Labor Market Power under Directed Search

Xincheng Qiu* Yu Yi† Shengxing Zhang‡

March 26, 2026

Abstract

We study labor market power in a directed-search model with segmented local labor markets and a finite number of firms. Workers direct applications across posted wage–vacancy pairs, hiring is frictional, and production exhibits decreasing returns. We derive closed-form wage markdown formulas showing that the wedge between wages and marginal product depends on an accessibility-adjusted market share and the elasticity of matching. By contrast, the vacancy-posting first-order condition is identical to the competitive-search benchmark, so finite-firm market power distorts wages but not the surplus condition governing vacancies. With heterogeneous firms, more productive employers attract larger applicant pools, operate at higher tightness, and generate dispersion in firm-level tightness that shifts the Beveridge curve outward for a given vacancy stock. Quantitatively, we map the model to the cross-location distribution of employer counts, concentration, mean firm size, application elasticities, steady-state unemployment, and aggregate tightness. The calibrated economy matches these moments closely and uses the Beveridge curve as an out-of-sample validation of the model's reallocation mechanism. The same markdown mechanism that shifts the Beveridge curve also flattens the wage Phillips curve: oligopsony absorbs demand shocks through the markdown margin rather than through wage adjustments, so that the calibrated economy exhibits a nearly flat wage–unemployment trade-off even though the competitive-search benchmark produces a steep one.

Keywords: directed search, oligopsony, wage markdowns, firm heterogeneity, Beveridge curve, Phillips curve.

JEL Codes: E24, E31, J31, J63, L13.

*Peking University, Guanghua School of Management.

†Nankai University.

‡Carnegie Mellon University, Tepper School of Business.

1 Introduction

Standard search models are typically competitive in the sense that an individual firm takes the value of search as given. In random-search environments, workers draw from an exogenous wage-offer distribution. In competitive directed-search models, a continuum of firms implies that no single employer affects workers' outside option. These frameworks are useful for studying matching frictions, but they are less well suited to analyzing labor market power in environments where firms post both wages and vacancies and workers direct applications across employers. This omission is increasingly important in light of evidence on employer concentration and wage markdowns. The question we study is therefore straightforward: how does labor market power operate when local labor markets are segmented, workers direct applications toward posted wage–vacancy menus, and each market contains only finitely many employers?

We develop a directed-search model with a continuum of segmented locations. Each location contains a finite number of firms and a mass of workers who search only within that location. Firms post wage–vacancy menus, workers choose where to apply according to a logit application rule, and matches form through a constant-returns matching technology. Production exhibits decreasing returns. The finite number of firms is the key departure from the standard directed-search benchmark: an individual employer internalizes that its wage posting affects not only its own applicant pool but also the value of search at rival firms.

The paper delivers three main theoretical results. **First**, we derive closed-form sufficient-statistics formulas for firm-specific wage markdowns. The markdown depends on an accessibility-adjusted market share and on the elasticity of the matching function, so larger or less contestable firms pay lower wages relative to marginal product. **Second**, we establish a *vacancy invariance* result: the first-order condition for vacancy posting is identical to the competitive-search benchmark even when wages reflect oligopsony power. In our environment, market power distorts wages directly, but it does not alter the surplus condition governing vacancies. **Third**, with heterogeneous firms, more productive employers attract more applicants, post more vacancies, and operate at higher tightness. Because tightness is no longer equalized across firms, aggregate matching efficiency falls for a given vacancy stock, generating a structural channel through which the Beveridge curve can shift even when the matching technology is unchanged.

The paper also provides a quantitative implementation of the model. We map one model market to one segmented local labor market and discipline the distribution of firm counts across locations using Berger et al. (2022). We then choose productivity dispersion and levels to match concentration and mean firm size, discipline worker substitution across firms using application elasticities, and choose matching efficiency and vacancy costs to match steady-state unemployment and aggregate tightness. A key feature of the quantitative block is that cross-location labor supply responds to

the promised value of search in steady state, while Beveridge-curve exercises hold location labor supply fixed and are treated as validation rather than calibration targets. This design keeps distinct the paper's long-run allocation of workers across markets and its short-run vacancy–unemployment implications.

