.

# B Additional Application Results

## B.1 Prompts and Generation Settings

Three separate calls to a language model appear in the application, and only the first affects the partitions. This subsection records all three, since the corpus is the one input to the exercise that a reader cannot inspect directly.

**Generating the descriptions.** Each of the 363 descriptions comes from a single call to `claude-opus-4-8` with a 1,500-token cap. The user message is

```
Summarize the economic development of {city} from 1979-2019 in around
500 words with a focus on the labor market.  Focus on broad trends
rather than specific numbers, and on aspects that make it unique
relative to the US as a whole.
```

and the system message is

```
Write a substantive, best-effort description using your general knowledge
of the place.  You do not need verified statistics or sources, and
you should not refuse or add disclaimers about lacking data or access
-- always produce a description, doing the best you can even when
you are uncertain.  Start right away with the description, without
an opening sentence like 'Here's a 500-word summary of Philadelphia's
economic development:'.
```

Four features of this are deliberate and worth stating.

First, "broad trends rather than specific numbers" is there to discourage invented statistics. The exercise needs qualitative economic structure, not a list of numbers, and the model is known to hallucinate statistics. Further, it will tend to recall numbers for locations it knows well and revert to broad trends for those it knows less well anyways, which would systematically change the structure of the description across locations.

Second, and relatedly, the refusal suppression is what makes the corpus complete. Without it the model declines or hedges for the smaller CBSAs it knows least about.

Third, "unique relative to the US as a whole" pushes the descriptions toward cross-sectional differentiation. A prompt asking simply for an economic history would return text dominated by features common to all US metros—national recessions, the general decline of manufacturing—which contribute a near-constant component to every document embedding and so cannot separate cities.

Fourth, the instruction to start immediately, without a preamble, matters for the same reason. A boilerplate opening clause repeated across 363 documents is a constant vector added to every average of word embeddings, which shrinks the relative distances the clustering depends on.

The resulting corpus is 363 descriptions of 476 to 510 words, median 495, so document length is close to constant.

**Labelling the clusters.** The cluster names in Section B.10 come from a second call, made once per cluster per method, to `claude-opus-5` at low reasoning effort with a 2,000-token cap. It is given the member city names and their descriptions and asked

> `Your response should include less than 60 words. Start right away to list the common themes. Cities: {cities}. Based on the following documents, what do these cities have in common? Identify the common themes: {documents}`

These labels are purely *ex post*. They are produced after the partitions are fixed, play no part in forming them, and enter no statistic reported anywhere in the paper—they exist only to interpret the clusters.

**The direct-LLM alternative.** The comparison in Section 4's footnote asks `claude-opus-5`, with a 4,000-token cap, to partition the cities itself rather than embed them: it receives all 363 descriptions at once and is asked to return exactly $L$ clusters as JSON, each with a name, a member list and a short explanation, with the instruction that every city belong to exactly one cluster. It did not comply at this scale. The returned partition covered only part of the sample and included place names absent from the input list, which is why the paper routes through embeddings instead.

## B.2 Application Embedding Implementation

This section specifies the five embedding methods compared in Section 4.

**Common to all five.** Every method starts from the same 363 descriptions and differs only in the map from a description to a vector. Text is lower-cased and tokenized on the regular expression `[A-Za-z]+(?:'[A-Za-z]+)?`, which keeps letter strings with internal apostrophes and drops digits and punctuation. The result is $183{,}633$ tokens and $7{,}411$ distinct word types, with documents between $485$ and $525$ tokens long (median $506$), so the averages below are taken over a near-constant number of terms. The four word-level methods embed words and then take the unweighted within-document average, which is the $\mu_d$ of Proposition 3; the fifth embeds the document directly. Each of the five is then clustered by $k$-means with $L = 5$, `k-means++` seeding [Arthur and Vassilvitskii, 2007] and $500$ restarts, keeping the best fit.

### B.2.1 Closed-Form SVD-$\beta$ Embedding

We directly estimate $\widehat{R}$ from the corpus word-context co-occurrence matrix (taking the full document as context), center to $\widehat{R} - \mathbf{1}_V \mathbf{1}_V^\top$, and take the top-$r$ eigendecomposition. The resulting $\widehat{\beta} = \Phi \Lambda^{1/2}$ satisfies $\widehat{\beta}\widehat{\beta}^\top = \widehat{R} - \mathbf{1}_V \mathbf{1}_V^\top$ on the top-$r$ eigenspace. Document embeddings are the within-document average of $\widehat{\beta}_v$. The eigenpairs are computed by Lanczos iteration.

**Vocabulary filter.** We restrict to words appearing in at least 5 documents, giving $V = 2{,}440$ types and $94.1\%$ token coverage. Rare words have very small marginal probabilities $P(w)$, inflating $1/P(w)$ in the centered ratio, making it computationally unstable, and dominating the resulting eigendecomposition.

**Choice of dimension.** Proposition 1 gives rank $K - 1$, and $K$ is not observable for the CBSA corpus. By Remark 1 we do not need it: any $r \geq K - 1$ recovers the same geometry in population, so the dimension only has to be chosen large enough. We set $r = 50$ as the primary specification. Section B.6 reports the sweep over $r \in \{10, 25, 50, 100, 200\}$ in place of the elbow the spectrum does not supply.

### B.2.2 Corpus-Trained Word2Vec: SGNS

Skip-gram with negative sampling using the `gensim` implementation [Řehůřek and Sojka, 2010]: $r = 50$ dimensions, `min_count` $= 2$, 5 epochs, and the full document as context, implemented as window $J = 10{,}000$, which exceeds every document length. `gensim` draws the effective half-window uniformly on $\{1, \ldots, \texttt{window}\}$ per token, so about $95\%$ of tokens see the entire document and the rest a narrower one. Section B.8 reports the sweep over $J \in \{5, 10, 20, 50\}$ The remaining hyperparameters are `gensim` defaults: $\nu = 5$ negative samples per update drawn from the unigram distribution raised to the $0.75$ power, frequent-word downsampling at threshold $10^{-3}$, and an initial learning rate of $0.025$ decayed linearly.

**Epochs.** The $5$ epochs are set against the $200$ of the CBOW arm (Section B.2.3) because a common count would not be equal training. CBOW makes one update per target token, while SGNS makes one per (target, context) pair, so at full-document context SGNS performs roughly $500$ times as many updates per epoch. Matching CBOW's $200$ epochs would give SGNS about $500$ times CBOW's total training rather than the same amount. At $5$ epochs SGNS still performs about $12$ times CBOW's total updates, so it is not the under-trained arm. At the narrower windows of Section B.8 the ratio is small enough that both objectives run $200$ epochs.

**Coverage.** After `min_count` $= 2$, the trained vocabulary contains $4{,}876$ word types covering $98.6\%$ of the $183{,}633$ tokens, against $86.3\%$ for the pretrained vectors of Section B.2.4, because the vocabulary is learned from the data rather than inherited. The trade-off is corpus size: $184$ thousand tokens is arguably small to train a Word2Vec model from scratch.

### B.2.3 Corpus-Trained Word2Vec: CBOW

**Settings.** $r = 50$ dimensions, `min_count` $= 2$, $200$ epochs, and the full document as context. Full-document context is implemented as window $J = 10{,}000$. The remaining hyperparameters are `gensim` defaults: $5$ negative samples per update drawn from the unigram distribution raised to the $0.75$ power, frequent-word downsampling at threshold $10^{-3}$, and an initial learning rate of $0.025$ decayed linearly.

### B.2.4 Pretrained Word2Vec: Google News

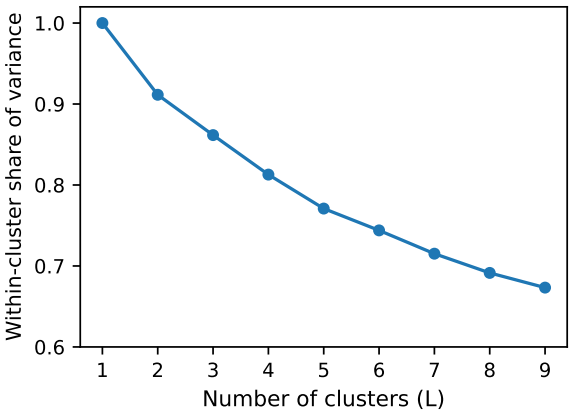
We average pretrained Google News Word2Vec vectors—$300$-dimensional, trained on part of a Google News corpus of about 100 billion words [Mikolov et al., 2013b] and distributed through `gensim`—over each CBSA description, yielding 363 embeddings in $\mathbb{R}^{300}$. These vectors cover $86.3\%$ of the corpus tokens; the remainder are dropped from the average.

### B.2.5 LLM Embedding-Based Clustering

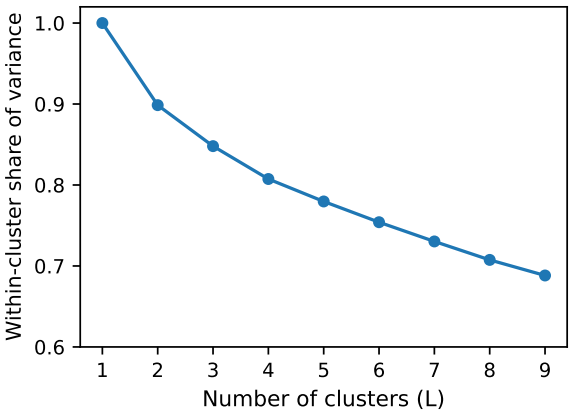
Using OpenAI's `text-embedding-3-large`, I embed each of the 363 descriptions into $\mathbb{R}^{3072}$. This is a contextual transformer rather than an average of word vectors, so none of the results of Section 2 technically apply to it; it enters as a benchmark for current practice.

## B.3 Choice of $L$ for the Application Embeddings

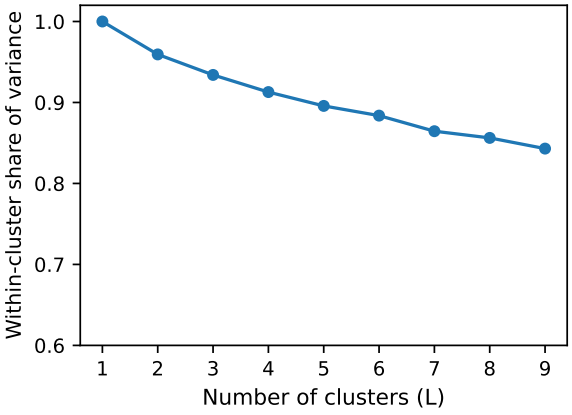
For each embedding method in Section 4, I use $L = 5$ clusters. Figure OA.1 reports the within-cluster sum of squares (WCSS) as a function of $L$ for all five methods. In each case the elbow is gradual, but $L = 5$ provides a reasonable trade-off. The absence of a sharp elbow is itself informative: it is what one expects when the document embeddings are spread over the simplex rather than concentrated at a small number of vertices, which is the regime Section 2 distinguishes from exact recovery. The vertical axis is the share of embedding variance within clusters.



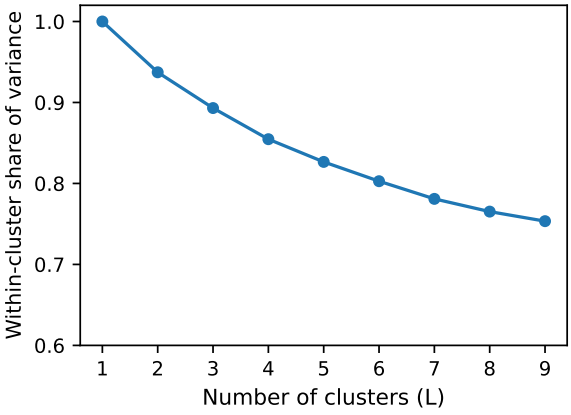

**(a)** SVD-$\beta$ ($r = 50$)



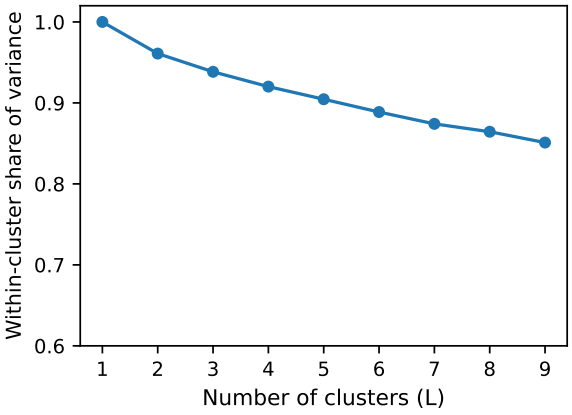

**(b)** Corpus SGNS




**(c)** Corpus CBOW



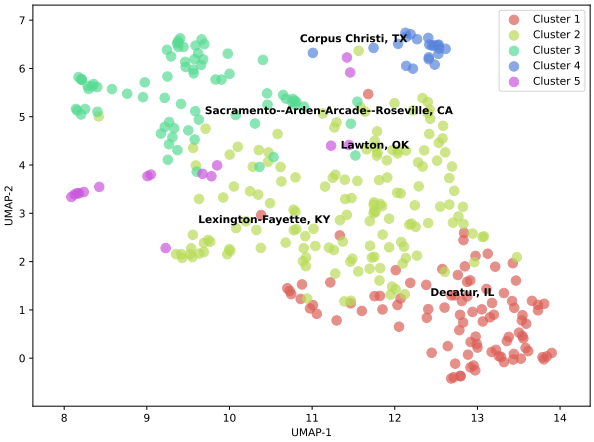

**(d)** Pretrained (Google News)



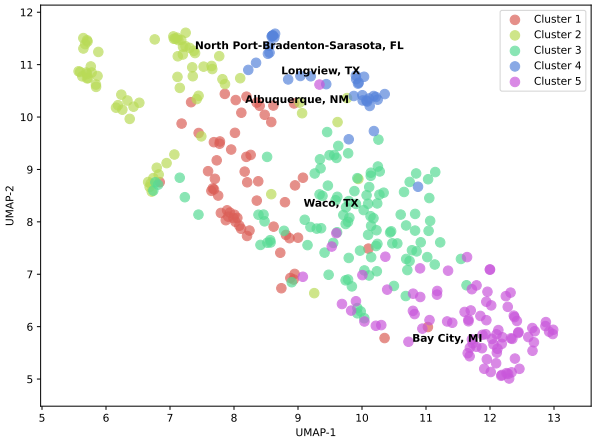

**(e)** LLM embeddings

**Online Appendix Figure OA.1:** Within-cluster sum of squares as a function of the number of clusters $L$ for each of the five embedding methods (363 CBSAs), normalized by the total sum of squares. For $k$-means the centroids are the cluster means, so $\text{TSS} = \text{WCSS} + \text{BCSS}$ holds exactly and the plotted quantity is the share of embedding variance still within clusters; one minus it is the share the partition explains.