The calibrated economy matches its targets closely: the model produces a payroll-weighted HHI of 0.114 (target 0.110), mean firm size of 22.8 (target 22.8), an application elasticity of 0.756 (target 0.77), steady-state unemployment of 5.9 percent (target 6 percent), and aggregate tightness of 69.0 (target 70). The employment-weighted average wage markdown is $w/MPL = 0.40$. Two counterfactual exercises then use the Beveridge curve to disentangle the roles of productivity dispersion and market structure. Doubling within-location productivity dispersion shifts the curve outward at a given unemployment rate, lowering aggregate matching efficiency through greater tightness heterogeneity. By contrast, a sharp increase in concentration—halving the average number of firms per location so that the payroll-weighted HHI rises from 0.114 to 0.423—shifts the curve only modestly when per-location labor supply is held fixed. The quantitative dominance of the productivity-dispersion channel over the concentration channel suggests that local firm counts shift the Beveridge curve primarily through the equilibrium reallocation of workers across markets, a margin shut down by construction in the fixed-labor-supply exercise. Disaggregating by location size reveals a monotone ordering: markets with fewer firms display steeper, outward-displaced Beveridge curves, with monopoly locations producing near-zero unemployment because the monopsonist's extreme markdown induces vacancy posting sufficient to match nearly all available workers.

We complement the quantitative analysis with a discussion of the model's testable predictions and supporting evidence. At the firm level, the markdown formula links wage markdowns to application shares, vacancy yields, and local concentration—predictions that align with evidence documenting that markdowns rise with firm size and productivity (Yeh et al., 2022; Seegmiller, 2021; Mertens and Mottironi, 2025). At the market level, we show that U.S. states with above-median employer concentration display an outward-shifted Beveridge curve relative to less concentrated states, consistent with the model's prediction that monopsonistic wage-setting generates higher vacancy rates at every unemployment level. Event-study evidence on mergers and employer consolidation further supports the link between reductions in the local number of firms and depressed wages (Prager and Schmitt, 2021; Arnold, 2019; Dodini et al., 2024).

The same markdown mechanism that shifts the Beveridge curve also flattens the wage Phillips curve. When productivity rises and unemployment falls, the competitive-search benchmark predicts a steep decline in wages—roughly a 22 percent change per unit of unemployment. Under oligopsony, firms absorb the productivity gain almost entirely through the markdown margin, leaving the average wage nearly unchanged. The calibrated oligopsony Phillips curve is essentially flat. This result provides a structural explanation for the well-documented flattening of the Phillips curve (Blanchard,

2016; Hazell et al., 2022) that does not rely on expectations anchoring or nominal rigidities. Instead, labor market power operates as a real channel: the markdown wedge between marginal product and wages buffers wage responses to demand shocks, and the size of this buffer rises with employer concentration. The model's cross-location Phillips curves map directly to the cross-state estimates in Hazell et al. (2022), who find a structural slope $\kappa \approx 0.006$ using state-level nontradeable inflation data and interpret this as requiring a large amount of real rigidity—precisely the role played by the markdown margin in our framework.

Literature. The paper bridges two literatures. In the labor market power literature, Berger et al. (2022) study oligopsonistic labor markets with a finite number of firms and estimate substantial wage markdowns and welfare costs, but their framework abstracts from search frictions and vacancy posting. Yeh et al. (2022) document concentration and wage effects using vacancy data. Jarosch et al. (2023) analyze granular labor markets and wage inequality, while Gottfries and Jarosch (2025) introduce monopsony in a *random*-search environment. A growing empirical literature documents that local employer concentration depresses wages (Azar et al., 2022, 2020; Benmelech et al., 2022) and that employer consolidation causally reduces worker earnings (Prager and Schmitt, 2021; Arnold, 2019); our framework provides a structural interpretation of these findings by linking observed concentration to the sufficient-statistics markdown formula. None of these papers study wage–vacancy posting under directed search with finitely many employers, and none link labor market power to the slope of the Phillips curve. In the Phillips curve literature, Blanchard (2016) surveys the flattening of the U.S. Phillips curve and Hazell et al. (2022) estimate its slope using cross-state variation in nontradeable inflation, finding a structural slope $\kappa \approx 0.006$ that they interpret as requiring substantial real rigidity; Beraja et al. (2019) estimate regional wage Phillips curves using scanner data and document cross-region slope heterogeneity. These papers attribute the flatness primarily to expectations anchoring and nominal rigidity rather than to market structure. Our framework provides a complementary channel—the markdown margin—and a concrete prediction that the cross-state slope should covary with employer concentration. In the directed-search literature, Moen (1997), Burdett et al. (2001), Lester (2010), and Wright et al. (2017) provide the competitive benchmark in which a continuum of firms eliminates strategic effects on workers' outside options.