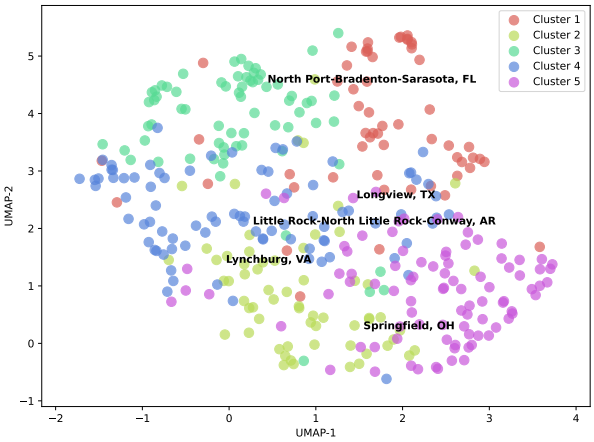
## B.4 Cluster Visualizations in UMAP Space



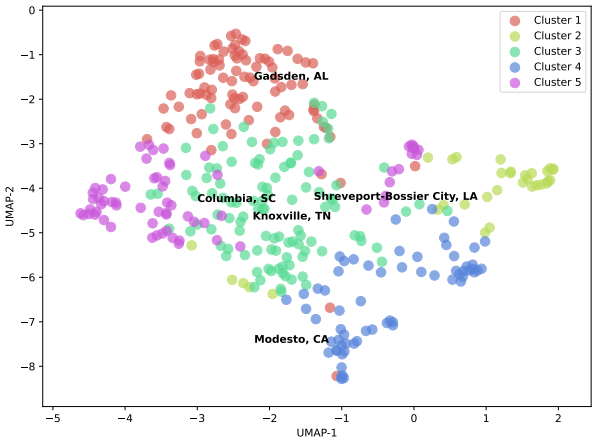

**(a)** SVD-$\beta$ ($r = 50$)



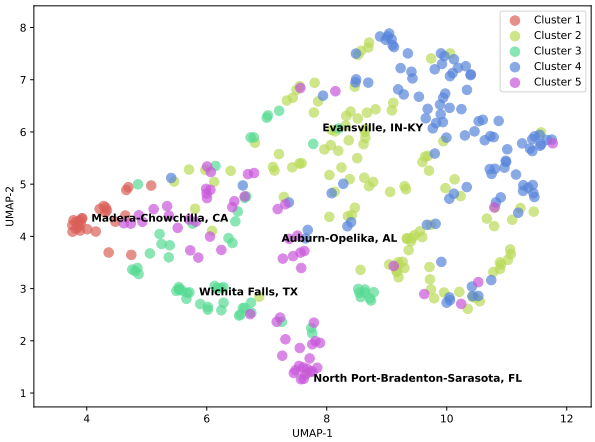

**(b)** Corpus SGNS




**(c)** Corpus CBOW




**(d)** Pretrained (Google News)




**(e)** LLM embeddings

**Online Appendix Figure OA.2:** Document embeddings projected to two dimensions with UMAP [McInnes et al., 2018]. Labels corresponds to the CBSA closest to that cluster's centroid in the full embedding space.

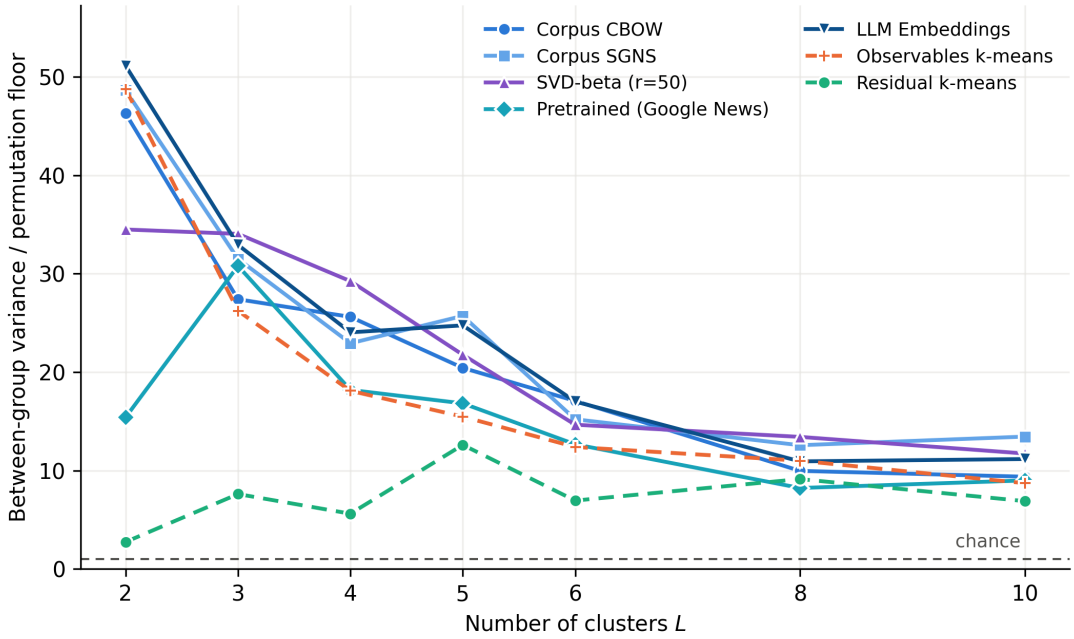
## B.5 Sensitivity to the Number of Clusters

Section 4 fixes $L = 5$. Figure OA.3 sweeps $L \in \{2, 3, 4, 5, 6, 8, 10\}$ for each method, recomputing the permutation floor at every $L$ so the ratios are comparable down a column.

First, every method stays well above its floor at every $L$, so the finding that these groupings separate employment dynamics more sharply than chance is not an artifact of the cluster count.

Second, a text method leads at every $L$: LLM embeddings at $L = 2$ and $L = 6$, SVD-$\beta$ at $L = 3$, 4, and 8, corpus CBOW at $L = 6$, corpus SGNS at $L = 5$ and 10. Residual $k$-means is below all five text methods at every $L$ except $L = 8$, and below all three corpus-based methods across the grid. Observables $k$-means tends to fall between residual $k$-means and the text-based methods.

We conclude that the ordering *among* the three corpus-based text methods is not stable across the grid. But the finding that text-based groupings separate employment dynamics better than our benchmarks does not depend on the cluster count.

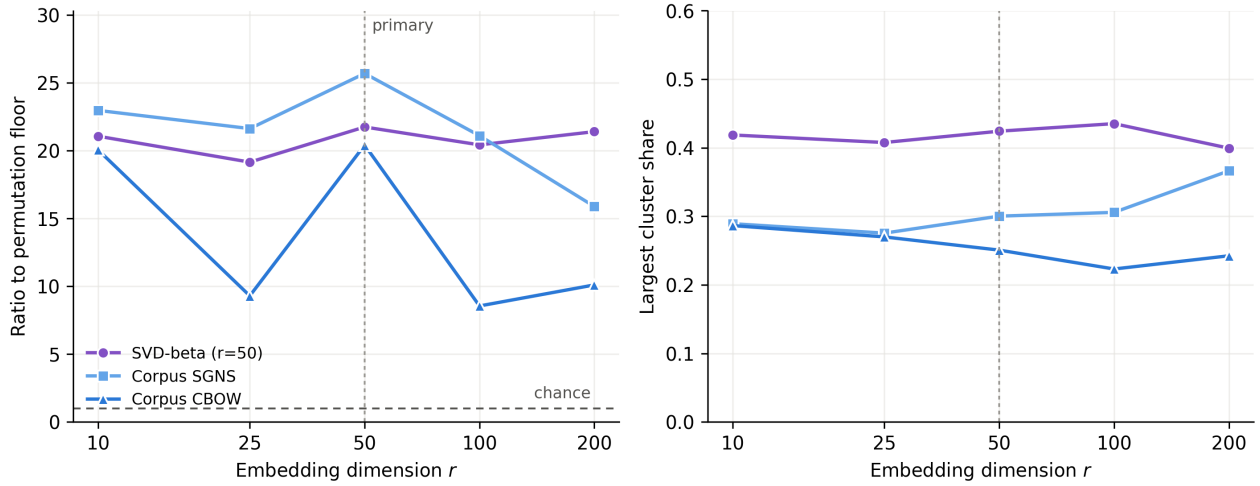



**Online Appendix Figure OA.3:** Between-group variance of unit-specific AR(2) impulse responses relative to the size-matched permutation floor, against the number of clusters $L$. The floor is recomputed at every $L$, so ratios are comparable along each line. Dashed lines represent the non-text benchmarks.

## B.6 Sensitivity to the Embedding Dimension

Remark 1 shows that any $r \geq K - 1$ recovers the same geometry in population, so the dimension needs only to be chosen large enough—which matters because $K$ is not observable for this

corpus. Figure OA.4 sweeps $r \in \{10, 25, 50, 100, 200\}$ for the three corpus-based methods at the full-document window and $L = 5$.




**Online Appendix Figure OA.4:** Sensitivity to the embedding dimension $r$ at $L = 5$, $p = 2$ and the full-document window, for the three corpus-based methods. Left: between-group variance relative to the size-matched permutation floor. Right: share of CBSAs in the largest cluster. SVD-$\beta$ and corpus SGNS are flat from $r = 10$ to $r = 200$, which is what justifies the $r = 50$ primary specification (dotted) without identifying $K$. Corpus CBOW alternates between two families of near-equivalent $k$-means partitions.

SVD-$\beta$ and corpus SGNS are flat across the entire range: neither ratio trends in $r$, and their partitions stay comparably balanced throughout. This is what Remark 1 predicts once $\widehat{R}$ is well estimated—coordinates beyond $K - 1$ contribute nothing in population and, at the full-document window, little in sample either. It justifies $r = 50$ without an estimate of $K$: any choice across this range would serve.

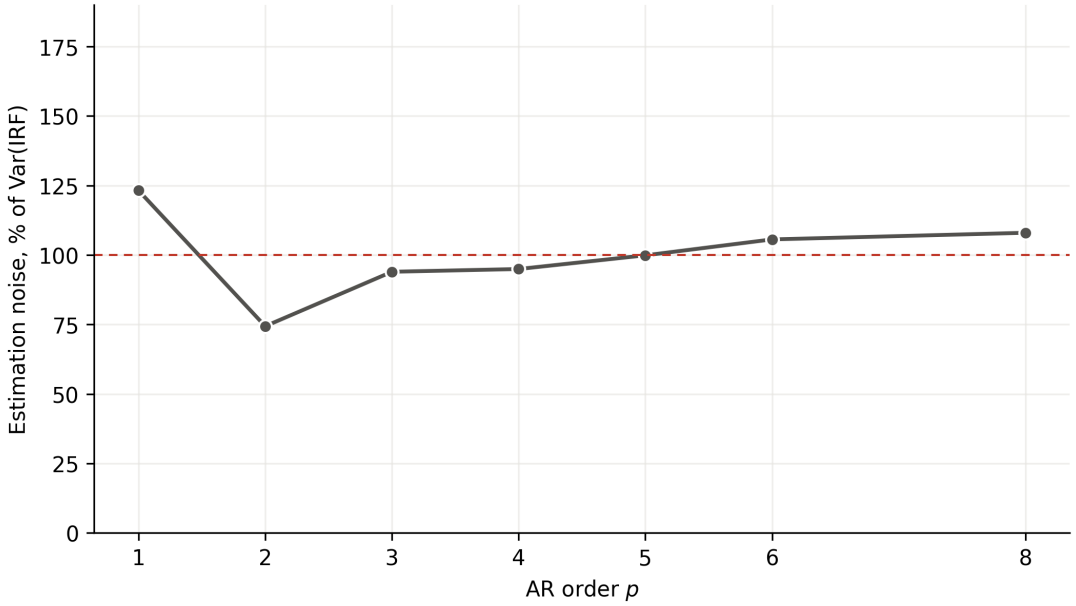
Corpus CBOW is not flat, but in this case it reflects the instability of k-means. For the CBOW embeddings, the algorithm is choosing among near-equivalent optima, and which one it lands on moves the between-group ratio by a factor of two.

## B.7 Sensitivity to the Autoregressive Order

The AR order $p$ is a nuisance parameter of the same kind as the cluster number $L$, context window $J$ and embedding dimension $r$. In this section I explore the choice of $p$.

Figure OA.5 shows the share of cross-city IRF dispersion attributable to estimation error. To compute this, I hold the DGP fixed at a common-slope $AR(8)$ whose coefficients are the pooled FE estimates on the same panel, and whose innovations are each unit's own $AR(8)$ residuals resampled $i.i.d.$ The fitted model is then $AR(p)$, unit by unit, on the simulated paths, averaged over 20 replications. Cross-city IRF dispersion is defined as the mean squared deviation from the cross-unit mean

response, horizon by horizon, and Figure OA.5 depicts the ratio between the simulated dipersion (under homogeneity) and the dispersion observed in the data.



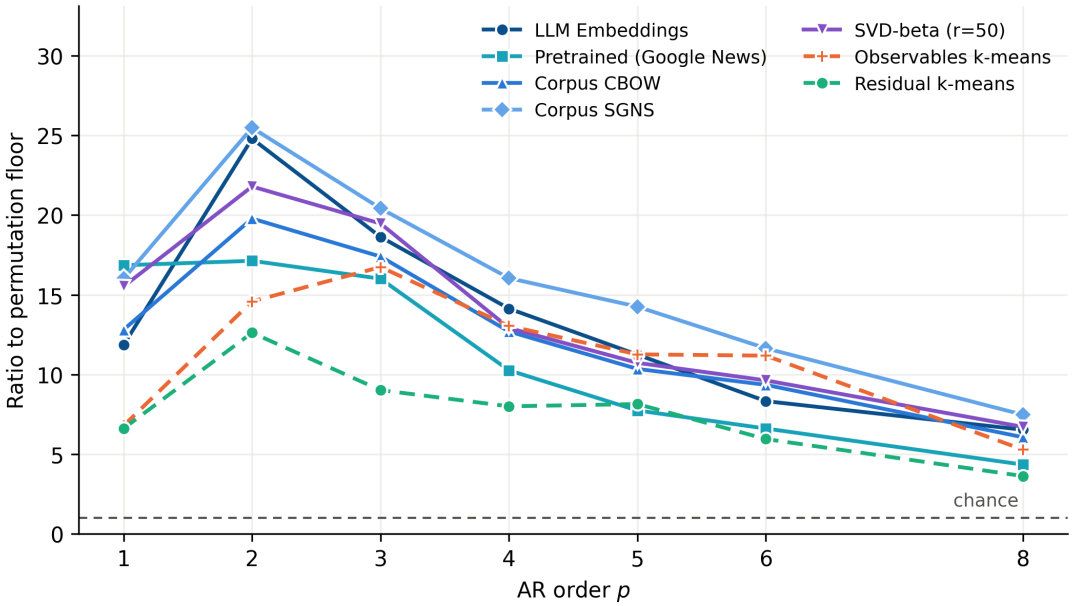

**Online Appendix Figure OA.5:** Share of total cross-city IRF variance attributable to estimation error, against the autoregressive order $p$. The benchmark holds the DGP fixed at a common-slope AR(8) and fits AR($p$) to the simulated paths. The dashed line marks the point at which the simulated data generates as much dispersion as the observed data.

On that benchmark the share is minimised at $p = 2$ (74%) and is at least 94% at every other order, reaching 100% at $p = 5$ and exceeding it from $p = 6$. One note of caution: We redraw innovations independently across units; employment shocks are not (and likely positively correlated). That would bias all depicted ratios up and may be why the share exceeds 100% from $p = 5$ on.

Figure OA.6 depicts the ratio from Table 4 for different values of $p$. Seven of the eight approaches peak at $p = 2$, though they all remain above 1 throughout.

We thus read $p = 2$ as the best low-dimensional *projection* of the dynamics for cross-sectional comparison rather than as a correctly specified model: the pooled panel clearly wants more lags, but fitting them spends per-unit degrees of freedom that a $T = 51$ series does not have, and a parsimonious fit concentrates whatever genuine heterogeneity exists into few well-estimated coefficients instead of dispersing it across many noisy ones.

One caveat applies to both figures. The Nickell bias in $\hat{\rho}$ is $O(1/T)$, but the horizon-$h$ impulse response is a compounding function of it, so the bias in the object plotted here is of order $h/T$; at $h = 10$ and $T = 51$ that is not negligible. Correcting it—by split-panel jackknife or an analytic adjustment—would plausibly shift the levels in both at the expense of even noisier estimates. We




**Online Appendix Figure OA.6:** Between-group variance of the unit-specific IRFs relative to the size-matched permutation floor, against the autoregressive order $p$, at $L = 5$. Chance is at $1\times$. Dashed lines represent the two non-text benchmarks.

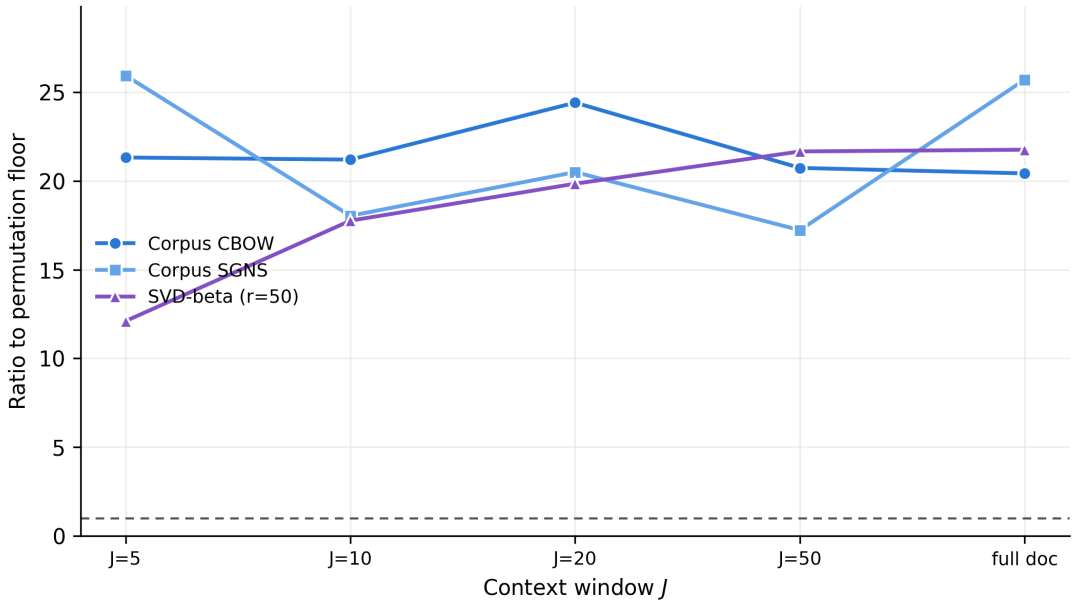
leave this for future work.

## B.8 Sensitivity to the Context Window

Section 4 uses the full document as context for both corpus-trained arms and SVD-$\beta$, on the grounds that positions within a document are exchangeable under the model, so $R$ does not depend on the window in population. Figure OA.7 tests that choice against $J \in \{5, 10, 20, 50\}$, holding $L = 5$.

Neither corpus CBOW nor corpus SGNS exhibit a clear trend. SVD-$\beta$ does trend. It holds $21.8\times$ at the full document and $21.7\times$ at $J = 50$, but falls to $12.1\times$ at $J = 5$—roughly half. This is in line with our findings from our numerical exercise: With $V$ in the low thousands, $\widehat{R}$ has on the order of six million cells, and the ordered word–context pairs the 363 descriptions supply fall from roughly eighty million at the full-document window to a few million at $J = 5$. With that sample size the estimated SVD-$\beta$ embeddings start to become somewhat noisier, reflecting the difficulty in directly matrix-factorizing the noisy sample co-occurrence structure.

One implementation detail matters for reading the two Word2Vec lines: gensim's window parameter is a maximum, not a fixed width. For each token the effective window is drawn uniformly between one and that maximum, which is equivalent to weighting contexts by their distance from the focus word [Levy et al., 2015]. Any direct comparison to SVD-$\beta$ based on Figure OA.7 is thus to be treated with caution.




**Online Appendix Figure OA.7:** Sensitivity to the context window $J$ at $L = 5$ and $p = 2$. Between-group variance of the unit-specific IRFs relative to the size-matched permutation floor, with chance at $1\times$. The SVD-$\beta$ ratio falls by about half as the window narrows to $J = 5$; neither corpus-trained arm trends either way.

## B.9 Dating the Observables: Covariate Vintage

The covariate set consists of ten variables, all at the CBSA level. Four are employment shares—manufacturing, total government, military, and federal civilian. Three are population shares—foreign-born, Black, and Hispanic. Two are education shares—the share with no high-school diploma and the share with a four-year college degree. The tenth covariate is population per square mile. We standardize each to mean zero and unit variance and then apply $k$-means with $L = 5$ (`k-means++`, 500 restarts, seed 42).

The observables benchmark in Section 4.1 clusters CBSAs on covariates measured at a single Census vintage. We take 1970 as the primary specification: it is the only vintage that is effectively predetermined, being measured one year into a fifty-one-year panel. In this section, I sweep the vintage to measure the contamination induced by using concurrent Census vintages. Table OA.1 reports the sweep.Population density enters at 1970 throughout, being the only vintage available. The other nine (industry-mix, demographic and education shares) are avilable at 1970, 1980, 1990 and 2000 with complete coverage of the 363 CBSAs.

The ratio rises monotonically in the vintage, from $15.5\times$ the permutation floor at 1970 to $18.6\times$ at 2000—a gain of about $20\%$ from dating the same nine covariates thirty years later.

We read this as a measurement of look-ahead: later measurement is itself partly an outcome of the dynamics being explained. On this metric the advantage is worth roughly a fifth of the benchmark's

| Vintage | Years elapsed | Var(Between) | Ratio | Max share |
|---|---|---|---|---|
| 1970 | 1 | 0.0209 | 15.5× | 36% |
| 1980 | 11 | 0.0238 | 17.2× | 46% |
| 1990 | 21 | 0.0243 | 17.4× | 44% |
| 2000 | 31 | 0.0248 | 18.6× | 41% |

**Online Appendix Table OA.1:** Observables benchmark at each Census vintage, $L = 5$ and $p = 2$. "Years elapsed" is the number of years of the 1969–2019 outcome window already observed at the measurement date. "Ratio" is between-group variance over the size-matched permutation floor. All nine covariates are complete at every vintage; population density enters at 1970 throughout, being the only vintage available.

apparent performance. That matters beyond the choice of covariates, because the same objection applies to the text. The descriptions span 1979–2019 and so are concurrent with the outcome window rather than predetermined, a limitation Section 4 states but cannot quantify. The vintage sweep gives an order of magnitude for what such concurrency can buy—roughly $20\%$ here—which is well short of the margin by which the text-based groupings exceed the benchmark.

## B.10 Interpreting the Clusters

Table OA.2 reports the cluster assignment for each of the 363 CBSAs under seven clustering approaches. Clusters are numbered 1–5 within each method; descriptions for the text-based methods follow below. Figure OA.8 shows the geographic distribution of cluster assignments across the continental United States.

**Online Appendix Table OA.2:** Cluster assignments for all 363 CBSAs: corpus CBOW (CB), corpus SGNS (CS), closed-form SVD-$\beta$ at $r = 50$ (SB), pretrained Google News vectors (PW), LLM embedding (LE, $k$-means on OpenAI `text-embedding-3-large`), residual $k$-means (RK, $k$-means on pooled AR(2) residuals from Section 4.1), and observables $k$-means (OB, on the 1970 covariates of Section 4.1).Numeric labels correspond to the cluster descriptions below the table.