Our paper is closest to Hurst et al. (2025), who also combine directed search, finite heterogeneous firms, and worker-specific tastes over employers. Relative to that paper, our contribution is analytical and static rather than dynamic and policy-oriented. We derive closed-form markdown formulas and a vacancy-invariance result that isolate the roles of accessibility-adjusted market share and matching elasticity, and we use those results to study the Beveridge-curve implications of firm heterogeneity. By contrast, Hurst et al. (2025) solve a richer dynamic model with putty-clay capital,

worker heterogeneity, and policy counterfactuals. Our framework also nests a no-search-friction limit that yields a reduced-form oligopsony wedge analogous to Berger et al. (2022).

Roadmap. Section 2 introduces the mechanism in the simplest symmetric market. Section 3 contains the core analytical block of the paper: it states the general environment and equilibrium, derives the competitive-search and finite-firm results, and develops the heterogeneous-firm and Beveridge-curve implications that build intuition for the theory. Section 4 presents the quantitative analysis: calibration, baseline fit, counterfactual Beveridge-curve exercises, and wage Phillips curve implications. Section 5 discusses empirical implications and supporting evidence, including the connection between labor market power and the flattening Phillips curve. Section 6 concludes.

2 Example: Homogeneous Firms

This section introduces the core mechanism in the simplest finite-firm market. The objective is to show, in a symmetric environment with no firm heterogeneity, how a firm's wage and vacancy choices affect its applicant pool and therefore its equilibrium markdown. The key assumptions are a segmented location, a finite number of ex ante identical firms, directed applications, and a constant-returns matching technology. The section contributes a transparent benchmark that makes the strategic channel visible before the general model introduces heterogeneous firms and inputs.

Consider a representative segmented location with a unit mass of identical workers, each supplying one unit of labor, and $M \geq 2$ identical firms indexed by $i = 1, \ldots, M$. Firm i posts wage w_i and vacancies v_i, pays a linear vacancy cost cv_i, and produces zh_i when it hires h_i workers. Matches are generated by a constant-returns matching function $h_i = m(l_i, v_i)$, where l_i is the mass of applicants attracted by firm i. By constant returns, the job-finding probability of an applicant to firm i can be written as $m(\theta_i, 1)$, where

$$\theta_i = \frac{v_i}{l_i} \tag{2.1}$$

is firm i's market tightness.

Timing is as follows. First, firms simultaneously post (w_i, v_i). Second, workers observe all postings and direct one application to one firm. Third, matching occurs, wages are paid to matched workers, and production takes place.

Given postings, a worker who applies to firm i obtains expected income $w_i m(\theta_i, 1)$. A (pure-strategy) equilibrium in this location is a collection $\{(w_i, v_i, l_i)\}_{i=1}^{M}$ such that:

$$l_i > 0 \quad \text{only if} \quad w_i m(\theta_i, 1) = \max_{k=1,\ldots,M} w_k m(\theta_k, 1), \qquad \sum_{i=1}^{M} l_i = 1, \tag{2.2}$$

and each firm chooses (w_i, v_i) to maximize profits, taking rivals' postings as given and recognizing that applicant demand is induced by worker optimality:

$$\Pi_i = (z - w_i)m(\theta_i, 1)l_i - cv_i. \tag{2.3}$$

Because firms are ex ante identical, we focus on a symmetric equilibrium with $w_i = w$, $v_i = v$, and $l_i = 1/M$ for all i.