| | | **Clustering Method** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **CBSA** | **State** | **CB** | **CS** | **SB** | **PW** | **LE** | **RK** | **OB** |
| Abilene | TX | 4 | 3 | 2 | 3 | 3 | 2 | 2 |
| Akron | OH | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Albany | GA | 2 | 3 | 2 | 3 | 2 | 1 | 3 |
| Albany-Schenectady-Troy | NY | 5 | 3 | 2 | 5 | 5 | 3 | 2 |
| Albuquerque | NM | 3 | 1 | 2 | 5 | 3 | 3 | 2 |
| Alexandria | LA | 4 | 3 | 2 | 5 | 3 | 3 | 3 |
| Allentown-Bethlehem-Easton | PA-NJ | 5 | 5 | 1 | 1 | 4 | 1 | 4 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Altoona | PA | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Amarillo | TX | 4 | 3 | 2 | 3 | 3 | 3 | 2 |
| Ames | IA | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Anderson | IN | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Anderson | SC | 2 | 3 | 1 | 1 | 4 | 4 | 4 |
| Ann Arbor | MI | 4 | 1 | 2 | 5 | 2 | 4 | 2 |
| Anniston-Oxford | AL | 5 | 4 | 5 | 1 | 2 | 3 | 3 |
| Appleton | WI | 5 | 3 | 1 | 1 | 4 | 1 | 4 |
| Asheville | NC | 3 | 2 | 3 | 3 | 2 | 4 | 4 |
| Athens-Clarke County | GA | 2 | 3 | 2 | 5 | 2 | 1 | 3 |
| Atlanta-Sandy Springs-Marietta | GA | 2 | 1 | 2 | 3 | 5 | 1 | 3 |
| Atlantic City-Hammonton | NJ | 5 | 5 | 1 | 4 | 5 | 1 | 3 |
| Auburn-Opelika | AL | 2 | 3 | 2 | 3 | 2 | 1 | 3 |
| Augusta-Richmond County | GA-SC | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Austin-Round Rock-San Marcos | TX | 4 | 1 | 2 | 5 | 5 | 1 | 2 |
| Bakersfield-Delano | CA | 1 | 2 | 3 | 4 | 1 | 3 | 2 |
| Baltimore-Towson | MD | 2 | 3 | 2 | 5 | 2 | 3 | 3 |
| Bangor | ME | 2 | 3 | 2 | 3 | 2 | 3 | 4 |
| Barnstable Town | MA | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Baton Rouge | LA | 1 | 4 | 4 | 2 | 3 | 3 | 3 |
| Battle Creek | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Bay City | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Beaumont-Port Arthur | TX | 1 | 4 | 4 | 2 | 3 | 2 | 3 |
| Bellingham | WA | 3 | 2 | 3 | 4 | 2 | 3 | 2 |
| Bend | OR | 3 | 2 | 3 | 4 | 5 | 4 | 2 |
| Billings | MT | 4 | 3 | 2 | 2 | 3 | 3 | 2 |
| Binghamton | NY | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Birmingham-Hoover | AL | 2 | 3 | 2 | 1 | 2 | 3 | 3 |
| Bismarck | ND | 4 | 3 | 2 | 2 | 3 | 3 | 2 |
| Blacksburg-Christiansburg-Radford | VA | 4 | 3 | 2 | 5 | 2 | 1 | 4 |
| Bloomington | IN | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Bloomington-Normal | IL | 5 | 1 | 2 | 5 | 2 | 3 | 2 |
| Boise City-Nampa | ID | 4 | 1 | 2 | 3 | 5 | 3 | 2 |
| Boston-Cambridge-Quincy | MA-NH | 3 | 1 | 2 | 5 | 5 | 1 | 5 |
| Boulder | CO | 4 | 1 | 2 | 5 | 5 | 1 | 2 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Bowling Green | KY | 4 | 3 | 2 | 3 | 2 | 4 | 4 |
| Bremerton-Silverdale | WA | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Bridgeport-Stamford-Norwalk | CT | 2 | 5 | 1 | 3 | 5 | 1 | 5 |
| Brownsville-Harlingen | TX | 1 | 2 | 3 | 4 | 3 | 3 | 5 |
| Brunswick | GA | 2 | 3 | 3 | 3 | 2 | 1 | 3 |
| Buffalo-Niagara Falls | NY | 5 | 5 | 1 | 1 | 4 | 3 | 5 |
| Burlington | NC | 2 | 5 | 1 | 3 | 2 | 1 | 4 |
| Burlington-South Burlington | VT | 4 | 1 | 2 | 5 | 5 | 1 | 2 |
| Canton-Massillon | OH | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Cape Coral-Fort Myers | FL | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Cape Girardeau-Jackson | MO-IL | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| Carson City | NV | 4 | 3 | 2 | 5 | 2 | 4 | 2 |
| Casper | WY | 4 | 4 | 4 | 2 | 3 | 2 | 2 |
| Cedar Rapids | IA | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| Champaign-Urbana | IL | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Charleston | WV | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Charleston-North Charleston-Summerville | SC | 3 | 1 | 2 | 3 | 5 | 3 | 1 |
| Charlotte-Gastonia-Rock Hill | NC-SC | 2 | 3 | 2 | 3 | 5 | 1 | 4 |
| Charlottesville | VA | 3 | 1 | 2 | 5 | 2 | 1 | 3 |
| Chattanooga | TN-GA | 2 | 5 | 1 | 3 | 4 | 4 | 4 |
| Cheyenne | WY | 1 | 4 | 2 | 2 | 3 | 3 | 2 |
| Chicago-Joliet-Naperville | IL-IN-WI | 2 | 5 | 1 | 3 | 4 | 1 | 5 |
| Chico | CA | 4 | 2 | 3 | 4 | 1 | 3 | 2 |
| Cincinnati-Middletown | OH-KY-IN | 2 | 3 | 2 | 3 | 4 | 1 | 4 |
| Clarksville | TN-KY | 1 | 4 | 5 | 3 | 3 | 4 | 1 |
| Cleveland | TN | 2 | 3 | 2 | 3 | 2 | 4 | 4 |
| Cleveland-Elyria-Mentor | OH | 5 | 5 | 1 | 1 | 4 | 1 | 5 |
| Coeur d'Alene | ID | 3 | 2 | 3 | 4 | 5 | 4 | 2 |
| College Station-Bryan | TX | 4 | 1 | 2 | 5 | 3 | 2 | 3 |
| Colorado Springs | CO | 4 | 4 | 5 | 5 | 3 | 1 | 1 |
| Columbia | MO | 4 | 3 | 2 | 5 | 2 | 3 | 2 |
| Columbia | SC | 2 | 3 | 2 | 5 | 2 | 1 | 3 |
| Columbus | GA-AL | 2 | 3 | 2 | 3 | 2 | 3 | 1 |
| Columbus | IN | 4 | 5 | 1 | 1 | 4 | 4 | 4 |
| Columbus | OH | 4 | 3 | 2 | 3 | 2 | 1 | 2 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Corpus Christi | TX | 1 | 4 | 4 | 2 | 3 | 2 | 2 |
| Corvallis | OR | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Crestview-Fort Walton Beach-Destin | FL | 3 | 2 | 3 | 4 | 5 | 3 | 1 |
| Cumberland | MD-WV | 5 | 5 | 1 | 3 | 4 | 3 | 4 |
| Dallas-Fort Worth-Arlington | TX | 4 | 1 | 2 | 3 | 5 | 1 | 4 |
| Dalton | GA | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Danville | IL | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Danville | VA | 5 | 5 | 1 | 1 | 4 | 4 | 3 |
| Davenport-Moline-Rock Island | IA-IL | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Dayton | OH | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Decatur | AL | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Decatur | IL | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Deltona-Daytona Beach-Ormond Beach | FL | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Denver-Aurora-Broomfield | CO | 4 | 1 | 4 | 2 | 5 | 1 | 2 |
| Des Moines-West Des Moines | IA | 4 | 1 | 2 | 3 | 5 | 3 | 2 |
| Detroit-Warren-Livonia | MI | 5 | 5 | 1 | 1 | 4 | 4 | 5 |
| Dothan | AL | 2 | 3 | 2 | 3 | 2 | 4 | 3 |
| Dover | DE | 1 | 3 | 2 | 5 | 2 | 3 | 1 |
| Dubuque | IA | 4 | 3 | 2 | 1 | 4 | 3 | 4 |
| Duluth | MN-WI | 5 | 5 | 1 | 1 | 4 | 3 | 2 |
| Durham-Chapel Hill | NC | 4 | 1 | 2 | 5 | 2 | 3 | 3 |
| Eau Claire | WI | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| El Centro | CA | 1 | 2 | 3 | 4 | 1 | 3 | 5 |
| Elizabethtown | KY | 1 | 3 | 2 | 3 | 2 | 3 | 1 |
| Elkhart-Goshen | IN | 4 | 2 | 2 | 1 | 4 | 4 | 4 |
| Elmira | NY | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| El Paso | TX | 1 | 2 | 2 | 1 | 3 | 3 | 5 |
| Erie | PA | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Eugene-Springfield | OR | 4 | 3 | 3 | 4 | 2 | 4 | 2 |
| Evansville | IN-KY | 2 | 5 | 1 | 3 | 4 | 3 | 4 |
| Fargo | ND-MN | 4 | 1 | 2 | 3 | 2 | 3 | 2 |
| Farmington | NM | 1 | 4 | 4 | 2 | 3 | 2 | 2 |
| Fayetteville | NC | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Fayetteville-Springdale-Rogers | AR-MO | 4 | 1 | 2 | 3 | 2 | 1 | 4 |
| Flagstaff | AZ | 3 | 2 | 3 | 4 | 2 | 1 | 2 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Flint | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Florence | SC | 2 | 3 | 2 | 3 | 2 | 1 | 3 |
| Florence-Muscle Shoals | AL | 5 | 3 | 1 | 3 | 2 | 3 | 4 |
| Fond du Lac | WI | 4 | 3 | 2 | 1 | 4 | 3 | 4 |
| Fort Collins-Loveland | CO | 4 | 1 | 2 | 3 | 5 | 1 | 2 |
| Fort Smith | AR-OK | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Fort Wayne | IN | 5 | 5 | 1 | 3 | 4 | 4 | 4 |
| Fresno | CA | 1 | 2 | 3 | 4 | 1 | 3 | 2 |
| Gadsden | AL | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Gainesville | FL | 4 | 1 | 2 | 5 | 2 | 1 | 2 |
| Gainesville | GA | 3 | 2 | 3 | 3 | 2 | 1 | 4 |
| Glens Falls | NY | 5 | 3 | 2 | 3 | 4 | 3 | 4 |
| Goldsboro | NC | 1 | 3 | 2 | 4 | 3 | 3 | 3 |
| Grand Forks | ND-MN | 4 | 3 | 2 | 3 | 3 | 3 | 5 |
| Grand Junction | CO | 4 | 2 | 4 | 2 | 3 | 2 | 2 |
| Grand Rapids-Wyoming | MI | 4 | 1 | 2 | 3 | 4 | 4 | 4 |
| Great Falls | MT | 4 | 3 | 2 | 2 | 3 | 3 | 2 |
| Greeley | CO | 1 | 2 | 4 | 2 | 3 | 1 | 2 |
| Green Bay | WI | 5 | 3 | 2 | 3 | 4 | 3 | 4 |
| Greensboro-High Point | NC | 2 | 5 | 1 | 1 | 4 | 4 | 4 |
| Greenville | NC | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Greenville-Mauldin-Easley | SC | 2 | 5 | 1 | 3 | 4 | 1 | 4 |
| Gulfport-Biloxi | MS | 1 | 4 | 4 | 2 | 5 | 3 | 1 |
| Hagerstown-Martinsburg | MD-WV | 5 | 3 | 2 | 3 | 2 | 3 | 4 |
| Hanford-Corcoran | CA | 1 | 2 | 3 | 4 | 1 | 3 | 1 |
| Harrisburg-Carlisle | PA | 4 | 3 | 2 | 3 | 2 | 3 | 2 |
| Harrisonburg | VA | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| Hartford-West Hartford-East Hartford | CT | 5 | 1 | 2 | 1 | 5 | 1 | 5 |
| Hattiesburg | MS | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Hickory-Lenoir-Morganton | NC | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Hinesville-Fort Stewart | GA | 1 | 4 | 5 | 5 | 3 | 5 | 1 |
| Holland-Grand Haven | MI | 4 | 1 | 2 | 1 | 4 | 4 | 4 |
| Hot Springs | AR | 3 | 2 | 3 | 4 | 2 | 4 | 4 |
| Houma-Bayou Cane-Thibodaux | LA | 1 | 4 | 4 | 2 | 3 | 2 | 3 |
| Houston-Sugar Land-Baytown | TX | 4 | 4 | 4 | 2 | 3 | 2 | 3 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Huntington-Ashland | WV-KY-OH | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Huntsville | AL | 2 | 1 | 5 | 5 | 2 | 1 | 1 |
| Idaho Falls | ID | 4 | 3 | 2 | 3 | 3 | 3 | 2 |
| Indianapolis-Carmel | IN | 2 | 5 | 2 | 3 | 4 | 1 | 4 |
| Iowa City | IA | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Ithaca | NY | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Jackson | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Jackson | MS | 2 | 3 | 2 | 5 | 2 | 1 | 3 |
| Jackson | TN | 2 | 3 | 2 | 3 | 2 | 4 | 3 |
| Jacksonville | FL | 2 | 1 | 2 | 3 | 2 | 1 | 3 |
| Jacksonville | NC | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Janesville | WI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Jefferson City | MO | 4 | 3 | 2 | 5 | 2 | 3 | 2 |
| Johnson City | TN | 2 | 3 | 2 | 3 | 2 | 4 | 4 |
| Johnstown | PA | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Jonesboro | AR | 2 | 3 | 2 | 3 | 2 | 3 | 4 |
| Joplin | MO | 5 | 3 | 2 | 3 | 2 | 4 | 4 |
| Kalamazoo-Portage | MI | 5 | 1 | 2 | 3 | 2 | 3 | 4 |
| Kankakee-Bradley | IL | 5 | 5 | 1 | 3 | 4 | 3 | 4 |
| Kansas City | MO-KS | 4 | 3 | 2 | 3 | 2 | 1 | 2 |
| Kennewick-Pasco-Richland | WA | 1 | 2 | 2 | 4 | 3 | 3 | 2 |
| Killeen-Temple-Fort Hood | TX | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Kingsport-Bristol-Bristol | TN-VA | 5 | 5 | 1 | 3 | 4 | 3 | 4 |
| Kingston | NY | 3 | 5 | 1 | 1 | 4 | 3 | 4 |
| Knoxville | TN | 2 | 3 | 2 | 3 | 2 | 3 | 4 |
| Kokomo | IN | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| La Crosse | WI-MN | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| Lafayette | IN | 4 | 1 | 2 | 1 | 4 | 3 | 2 |
| Lafayette | LA | 4 | 4 | 4 | 2 | 3 | 2 | 3 |
| Lake Charles | LA | 1 | 4 | 4 | 2 | 3 | 3 | 3 |
| Lake Havasu City-Kingman | AZ | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Lakeland-Winter Haven | FL | 3 | 2 | 3 | 4 | 5 | 1 | 3 |
| Lancaster | PA | 4 | 1 | 2 | 3 | 2 | 1 | 4 |
| Lansing-East Lansing | MI | 4 | 3 | 2 | 5 | 2 | 3 | 2 |
| Laredo | TX | 1 | 2 | 3 | 4 | 3 | 2 | 5 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Las Cruces | NM | 1 | 2 | 3 | 4 | 3 | 3 | 1 |
| Las Vegas-Paradise | NV | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Lawrence | KS | 3 | 1 | 2 | 5 | 2 | 3 | 2 |
| Lawton | OK | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Lebanon | PA | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Lewiston | ID-WA | 1 | 3 | 2 | 3 | 2 | 3 | 4 |
| Lewiston-Auburn | ME | 2 | 5 | 1 | 3 | 4 | 1 | 4 |
| Lexington-Fayette | KY | 2 | 3 | 2 | 3 | 2 | 4 | 2 |
| Lima | OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Lincoln | NE | 4 | 3 | 2 | 5 | 2 | 3 | 2 |
| Little Rock-North Little Rock-Conway | AR | 4 | 3 | 2 | 5 | 2 | 3 | 3 |
| Logan | UT-ID | 4 | 1 | 2 | 3 | 2 | 3 | 2 |
| Longview | TX | 1 | 4 | 4 | 2 | 3 | 2 | 3 |
| Longview | WA | 1 | 2 | 3 | 2 | 4 | 1 | 4 |
| Los Angeles-Long Beach-Santa Ana | CA | 3 | 2 | 3 | 1 | 5 | 1 | 5 |
| Louisville/Jefferson County | KY-IN | 2 | 3 | 2 | 3 | 2 | 1 | 4 |
| Lubbock | TX | 4 | 3 | 2 | 3 | 3 | 3 | 2 |
| Lynchburg | VA | 2 | 3 | 2 | 3 | 2 | 1 | 4 |
| Macon | GA | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Madera-Chowchilla | CA | 1 | 2 | 3 | 4 | 1 | 3 | 4 |
| Madison | WI | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Manchester-Nashua | NH | 2 | 1 | 2 | 5 | 5 | 1 | 4 |
| Manhattan | KS | 1 | 1 | 5 | 5 | 2 | 3 | 1 |
| Mankato-North Mankato | MN | 4 | 3 | 2 | 3 | 2 | 3 | 2 |
| Mansfield | OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| McAllen-Edinburg-Mission | TX | 1 | 2 | 3 | 4 | 3 | 3 | 5 |
| Medford | OR | 3 | 2 | 3 | 4 | 2 | 4 | 2 |
| Memphis | TN-MS-AR | 2 | 3 | 2 | 3 | 2 | 1 | 3 |
| Merced | CA | 1 | 2 | 3 | 4 | 1 | 3 | 1 |
| Miami-Fort Lauderdale-Pompano Beach | FL | 3 | 2 | 3 | 4 | 5 | 1 | 5 |
| Michigan City-La Porte | IN | 5 | 5 | 1 | 3 | 4 | 1 | 4 |
| Midland | TX | 1 | 4 | 4 | 2 | 3 | 2 | 2 |
| Milwaukee-Waukesha-West Allis | WI | 2 | 5 | 1 | 3 | 4 | 1 | 5 |
| Minneapolis-St. Paul-Bloomington | MN-WI | 2 | 1 | 2 | 5 | 5 | 1 | 2 |
| Missoula | MT | 3 | 2 | 3 | 3 | 2 | 3 | 2 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Mobile | AL | 2 | 3 | 2 | 2 | 2 | 3 | 3 |
| Modesto | CA | 1 | 2 | 3 | 4 | 1 | 3 | 4 |
| Monroe | LA | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Monroe | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Montgomery | AL | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Morgantown | WV | 4 | 3 | 2 | 5 | 2 | 3 | 2 |
| Morristown | TN | 2 | 3 | 1 | 1 | 4 | 4 | 4 |
| Mount Vernon-Anacortes | WA | 1 | 2 | 3 | 4 | 1 | 4 | 2 |
| Muncie | IN | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Muskegon-Norton Shores | MI | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Myrtle Beach-North Myrtle Beach-Conway | SC | 3 | 2 | 3 | 4 | 5 | 3 | 3 |
| Napa | CA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Naples-Marco Island | FL | 3 | 2 | 3 | 4 | 5 | 4 | 2 |
| Nashville-Davidson–Murfreesboro–Franklin | TN | 3 | 1 | 2 | 3 | 2 | 4 | 3 |
| New Haven-Milford | CT | 2 | 5 | 1 | 5 | 5 | 1 | 5 |
| New Orleans-Metairie-Kenner | LA | 3 | 4 | 4 | 2 | 5 | 3 | 3 |
| New York-Northern New Jersey-Long Island | NY-NJ-PA | 3 | 1 | 2 | 5 | 5 | 1 | 5 |
| Niles-Benton Harbor | MI | 5 | 5 | 1 | 3 | 4 | 4 | 4 |
| North Port-Bradenton-Sarasota | FL | 3 | 2 | 3 | 4 | 5 | 4 | 2 |
| Norwich-New London | CT | 1 | 4 | 5 | 1 | 5 | 3 | 2 |
| Ocala | FL | 3 | 2 | 3 | 4 | 2 | 1 | 3 |
| Ocean City | NJ | 3 | 2 | 3 | 4 | 5 | 3 | 2 |
| Odessa | TX | 1 | 4 | 4 | 2 | 3 | 2 | 2 |
| Ogden-Clearfield | UT | 4 | 3 | 2 | 3 | 5 | 3 | 1 |
| Oklahoma City | OK | 4 | 4 | 4 | 2 | 3 | 3 | 2 |
| Olympia | WA | 1 | 3 | 2 | 5 | 2 | 3 | 2 |
| Omaha-Council Bluffs | NE-IA | 4 | 1 | 2 | 3 | 2 | 3 | 2 |
| Orlando-Kissimmee-Sanford | FL | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Oshkosh-Neenah | WI | 5 | 3 | 2 | 3 | 4 | 3 | 4 |
| Owensboro | KY | 5 | 3 | 1 | 3 | 4 | 1 | 4 |
| Oxnard-Thousand Oaks-Ventura | CA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Palm Bay-Melbourne-Titusville | FL | 3 | 1 | 5 | 5 | 5 | 1 | 2 |
| Palm Coast | FL | 3 | 2 | 3 | 4 | 5 | 1 | 3 |
| Panama City-Lynn Haven-Panama City Beach | FL | 3 | 2 | 3 | 4 | 5 | 3 | 1 |
| Parkersburg-Marietta-Vienna | WV-OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Pascagoula | MS | 1 | 4 | 5 | 2 | 3 | 2 | 4 |
| Pensacola-Ferry Pass-Brent | FL | 1 | 4 | 3 | 2 | 3 | 3 | 1 |
| Peoria | IL | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Philadelphia-Camden-Wilmington | PA-NJ-DE-MD | 2 | 5 | 2 | 3 | 5 | 1 | 5 |
| Phoenix-Mesa-Glendale | AZ | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Pine Bluff | AR | 5 | 5 | 1 | 1 | 4 | 3 | 3 |
| Pittsburgh | PA | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Pittsfield | MA | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Pocatello | ID | 5 | 3 | 1 | 1 | 2 | 3 | 2 |
| Portland-South Portland-Biddeford | ME | 3 | 3 | 2 | 3 | 5 | 1 | 2 |
| Portland-Vancouver-Hillsboro | OR-WA | 3 | 1 | 3 | 3 | 5 | 1 | 2 |
| Port St. Lucie | FL | 3 | 2 | 3 | 4 | 5 | 1 | 3 |
| Poughkeepsie-Newburgh-Middletown | NY | 3 | 3 | 2 | 3 | 4 | 3 | 2 |
| Prescott | AZ | 3 | 2 | 3 | 4 | 3 | 4 | 2 |
| Providence-New Bedford-Fall River | RI-MA | 2 | 5 | 1 | 1 | 4 | 1 | 5 |
| Provo-Orem | UT | 4 | 1 | 2 | 5 | 5 | 1 | 2 |
| Pueblo | CO | 5 | 5 | 1 | 1 | 4 | 3 | 2 |
| Punta Gorda | FL | 3 | 2 | 3 | 4 | 5 | 4 | 2 |
| Racine | WI | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Raleigh-Cary | NC | 3 | 1 | 2 | 5 | 5 | 1 | 3 |
| Rapid City | SD | 3 | 3 | 2 | 4 | 3 | 3 | 2 |
| Reading | PA | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Redding | CA | 3 | 3 | 3 | 4 | 1 | 3 | 2 |
| Reno-Sparks | NV | 3 | 2 | 3 | 2 | 5 | 1 | 2 |
| Richmond | VA | 2 | 3 | 2 | 5 | 2 | 1 | 3 |
| Riverside-San Bernardino-Ontario | CA | 1 | 2 | 3 | 4 | 5 | 1 | 2 |
| Roanoke | VA | 5 | 3 | 2 | 3 | 2 | 1 | 4 |
| Rochester | MN | 4 | 1 | 2 | 5 | 2 | 3 | 2 |
| Rochester | NY | 5 | 5 | 1 | 1 | 2 | 3 | 4 |
| Rockford | IL | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Rocky Mount | NC | 2 | 5 | 1 | 1 | 4 | 4 | 3 |
| Rome | GA | 2 | 3 | 2 | 3 | 2 | 4 | 4 |
| Sacramento–Arden-Arcade–Roseville | CA | 3 | 2 | 3 | 5 | 1 | 3 | 2 |
| Saginaw-Saginaw Township North | MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| St. Cloud | MN | 4 | 3 | 2 | 3 | 2 | 3 | 4 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| St. George | UT | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| St. Joseph | MO-KS | 5 | 3 | 2 | 3 | 2 | 3 | 4 |
| St. Louis | MO-IL | 5 | 5 | 2 | 3 | 2 | 1 | 4 |
| Salem | OR | 3 | 3 | 3 | 4 | 2 | 3 | 2 |
| Salinas | CA | 3 | 2 | 3 | 4 | 1 | 3 | 1 |
| Salisbury | MD | 4 | 3 | 2 | 3 | 2 | 1 | 3 |
| Salt Lake City | UT | 4 | 1 | 2 | 5 | 5 | 1 | 2 |
| San Angelo | TX | 1 | 4 | 2 | 2 | 3 | 3 | 2 |
| San Antonio-New Braunfels | TX | 1 | 1 | 2 | 5 | 3 | 3 | 1 |
| San Diego-Carlsbad-San Marcos | CA | 3 | 1 | 3 | 5 | 5 | 1 | 1 |
| Sandusky | OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| San Francisco-Oakland-Fremont | CA | 3 | 1 | 3 | 1 | 5 | 1 | 5 |
| San Jose-Sunnyvale-Santa Clara | CA | 3 | 1 | 3 | 5 | 5 | 1 | 5 |
| San Luis Obispo-Paso Robles | CA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Santa Barbara-Santa Maria-Goleta | CA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Santa Cruz-Watsonville | CA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Santa Fe | NM | 3 | 2 | 3 | 5 | 3 | 3 | 2 |
| Santa Rosa-Petaluma | CA | 3 | 2 | 3 | 4 | 1 | 1 | 2 |
| Savannah | GA | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Scranton–Wilkes-Barre | PA | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Seattle-Tacoma-Bellevue | WA | 3 | 1 | 2 | 5 | 5 | 1 | 2 |
| Sebastian-Vero Beach | FL | 3 | 2 | 3 | 4 | 5 | 4 | 3 |
| Sheboygan | WI | 5 | 3 | 1 | 1 | 4 | 1 | 4 |
| Sherman-Denison | TX | 4 | 3 | 2 | 3 | 5 | 1 | 4 |
| Shreveport-Bossier City | LA | 1 | 4 | 4 | 2 | 3 | 3 | 3 |
| Sioux City | IA-NE-SD | 4 | 3 | 2 | 4 | 4 | 3 | 4 |
| Sioux Falls | SD | 4 | 1 | 2 | 3 | 2 | 3 | 2 |
| South Bend-Mishawaka | IN-MI | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Spartanburg | SC | 2 | 5 | 1 | 1 | 4 | 1 | 4 |
| Spokane | WA | 4 | 3 | 2 | 3 | 2 | 1 | 2 |
| Springfield | IL | 5 | 3 | 2 | 5 | 2 | 3 | 2 |
| Springfield | MA | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Springfield | MO | 4 | 3 | 2 | 3 | 2 | 1 | 4 |
| Springfield | OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| State College | PA | 4 | 1 | 2 | 5 | 2 | 3 | 2 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Steubenville-Weirton | OH-WV | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Stockton | CA | 1 | 2 | 3 | 4 | 1 | 3 | 2 |
| Sumter | SC | 2 | 3 | 1 | 3 | 2 | 3 | 3 |
| Syracuse | NY | 5 | 5 | 1 | 1 | 4 | 3 | 2 |
| Tallahassee | FL | 3 | 1 | 2 | 5 | 2 | 1 | 3 |
| Tampa-St. Petersburg-Clearwater | FL | 3 | 2 | 3 | 4 | 5 | 1 | 2 |
| Terre Haute | IN | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Texarkana, TX-Texarkana | AR | 5 | 3 | 2 | 3 | 3 | 1 | 3 |
| Toledo | OH | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Topeka | KS | 4 | 3 | 2 | 3 | 2 | 3 | 2 |
| Trenton-Ewing | NJ | 2 | 5 | 1 | 5 | 2 | 1 | 5 |
| Tucson | AZ | 1 | 2 | 3 | 5 | 3 | 1 | 2 |
| Tulsa | OK | 4 | 4 | 4 | 2 | 3 | 2 | 4 |
| Tuscaloosa | AL | 2 | 3 | 2 | 1 | 2 | 3 | 3 |
| Tyler | TX | 4 | 3 | 2 | 3 | 3 | 3 | 3 |
| Utica-Rome | NY | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Valdosta | GA | 2 | 3 | 2 | 3 | 2 | 3 | 3 |
| Vallejo-Fairfield | CA | 1 | 2 | 3 | 5 | 1 | 3 | 1 |
| Victoria | TX | 1 | 4 | 4 | 2 | 3 | 2 | 4 |
| Vineland-Millville-Bridgeton | NJ | 2 | 5 | 1 | 1 | 4 | 1 | 4 |
| Virginia Beach-Norfolk-Newport News | VA-NC | 1 | 4 | 5 | 2 | 5 | 3 | 1 |
| Visalia-Porterville | CA | 1 | 2 | 3 | 4 | 1 | 3 | 2 |
| Waco | TX | 4 | 3 | 2 | 3 | 3 | 3 | 3 |
| Warner Robins | GA | 1 | 4 | 5 | 5 | 3 | 3 | 1 |
| Washington-Arlington-Alexandria | DC-VA-MD-WV | 3 | 1 | 2 | 5 | 5 | 3 | 1 |
| Waterloo-Cedar Falls | IA | 5 | 3 | 2 | 1 | 4 | 3 | 4 |
| Wausau | WI | 4 | 3 | 2 | 3 | 2 | 3 | 4 |
| Wenatchee-East Wenatchee | WA | 3 | 2 | 3 | 4 | 1 | 3 | 2 |
| Wheeling | WV-OH | 5 | 5 | 1 | 1 | 4 | 3 | 4 |
| Wichita | KS | 5 | 3 | 1 | 1 | 3 | 1 | 2 |
| Wichita Falls | TX | 4 | 3 | 2 | 2 | 3 | 3 | 1 |
| Williamsport | PA | 5 | 5 | 1 | 1 | 4 | 1 | 4 |
| Wilmington | NC | 3 | 1 | 2 | 3 | 2 | 1 | 3 |
| Winchester | VA-WV | 4 | 3 | 2 | 3 | 2 | 4 | 4 |
| Winston-Salem | NC | 2 | 1 | 1 | 1 | 2 | 1 | 4 |

*Continued on next page*

*Table OA.2 continued*

| CBSA | State | CB | CS | SB | PW | LE | RK | OB |
|---|---|---|---|---|---|---|---|---|
| Worcester | MA | 2 | 5 | 1 | 3 | 4 | 1 | 4 |
| Yakima | WA | 1 | 2 | 3 | 4 | 1 | 3 | 2 |
| York-Hanover | PA | 5 | 3 | 1 | 3 | 4 | 1 | 4 |
| Youngstown-Warren-Boardman | OH-PA | 5 | 5 | 1 | 1 | 4 | 4 | 4 |
| Yuba City | CA | 1 | 2 | 3 | 4 | 1 | 3 | 1 |
| Yuma | AZ | 1 | 2 | 3 | 4 | 1 | 3 | 1 |

## Cluster Descriptions by Method

**Corpus CBOW (CB).**

1. *Amenity Migration and Population-Led Growth* (57 CBSAs). Sun Belt and Western metros whose growth is driven by in-migration of retirees and lifestyle migrants rather than by industry. Construction, real estate, retail, and healthcare are the core employers, which leaves these labor markets acutely exposed to housing cycles and to the 2008 shock. Examples: Asheville, Austin, Boise City, Cape Coral, Dallas.
2. *University and Hospital Anchors* (62 CBSAs). Flagship universities and major hospital systems dominate employment. Unemployment runs persistently below national averages and educational attainment is high, but the labor market is bifurcated between credentialed professionals and low-wage student and service workers. Examples: Ames, Ann Arbor, Boston, Boulder, Champaign–Urbana.
3. *Regional Hubs in Manufacturing Transition* (68 CBSAs). Legacy manufacturing—textiles, paper, steel, timber—has declined since 1979 and been partly replaced by "eds and meds," logistics along interstate and port corridors, and government or defense spending. These metros serve as retail and medical hubs for rural hinterlands, with below-average wages and uneven recoveries. Examples: Allentown, Birmingham, Charlotte, Chattanooga, Cincinnati.
4. *Severe Deindustrialization* (85 CBSAs). Steel, auto, textile, and tire dependence followed by heavy job losses after 1979. Loss of high-wage unionized work, wage stagnation, population decline, and out-migration of younger workers, with "eds and meds," retail, and logistics as partial replacements. Examples: Akron, Buffalo, Chicago, Cleveland, Detroit.
5. *Agriculture and Resource Extraction* (91 CBSAs). Economies anchored in agriculture, oil and gas, timber, or mining, with boom-bust cycles tied to commodity prices and weather. Seasonal low-wage work, chronically high unemployment and poverty, and large Hispanic and immigrant workforces, often near the border. Examples: Bakersfield, Brownsville, Corpus Christi, El Paso, Fresno.

**Corpus SGNS (CS).**

1. *Knowledge and Anchor-Institution Economies* (60 CBSAs). Universities, hospital systems and research employers dominate, with a shift out of manufacturing, resource extraction and defense into knowledge work and services. Unemployment is persistently low and recoveries—2008 in particular—are fast, but growth arrives with housing pressure and a labor market split between credentialed and low-wage service work. Examples: Albuquerque, Ames, Ann Arbor, Atlanta, Austin.
2. *Amenity Migration and Tourism* (71 CBSAs). Sun Belt, Western and coastal destinations growing through retiree and lifestyle in-migration rather than through industry. Tourism, hospitality, retail and healthcare dominate, construction and real estate amplify the cycle, and the 2008 housing shock hit hard. Seasonal low-wage work and large Hispanic workforces are common. Examples: Asheville, Bakersfield, Barnstable Town, Bend, Cape Coral.
3. *Regional Hubs in Transition* (109 CBSAs). The largest group: mid-sized metros that lost textiles, paper, timber, mining or tobacco employment and replaced part of it with "eds and meds," logistics along interstate corridors, and state or federal anchors. Each serves as the retail and medical hub of a broad rural hinterland. Examples: Abilene, Albany (GA), Albany–Schenectady–Troy, Amarillo, Appleton.
4. *Defense and Energy Single-Sector Metros* (34 CBSAs). One dominant sector—a military installation or oil, gas and petrochemicals—rather than a diversified base, so employment tracks Pentagon budgets or commodity prices. Blue-collar workforces, below-average wages, and Gulf Coast exposure to hurricanes and spills. Examples: Baton Rouge, Beaumont–Port Arthur, Bremerton–Silverdale, Casper, Colorado Springs.
5. *Severe Deindustrialization* (89 CBSAs). Steel, auto, textile, rubber, glass and coal employment collapsing from the early-1980s recessions onward, often around a single dominant employer. High-wage unionized work gives way to services, prisons and warehousing, alongside population loss and an aging workforce. Examples: Akron, Allentown, Altoona, Atlantic City, Binghamton.