Comparative statics. In a symmetric equilibrium, a unilateral increase in either v_i or w_i makes firm i more attractive to workers and increases its applicant pool. Differentiating the worker-allocation condition around the symmetric allocation yields

$$\frac{\partial l}{\partial v} = \frac{M-1}{M\theta}, \qquad \frac{\partial l}{\partial w} = \frac{(M-1)}{\varepsilon M^2 w}, \tag{2.4}$$

where $\varepsilon = \frac{\theta m'(\theta,1)}{m(\theta,1)}$ is the elasticity of the job-finding probability.

The firm's first-order conditions imply

$$w = \left[1 - \frac{M\varepsilon}{M - (1-\varepsilon)}\right] z, \qquad cv = \varepsilon y, \tag{2.5}$$

where $y = zm(\theta, 1)l$ is output at the representative firm. The wage expression shows directly how a finite number of firms generates a markdown even though workers direct search toward posted terms. The vacancy condition anticipates the general invariance result below: despite wage markdowns, vacancy posting is still pinned down by the elasticity of the matching technology and the marginal product of labor.

Proposition 1 (Existence and uniqueness in the homogeneous example)**.** *Suppose $\theta \mapsto m(\theta, 1)$ is continuously differentiable, strictly increasing, strictly concave, and satisfies $m(0, 1) = 0$. Let $\varepsilon(\theta) \equiv \theta m_\theta(\theta, 1)/m(\theta, 1) \in (0, 1)$, and assume*

$$zm_\theta(0, 1) > c > \lim_{\theta\to\infty} zm_\theta(\theta, 1). \tag{2.6}$$

Then the economy in this section admits a unique symmetric pure-strategy equilibrium. All firms are active, so $l_i = 1/M$, $v_i = \theta^*/M$, and $w_i = w^*$ for every i, where θ^* is the unique solution to

$$zm_\theta(\theta^*, 1) = c \tag{2.7}$$

and

$$w^* = \left[1 - \frac{M\varepsilon(\theta^*)}{M - (1 - \varepsilon(\theta^*))}\right] z. \tag{2.8}$$

The proposition makes the benchmark well posed. Concavity of the matching technology pins down a unique tightness, the wage first-order condition then pins down a unique markdown, and the equilibrium cannot leave a subset of ex ante identical firms inactive.

3 Full Model

This section states the general environment used in the analytical and quantitative results. The objective is to extend the example in Section 2 to heterogeneous firms, multiple inputs, and imperfect substitution across employers while keeping the directed-search structure explicit. The key assumptions are segmented locations, firm commitment to wage–vacancy postings, constant-returns matching within each input market, and decreasing returns in production. The section contributes the equilibrium concept that underlies both the competitive-search benchmark and the finite-firm results in Section 3.4.

3.1 Environment

The economy consists of a continuum of locations indexed by $\ell \in [0, 1]$. Each location contains one industry and a finite set of firms, indexed by $i = 1, \ldots, M_\ell$. Locations are completely segmented: workers residing in location ℓ direct applications only to firms in ℓ, and firms in one location do not recruit from another. To simplify notation, we characterize equilibrium within a generic location and suppress the location index until Section 4, where cross-location heterogeneity in M_ℓ is reintroduced explicitly.

Within a location, there are N search-relevant inputs indexed by $j = 1, \ldots, N$. For each type j, a unit mass of prospective suppliers chooses among the firms in that location. Firm i has productivity z_i and decreasing-returns technology

$$y_i = f(\{x_{ij}\}_j) = \prod_{j=1}^{N} x_{ij}^{\alpha_j}, \qquad 0 < \sum_j \alpha_j < 1. \tag{3.1}$$

Gross output is $z_i y_i$. Input j hired by firm i is

$$x_{ij} = m_j(\theta_{ij}, 1) l_{ij}, \qquad \theta_{ij} = \frac{v_{ij}}{l_{ij}}, \tag{3.2}$$

where l_{ij} is the mass of type-j applicants directed to firm i, v_{ij} is the number of type-j vacancies posted by firm i, and $m_j(\theta_{ij}, 1)$ is the job-finding probability for an applicant. In the quantitative application below we focus on one local labor market and one competitive capital input, but the analytical formulas are easier to present in this more general notation.