**Closed-Form SVD-$\beta$ at $r = 50$ (SB).**

1. *Deindustrialization* (97 CBSAs). Collapse of steel, auto, textile, and machinery employment; loss of unionized blue-collar jobs, population decline, and weak post-2008 recoveries. Examples: Akron, Allentown, Binghamton, Buffalo, Canton.
2. *Anchor Institutions and Regional Hubs* (154 CBSAs). "Eds and meds" as dominant, recession-resistant employers, often alongside state government, serving as service hubs for rural hinterlands. Low unemployment but below-average wages. Examples: Albuquerque, Ann Arbor, Atlanta, Austin, Madison.
3. *Amenity, Tourism and Agriculture* (75 CBSAs). Sun Belt and coastal in-migration destina-

tions driven by retirees and lifestyle migration, together with seasonal agricultural and border economies. Heavy tourism and construction exposure and a severe 2008 housing shock. Examples: Asheville, Bakersfield, Bend, Brownsville, Cape Coral.

4. *Energy and Resource Extraction* (21 CBSAs). Oil, gas, refining, and coal dependence, with boom-bust cycles tied to commodity prices rather than the national business cycle. Examples: Baton Rouge, Beaumont, Casper, Corpus Christi, Houston.
5. *Military and Defense Installations* (16 CBSAs). A dominant Army post, Navy base, or shipyard anchors each metro; fortunes track federal budgets and BRAC rounds rather than market conditions. Young, transient, diverse populations. Examples: Clarksville, Colorado Springs, Fayetteville, Huntsville, Killeen.

**Pretrained Google News Vectors (PW).**

1. *Deindustrialized Manufacturing* (89 CBSAs). Steel, auto, textile, rubber, and machinery dependence followed by sustained decline after 1979. High-wage unionized jobs give way to lower-paying service work, with single-employer vulnerability and sharp exposure to the 1981–82 and 2008–09 recessions. Examples: Akron, Allentown, Buffalo, Canton, Cleveland.
2. *Energy and Resource Extraction* (33 CBSAs). Oil, gas, coal, refining, and petrochemicals anchor these economies, so employment tracks commodity prices rather than the national cycle: the 1980s oil bust, the shale boom of the 2000s and 2010s, and the 2014–16 collapse. Diversification is incomplete. Examples: Baton Rouge, Casper, Corpus Christi, Houston, Lake Charles.
3. *Regional Service Hubs* (109 CBSAs). Small and mid-sized metros where legacy manufacturing and agriculture have given way to hospitals and anchor universities, retail and warehousing along interstate corridors, and tourism. They serve as hubs for multi-county rural hinterlands, with military bases and state government as stabilizers. Examples: Abilene, Asheville, Atlanta, Charlotte, Chattanooga.
4. *Seasonal and Amenity Economies* (64 CBSAs). Tourism, hospitality, and recreation with sharp seasonal swings, retiree and "snowbird" in-migration, and heavy construction and real-estate dependence that left these metros badly exposed to the 2008 housing crash. Agriculture and extraction appear alongside. Examples: Bakersfield, Bend, Brownsville, Cape Coral, Fresno.
5. *University and Government Centers* (68 CBSAs). Flagship universities, "eds and meds," state capitals, federal agencies, national labs, and military bases dominate employment. Unemployment sits below national averages and these metros were cushioned in both the 1980s and 2008 downturns. Examples: Albuquerque, Ann Arbor, Austin, Boston, Boulder.

**LLM Embeddings (LE).**

1. *Agricultural Economies* (25 CBSAs). Agriculture as the economic bedrock, often in high-value specialty crops, with limited diversification. Seasonal low-wage labor, pronounced employment cyclicality, large Latino and immigrant workforces, and above-average unemployment and poverty. Examples: Bakersfield, Fresno, Merced, Modesto, Salinas.
2. *Eds and Meds Regional Hubs* (114 CBSAs). Universities and hospital systems as dominant, recession-resistant employers, following the decline of textiles, timber, steel, and tobacco. Retail, hospitality, and logistics replace blue-collar work, and these metros serve as hubs for rural hinterlands. Examples: Ann Arbor, Asheville, Baltimore, Birmingham, Bloomington.
3. *Federal Anchors and Commodity Exposure* (59 CBSAs). Regional centers combining military bases, laboratories, and defense spending with boom-bust dependence on oil, gas, or agriculture. Healthcare and education are the growth sectors replacing extraction and manufacturing; wages sit below national averages. Examples: Albuquerque, Amarillo, Baton Rouge, Colorado Springs, Corpus Christi.
4. *Heavy Industry and Deindustrialization* (101 CBSAs). Metros dependent in 1979 on steel, autos, textiles, paper, chemicals, or coal, often around a single dominant employer, that suffered plant closures and the early-1980s recession shock. Globalization, automation, and offshoring are the recurring drivers. Examples: Akron, Buffalo, Canton, Chicago, Cleveland.
5. *Diversified Service and Knowledge Economies* (64 CBSAs). A shift from manufacturing and extraction toward services, with "eds and meds" as stable anchors and rising technology and professional-services sectors. In-migration drives growth, and construction exposure made these metros vulnerable in 2008. Examples: Atlanta, Austin, Boston, Boulder, Charlotte.

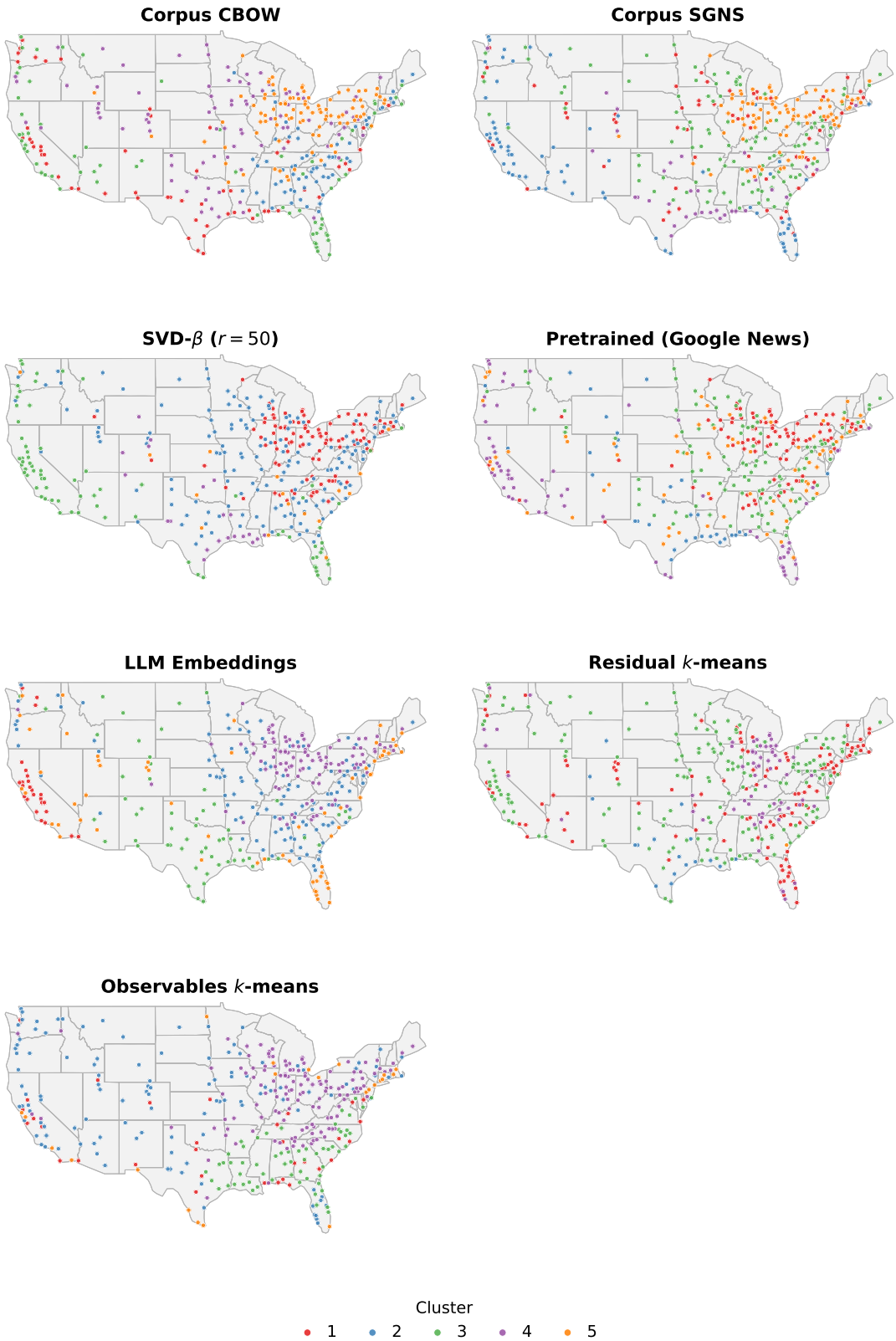
**Residual $k$-means (RK).** This purely outcome-based clustering applies $k$-means with $L = 5$ to the residuals from the pooled AR(2) specification with unit fixed effects (Section 4.1). Unlike the text-based methods, groups are formed to minimize within-group variation in employment dynamics rather than narrative similarity. The resulting clusters are less balanced than the text-based approaches—one group is a singleton and the largest contains 47% of all CBSAs—but an ex-post examination of their composition reveals interpretable economic patterns:

1. *Diversified, Higher-Density Metros* (113 CBSAs, e.g., Atlanta, Austin, Akron, Allentown, Barnstable Town). The densest group by some margin (1970 population density of 138 against 77–79 elsewhere), with moderate manufacturing ($20.6\%$ in 1980) and the second-highest college-educated share ($16.9\%$).
2. *Energy-Dependent* (17 CBSAs, e.g., Houston, Beaumont–Port Arthur, Corpus Christi, Lafayette, Casper). Concentrated in Texas and Louisiana with the lowest manufacturing share of any non-singleton group ($11.8\%$); employment dynamics shaped by oil and gas cycles rather than the national one.
3. *Government-Anchored* (170 CBSAs, e.g., Albuquerque, Baltimore, Ames, Amarillo, Augusta–

Richmond County). The largest group, holding 47% of the sample, with the highest government employment share ($23.4\%$) and the highest military share outside the singleton ($4.9\%$).

4. *Manufacturing-Heavy* (62 CBSAs, e.g., Battle Creek, Bay City, Chattanooga, Anderson, Asheville). The highest manufacturing share of any group ($26.2\%$) and the lowest college-educated share ($13.0\%$).
5. *Military Outlier* (1 CBSA: Hinesville–Fort Stewart, GA). A singleton with $61.0\%$ military and $79.3\%$ total government employment in 1980. Its impulse response is far from the pooled estimate, making it a residual outlier that no other CBSA resembles.

The alignment between these ex-post labels and the text-based cluster themes—particularly the energy and manufacturing archetypes—provides a form of external validation: qualitatively similar economic groupings emerge whether cities are classified by narrative content or by the time-series behavior of their employment.




**Online Appendix Figure OA.8:** Geographic distribution of cluster assignments for 363 CBSAs. Each dot is a CBSA, colored by its cluster. Cluster numbers are assigned within each method, so a color is comparable across CBSAs in one panel but not across panels.

# C Additional Numerical Results

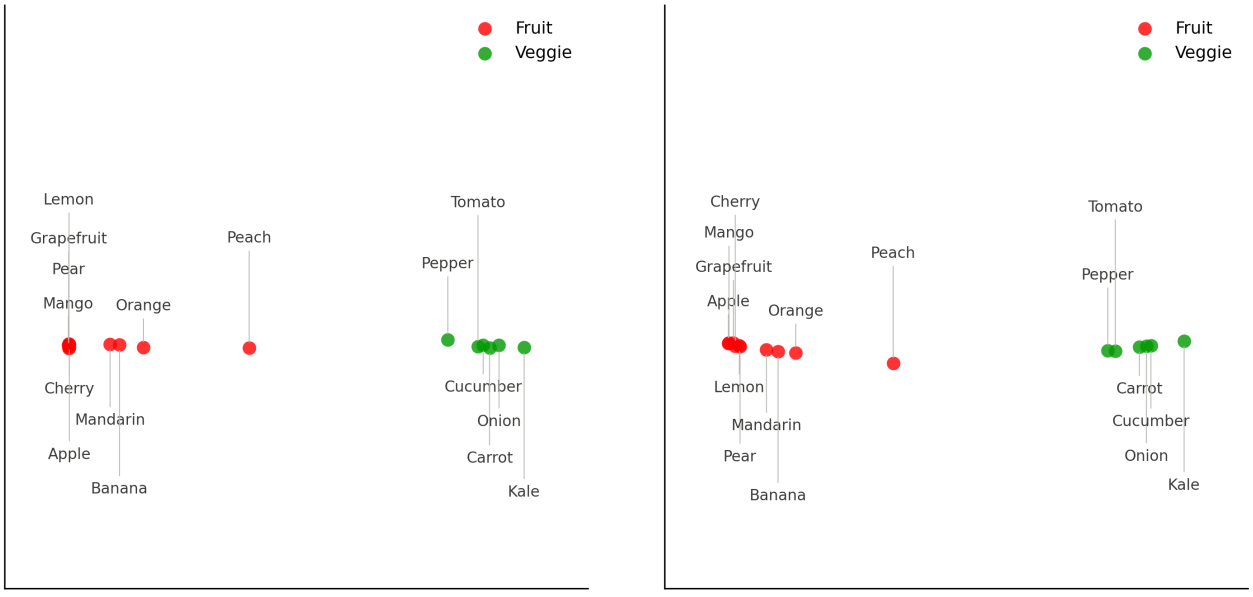
This appendix considers number of alternatives specifications for our numerical illustrations.

## C.1 Large Corpus

Figures OA.9–OA.11 repeat Figures 3–5 on the larger corpus ($D = 1{,}000$, $N_d = 10{,}000$, against $D = 363$ and $N_d = 487$). The comparison to make is panel against its main-text counterpart, and what changes differs by panel.

For SVD-$\beta$ (left) only the sample size changes. We keep the full document as context, so the number of ordered word–context pairs entering $\widehat{R}$ rises from $8.6 \times 10^7$ to $1.0 \times 10^{11}$, a factor of about $1{,}200$, on $56$ times as many tokens. The panel is therefore a pure sample-size comparison, and it is the cleanest picture of Proposition 1 in the paper: at $K = 2$ the cloud collapses onto a line.