Firms simultaneously post type-specific wage–vacancy pairs $\{w_{ij}, v_{ij}\}_{j=1}^{N}$. A type-$j$ worker who applies to firm i receives expected utility

$$u_{ij} = m_{ij} w_{ij} + (1 - m_{ij}) b_j, \tag{3.3}$$

where $m_{ij} \equiv m_j(\theta_{ij}, 1)$ and b_j is the payoff from not matching. The parameter $\beta > 0$ governs how sensitively workers reallocate applications in response to expected-payoff differences across firms. Given postings, type-j workers choose firms according to the logit shares

$$l_{ij} = \frac{\exp(\beta u_{ij})}{\sum_{k=1}^{M} \exp(\beta u_{kj})}, \qquad \sum_{i=1}^{M} l_{ij} = 1. \tag{3.4}$$

Firm i chooses $\{w_{ij}, v_{ij}\}_{j=1}^{N}$ to maximize

$$\Pi_i = z_i f(\{m_{ij}(\theta_{ij}, 1) l_{ij}\}_j) - \sum_j w_{ij} m_{ij}(\theta_{ij}, 1) l_{ij} - c \sum_j v_{ij}. \tag{3.5}$$

3.2 Equilibrium

The timing within a location is the natural extension of the example. First, firms simultaneously post $\{w_{ij}, v_{ij}\}_j$. Second, each type-j supplier observes all postings in the location and chooses one firm. Third, matching occurs independently across input markets, production takes place, and payoffs are realized.

A finite-firm equilibrium in a location is a collection

$$\left\{ \{w_{ij}, v_{ij}, l_{ij}, x_{ij}\}_{j=1}^{N} \right\}_{i=1}^{M}$$

such that:

1. for each type j, worker allocation satisfies the logit demand system above;
2. for each (i, j), consistency requires $\theta_{ij} = v_{ij}/l_{ij}$ and $x_{ij} = m_j(\theta_{ij}, 1) l_{ij}$;
3. for each firm i, the posting vector $\{w_{ij}, v_{ij}\}_j$ maximizes Π_i, taking rivals' postings as given and internalizing the induced response of applicant masses through the logit system.

This is the equilibrium notion used in Section 3.4. The competitive-search benchmark below keeps the same primitives but removes the firm's strategic effect on workers' outside option by taking the continuation value of search as given.

3.3 Competitive-Search Benchmark

With a continuum of firms, each individual firm is atomistic and therefore takes the logit denominator $\int_0^1 \exp(\beta u_{kj})\, dk$ as given. Relative to the finite-firm equilibrium defined above, the only change is that a single firm no longer internalizes its effect on the expected payoff available at other firms. The competitive-search first-order conditions are therefore a useful benchmark for isolating the role of oligopsonistic strategic interactions.

Proposition 2 (Input price markdown under competitive search)**.**

$$\frac{w_{ij}}{MPL_{ij}} = \frac{(1-\varepsilon_{ij})\beta w_{ij} m_{ij}}{1+\beta(w_{ij}-b_j)m_{ij}\varepsilon_{ij}+(1-\varepsilon_{ij})\beta w_{ij} m_{ij}}, \tag{3.6}$$

where $MPL_{ij} = \frac{\alpha_j z_i y_i}{m_{ij} l_{ij}}$ *and* $\varepsilon_{ij} = \frac{d\log m_{ij}}{d\log \theta_{ij}}$.

Proposition 3 (Vacancy posting under competitive search)**.**

$$cv_{ij} = \varepsilon_{ij}\Big(\alpha_j z_i y_i - m_{ij} l_{ij} b_j\Big). \tag{3.7}$$

Remark 1. Markdowns increase with $(1-\varepsilon_{ij})$, higher job substitutability (β), and lower unemployment benefits b_j. The vacancy condition is unaffected by β.