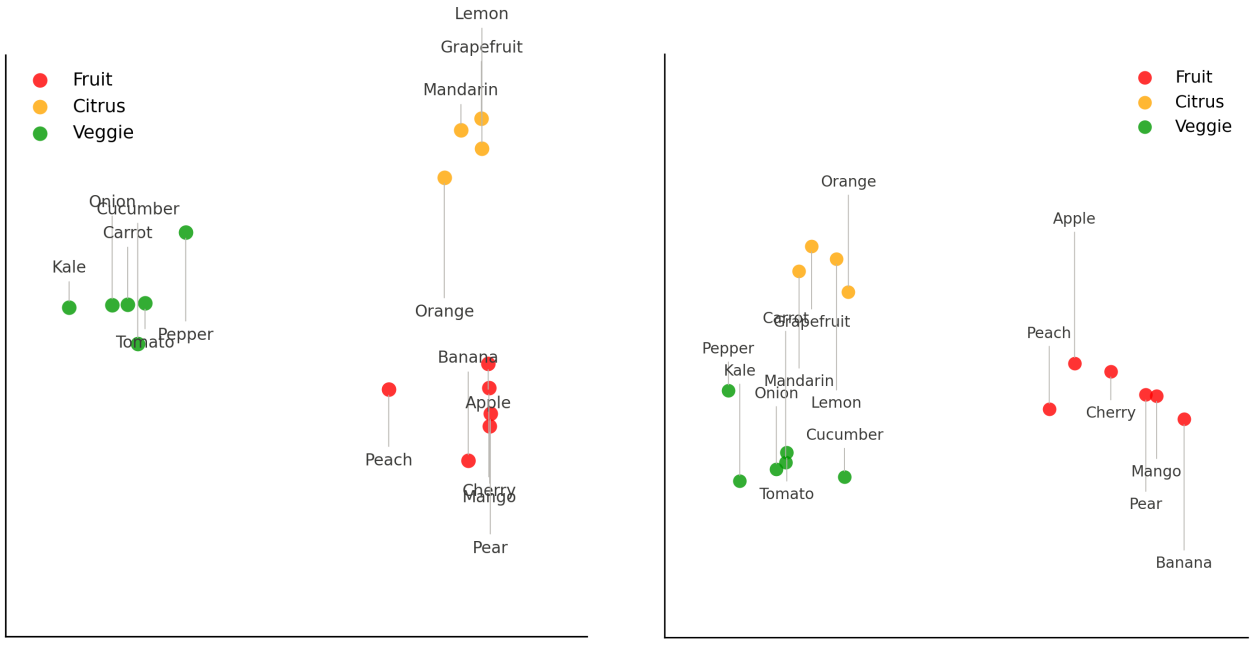
For SGNS (right) it does not. Full-document context is not feasible at this scale. We therefore run $J = 5$ (the default in `gensim`), and the two changes work against each other. `gensim` draws the effective half-window uniformly on $\{1, \ldots, J\}$, so at $J = 5$ each target sees about six contexts: the objective consumes roughly $6.0 \times 10^7$ pairs per epoch, against $8.6 \times 10^7$ in Figure 3. So effectively, the training sample is about $30\%$ *smaller* in pairs compared to its main-text counterpart.




**(a)** SVD-$\beta$, full document

**(b)** SGNS, $J = 5$

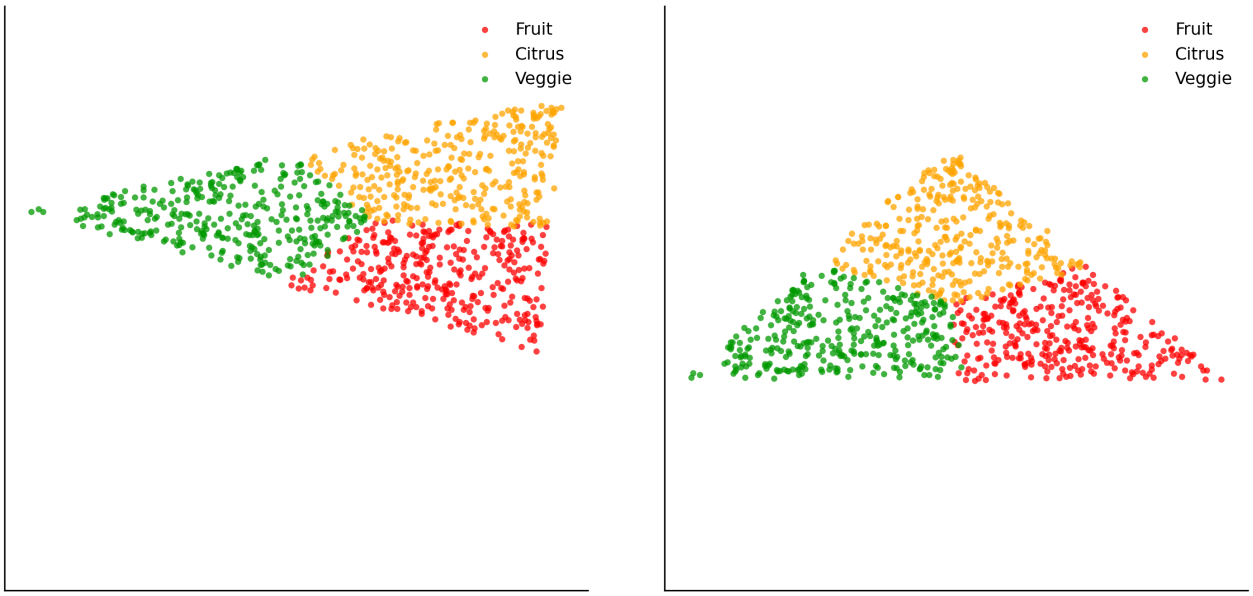
**Online Appendix Figure OA.9:** Word embeddings for Design 1 ($K = 2$) on the larger corpus ($D = 1{,}000$, $N_d = 10{,}000$). SVD-$\beta$ uses the full document as context, as in the main text; SGNS uses $J = 5$, which is what is feasible at this scale. Counterpart of Figure 3. Colors mark the dominant topic in the true $B$. Both panels are drawn at equal aspect: in population the embedding has rank $K - 1 = 1$, so the cloud is one-dimensional and the vertical spread is estimation noise.




**(a)** SVD-$\beta$, full document

**(b)** SGNS, $J = 5$

**Online Appendix Figure OA.10:** Word embeddings for Design 2 ($K = 3$) on the larger corpus. SVD-$\beta$ uses the full document as context; SGNS uses $J = 5$. Counterpart of Figure 4. Both methods separate the three topics. Each panel is projected onto its own leading two principal directions rather than onto the first two raw coordinates: for SVD-$\beta$ the two coincide, since $\beta$ is built from the ordered eigenvectors of $\widehat{R} - \mathbf{1}_V \mathbf{1}_V^\top$, but the CBOW coordinate basis is chosen by the optimizer and is arbitrary, so a raw two-coordinate slice would cut obliquely through the topic plane.




**(a)** SVD-$\beta$, full document

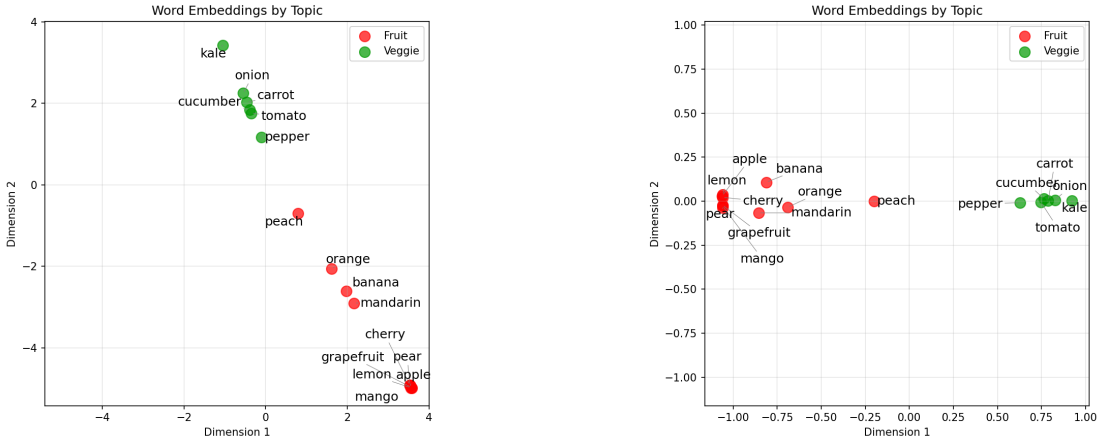
**(b)** SGNS, $J = 5$

**Online Appendix Figure OA.11:** Document embeddings for Design 2 ($K = 3$) on the larger corpus, SVD-$\beta$ at full-document context and SGNS at $J = 5$, colored by dominant topic. Counterpart of Figure 5. With $\alpha = 1$ the topic mixtures are spread across the simplex and the document embeddings fill the triangle spanned by the three topic centroids, as Proposition 3 predicts.

## C.2 Sparse Topic Mixtures: Dirichlet$(\alpha = 0.01)$

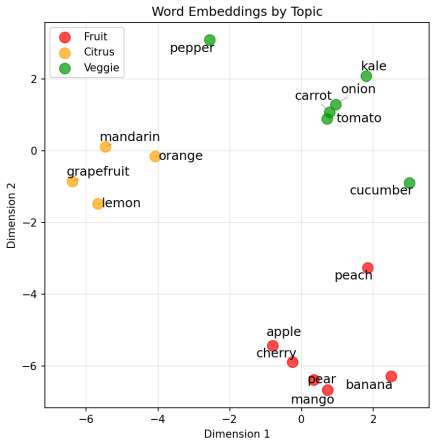
The simulations in Section 3 use a symmetric Dirichlet$(\alpha = 1)$ prior, which is uniform on the topic simplex. As a contrast, I repeat both designs with $\alpha = 0.01$ (Holding $D = 363$ and $N_d = 487$). Under Dirichlet$(0.01)$, almost all mass concentrates at the simplex vertices: most documents are nearly pure-topic, with topic mixtures $\Theta_{\bullet d}$ close to a single $e_k$. By Proposition 3, document embeddings $\mu_d = C\,\Theta_{\bullet d}$ then concentrate at the topic centroids $\{c_k\}$ rather than filling the centroid simplex.

Figures OA.12–OA.15 show the resulting embeddings. The $B$ matrices and pipeline are identical to Section 3; only the Dirichlet concentration parameter differs.

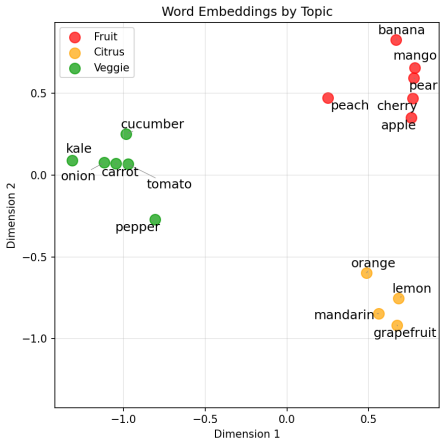



**(a)** CBOW **(b)** SVD-$\beta$

**Online Appendix Figure OA.12:** Word embeddings for Design 1 ($K = 2$) under Dirichlet$(\alpha = 0.01)$.



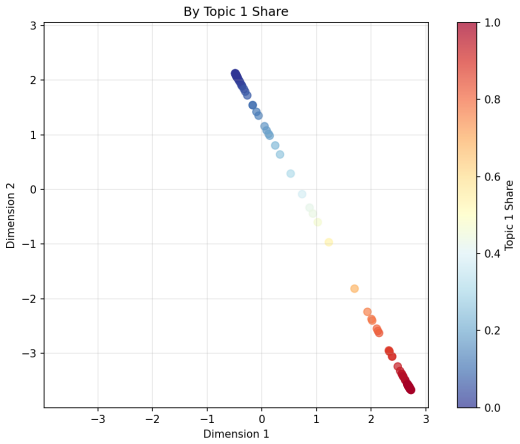

**(a)** CBOW



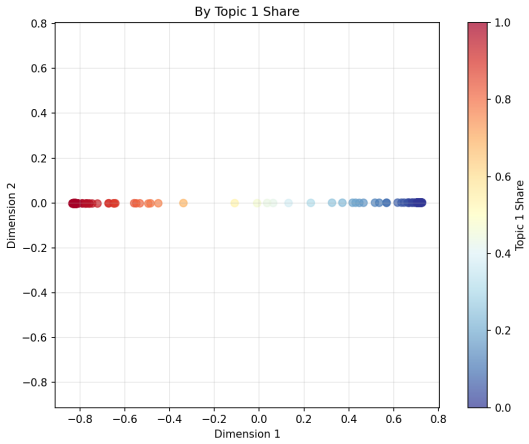

**(b)** SVD-$\beta$

**Online Appendix Figure OA.13:** Word embeddings for Design 2 ($K = 3$) under Dirichlet($\alpha = 0.01$).

The sparse-prior case illustrates the boundary regime for Lemma 2: when $\theta_{d,k^*(d)} \approx 1$, the dominant-topic threshold $\theta_{d,k^*(d)} > 1 - \eta$ is met for essentially every document, and $k$-means recovers the dominant-topic partition. At $\alpha = 0.01$, $\Pr(\theta_{d,k^*} > 0.5) \approx 1$ for both $K = 2$ and $K = 3$, consistent with the visible vertex concentration.

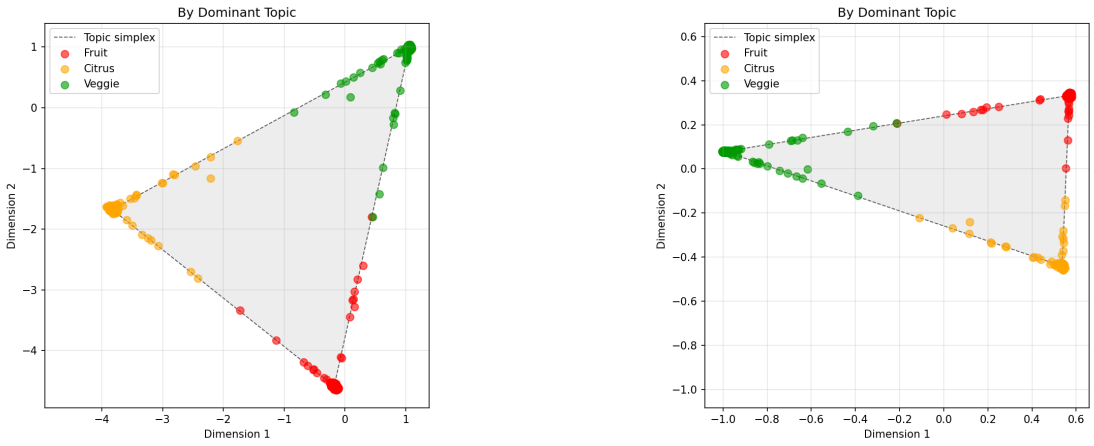



**(a)** CBOW




**(b)** SVD-$\beta$

**Online Appendix Figure OA.14:** Document embeddings for Design 1 ($K = 2$) under Dirichlet($\alpha = 0.01$). Documents cluster at the two endpoints of the line segment $[c_1, c_2]$ rather than filling it; color indicates the share of topic 1.




**(a)** CBOW **(b)** SVD-$\beta$

**Online Appendix Figure OA.15:** Document embeddings for Design 2 ($K = 3$) under Dirichlet($\alpha = 0.01$), colored by dominant topic. With most documents near-pure on a single topic, the dominant-topic partition is well-separated and is recovered cleanly by both embeddings ($k$-means with $K = 3$). The shaded region is the topic simplex $\mathrm{conv}\{c_1, c_2, c_3\}$; documents concentrate at its vertices rather than filling it, which is the contrast with Figure 5.

## C.3 Interior Archetypes ($L < K$)

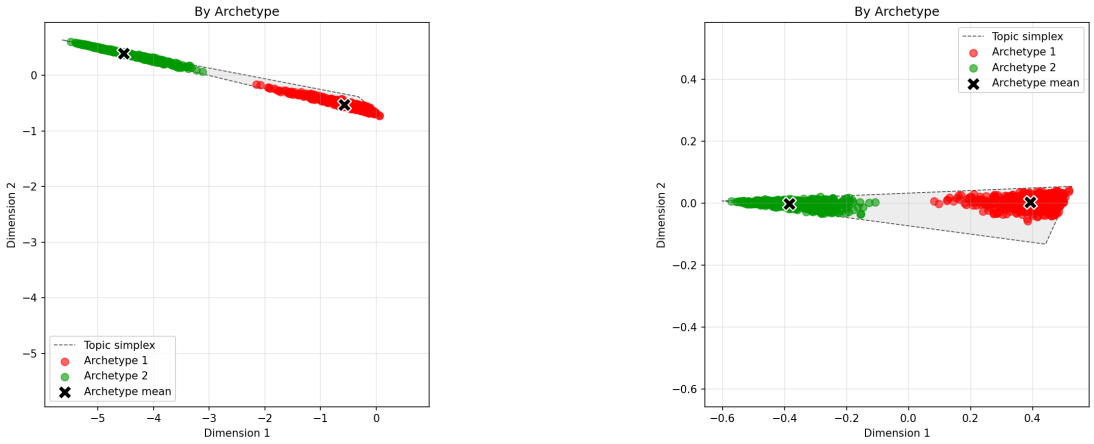
The simulations above illustrate the case $L = K$, where the number of clusters in document space matches the number of topics. Proposition 4 covers the more general case in which documents concentrate around $L$ *archetype mixtures* $\pi_1^*, \ldots, \pi_L^* \in \Delta_{K-1}$ that need not coincide with the simplex vertices. The empirical application in Section 4 sits in this regime: $k$-means on the CBSA corpus recovers $L = 5$ clusters whose centroids are interior mixtures of an unknown number of latent topics.

To illustrate this scaling, I repeat Design 2 ($K = 3$ topics, same $B$ matrix as in Section 3) but replace the Dirichlet($\alpha = 1$) prior on $\Theta_{\bullet d}$ with a two-archetype mixture. Each document is assigned uniformly to one of two archetypes,

$$\pi_1^* = (0.65,\ 0.25,\ 0.10), \qquad \pi_2^* = (0.10,\ 0.10,\ 0.80),$$

and its topic mixture is drawn from a tight Dirichlet centered on the assigned archetype, $\Theta_{\bullet d} \sim \text{Dirichlet}(20 \cdot \pi_\ell^*)$. Archetype 1 is fruit-heavy with some citrus; archetype 2 is nearly pure vegetable. Neither lies at a simplex vertex. The word geometry is not carried over unchanged. By Theorem 1 the word metric is $\Sigma_\Theta$-weighted. Figure OA.16 illustrates. The shaded region is the topic simplex $\text{conv}\{c_1, c_2, c_3\}$, computed from the true $B$ and the estimated word embeddings and projected onto the same two coordinates as the documents; Proposition 3 places every expected document embedding inside it. Crosses mark the two archetypes' expected embeddings, $\sum_k \pi_{\ell k}^* c_k$. Both sit in the interior, which is the $L < K$ regime of Proposition 4. The simplex is far flatter than in Figure OA.15 even though $B$ is the same, for the reason given above: $\Sigma_\Theta$ is nearly rank one here. That flattening is present in both panels and in the full $K$ coordinates, not an artifact of the projection.

The archetype design illustrates Proposition 4 in the $L = 2 < K = 3$ regime: cluster identity in embedding space is governed by the archetype assignment, not by the dominant-topic vertex. This mirrors the empirical situation in Section 4, where the choice of $L = 5$ is governed by the resolvable archetypes in the corpus and need not equal the (unobserved) number of latent topics.

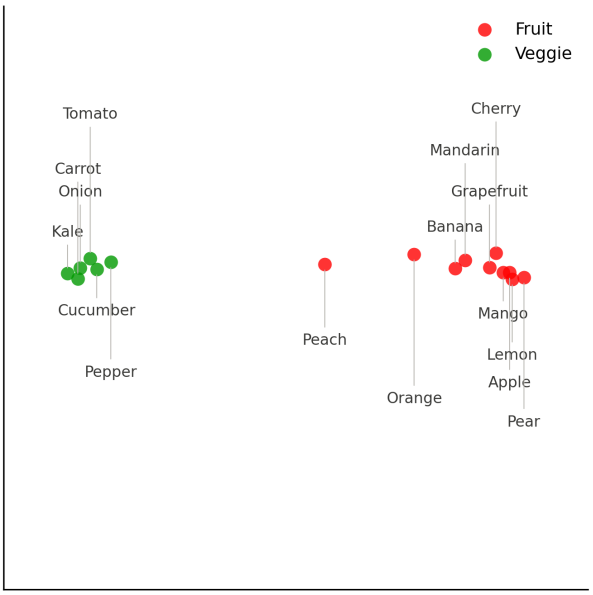



**(a)** CBOW **(b)** SVD-$\beta$

**Online Appendix Figure OA.16:** Document embeddings colored by archetype label. The shaded region is the topic simplex $\text{conv}\{c_1, c_2, c_3\}$, computed from the true $B$ and the estimated word embeddings and projected onto the same two coordinates as the documents. Crosses mark the two archetypes' expected embeddings, $\sum_k \pi^*_{\ell k}\, c_k$.

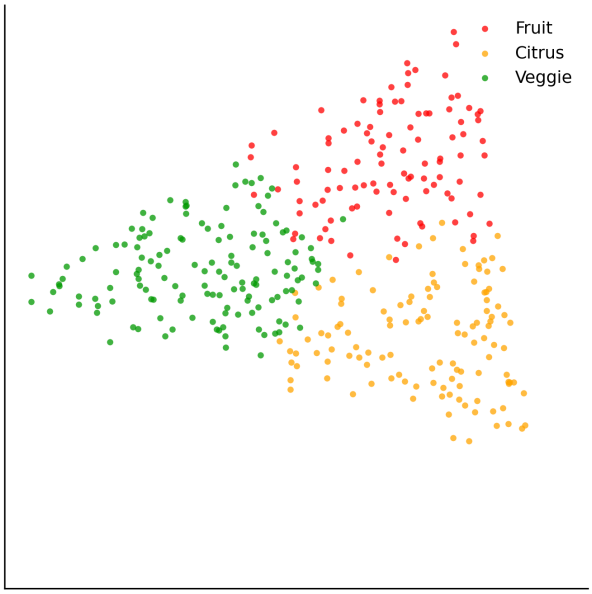
## C.4 CBOW

Here, I depict the counterpart of Figures 3-Figure 5 using continuous bag-of-words (CBOW) embeddings.




**(a)** Word embeddings, Design 1 ($K = 2$)


**(b)** Word embeddings, Design 2 ($K = 3$)

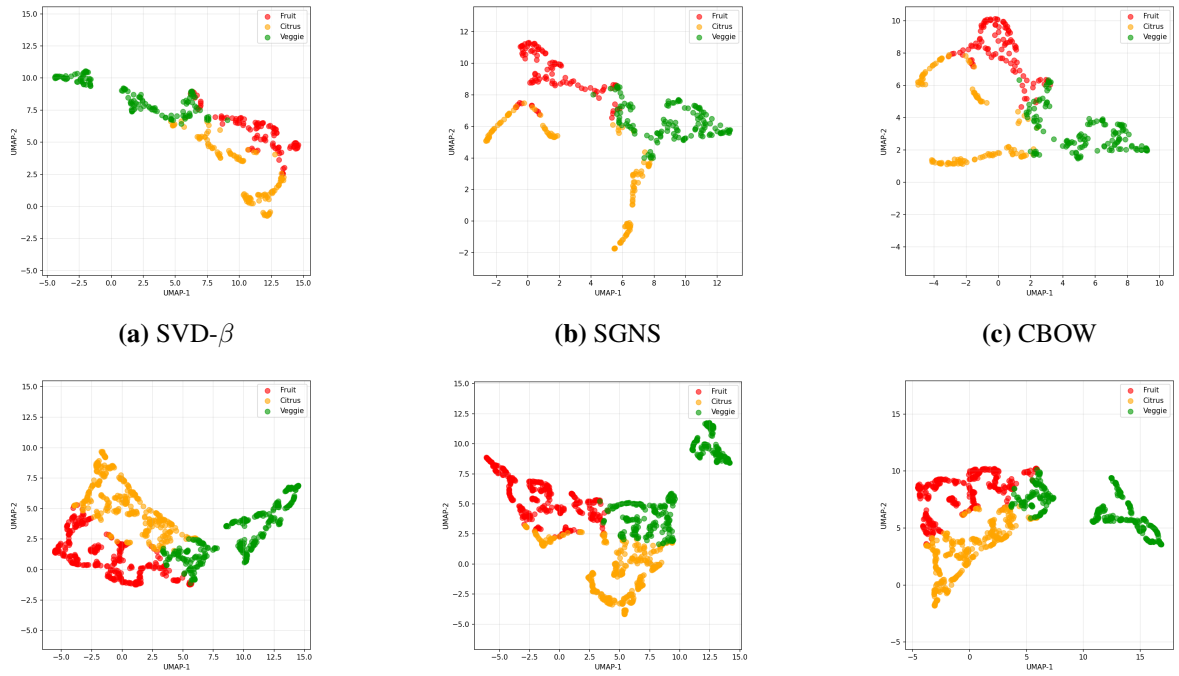



**(c)** Document embeddings, Design 2 ($K = 3$)

**Online Appendix Figure OA.17:** Estimated embedding using CBOW on applicaton-sized corpus. Colors mark the dominant topic in the true $B$.

## C.5 UMAP Projections

Figure 5 projects the 3-dimensional embeddings to two dimensions by retaining the first two coordinates. The empirical application in Section 4 instead uses UMAP [McInnes et al., 2018], a nonlinear dimensionality-reduction method, because the application's embeddings live in 50 dimensions (closed-form SVD-$\beta$ and both corpus-trained arms), 300 (pretrained Google News) or 3,072 (the LLM embeddings). For visual parallelism with the application figures, this section repeats the Design 2 ($K = 3$) embeddings under the same UMAP projection used in Section 4. UMAP's nonlinear transformation can curve or warp the simplex but generally preserves the cluster structure.




**(a)** SVD-$\beta$ **(b)** SGNS **(c)** CBOW

**(d)** SVD-$\beta$, large corpus **(e)** SGNS, large corpus **(f)** CBOW, large corpus

**Online Appendix Figure OA.18:** Document embeddings for Design 2 ($K = 3$), UMAP-projected and colored by dominant topic. Top row: application scale ($D = 363$, $N_d = 487$, full document as context), the same embeddings Figure 5 projects linearly. Bottom row: the larger corpus ($D = 1{,}000$, $N_d = 10{,}000$, $J = 5$), the counterpart of Figure OA.11. Cluster structure is preserved under UMAP at both scales.

## C.6 Rank of the SGNS Target under a Sparser Prior

Table 2 draws every column of $B$ and every topic mixture from a symmetric Dirichlet$(\alpha = 1)$, which puts no mass at the boundary of either simplex: each topic loads on every word and each document on every topic. Table OA.3 repeats that grid at $\alpha = 0.1$, where both are more concentrated.

Two things change. The level falls sharply: $\pi_{K-1}$ runs from $0.68$ to $0.89$, against $0.91$ to $0.99$ at $\alpha = 1$, so a rank-$(K-1)$ approximation now misses between a tenth and a third of the spectral mass of $\log R$. And the invariance to $V$ breaks for moderately sized $K$. This suggests the cost of the rank constraint grows with the vocabulary under a sparse prior.

Both follow from a larger departure from independence. The diagnostic $\varrho = \|R - \mathbf{1}_V \mathbf{1}_V^\top\|_\infty$ stays below $0.6$ in every cell at $\alpha = 1$, but ranges from $0.84$ to $6.5$ here and exceeds one in $25$ of the $30$ cells. Remark 3 expands $\log R = \beta\beta^\top + O(\varrho^2)$ under $\varrho < 1$, so that expansion does not cover most of this table.

| $V$ | $K = 2$ | $K = 3$ | $K = 5$ | $K = 10$ | $K = 25$ | $K = 50$ |
|---|---|---|---|---|---|---|
| 100 | 0.677 | 0.693 | 0.728 | 0.779 | 0.831 | 0.890 |
| 250 | 0.677 | 0.692 | 0.715 | 0.743 | 0.788 | 0.827 |
| 500 | 0.676 | 0.691 | 0.711 | 0.731 | 0.755 | 0.808 |
| 1,000 | 0.677 | 0.692 | 0.705 | 0.721 | 0.726 | 0.786 |
| 2,500 | 0.676 | 0.691 | 0.704 | 0.712 | 0.702 | 0.747 |

**Online Appendix Table OA.3:** Share of the spectral mass of $\log R$ carried by its leading $K-1$ eigenvalues, $\pi_{K-1} = \sum_{i \le K-1} |\lambda_i| / \sum_i |\lambda_i|$, at $\alpha = 0.1$. Each column of $B$ and each topic mixture $\Theta_{\bullet d}$ is drawn from a symmetric Dirichlet$(\alpha)$; $R$ is then formed in population from $B$ and $G$, so no corpus is sampled. Means over 8 draws; the largest standard deviation in any cell is $0.016$. The negative-sampling shift contributes one further eigenvalue, of size $V \log \nu$, and is excluded.

# References


David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, volume 14, pages 617–624, 2001.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, pages 2177–2185, 2014.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013a.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. word2vec: Tool for computing continuous distributed representations of words. Google Code archive, 2013b. URL `https://code.google.com/archive/p/word2vec/`. Pre-trained `GoogleNews-vectors-negative300` model.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, 2013c.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.