Remark 2 (Wage homogeneity in competitive search). In canonical competitive search models with constant marginal vacancy posting costs and no preference shocks, the wage rate is homogeneous across firms: the free entry of market makers (or the zero-profit condition) pins down a single wage. Kaas and Kircher (2015) introduce increasing marginal vacancy posting costs and heterogeneous firms with decreasing returns to labor; in that environment, wages depend on productivity and job-filling rates vary with firm characteristics.

3.4 Imperfect Competition with Finite Firms

This subsection introduces strategic interaction among a finite number of employers. The objective is to show how the competitive-search benchmark is modified once an individual firm internalizes its effect on the distribution of workers across rival firms. The key assumptions are the finite-firm equilibrium defined above, logit application choices, and constant-returns matching at the firm-input level. The subsection contributes the paper's main sufficient-statistics formulas: a markdown equation that depends on firm-specific market power and a vacancy equation that is unchanged relative to competitive search.

The finite-firm case differs from the competitive-search benchmark because a firm's wage posting now affects not only its own applicant pool but also the expected value of search at rival

firms. That interaction is summarized by the accessibility-adjusted market share

$$s_{ij} = \frac{\frac{l_{ij}}{1+\beta(w_{ij}-b_j)m_{ij}\varepsilon_{ij}}}{\sum_k \frac{l_{kj}}{1+\beta(w_{kj}-b_j)m_{kj}\varepsilon_{kj}}}, \qquad A_{ij} = 1 - s_{ij}. \tag{3.8}$$

The object s_{ij} is larger when firm i attracts many applicants and when those applicants have relatively few close substitutes among competing firms. The complement A_{ij} measures the residual elasticity of labor supply to firm i: when A_{ij} is small, firm i faces little competitive pressure from nearby alternatives and can sustain a larger markdown.

Proposition 4 (Markdown under imperfect competition)**.**

$$\frac{w_{ij}}{MPL_{ij}} = \frac{(1-\varepsilon_{ij})\beta w_{ij} m_{ij} A_{ij}}{1+\beta(w_{ij}-b_j)m_{ij}\varepsilon_{ij} + (1-\varepsilon_{ij})\beta w_{ij} m_{ij} A_{ij}}. \tag{3.9}$$

Relative to the competitive-search expression, the only additional sufficient statistic is A_{ij}. This is useful economically because it isolates the channel through which local concentration matters. A firm with a larger effective market share has a lower A_{ij}, and therefore pays a lower wage relative to marginal product. By contrast, when the market is highly contestable and A_{ij} is close to one, the markdown converges to the competitive-search benchmark.

Proposition 5 (Vacancy invariance)**.** *The vacancy FOC remains*

$$cv_{ij} = \varepsilon_{ij}\Big(\alpha_j z_i y_i - m_{ij} l_{ij} b_j\Big). \tag{3.10}$$

The second proposition shows that oligopsonistic distortions operate through wages rather than through the vacancy margin. Intuitively, a wage increase changes the attractiveness of the firm relative to all alternatives and therefore has a strategic effect on worker reallocation. An additional vacancy, by contrast, changes hiring only through the within-firm congestion term embedded in the matching function. Under constant returns, that congestion effect is fully summarized by ε_{ij}, so the firm's strategic effect on rivals cancels out of the vacancy first-order condition.

Remark 3. Markdowns rise with greater market share s_{ij} (lower A_{ij}) and smaller ε_{ij}. The implication is that the most productive or best-connected firms can simultaneously attract more workers, post more vacancies, and yet pay lower wages relative to marginal product.

3.5 Heterogeneous Productivity

This subsection isolates the role of productivity dispersion within a local market. The objective is to show, in the simplest analytically transparent case, how heterogeneity in z_i maps into heterogeneity

in wages, applicant flows, and tightness under finite-firm competition. The key assumptions are a single labor input, zero outside option, perfect observability of posted payoffs, and a power matching function. The subsection contributes a sharp comparative-static example that connects the general markdown formula to an empirically relevant prediction: more productive firms can be larger and tighter while paying lower labor shares.

Consider $N = 1$, $b = 0$, $\beta = \infty$, and $m_i = e\theta_i^{\varepsilon_i}$. The limit $\beta = \infty$ implies that workers direct applications toward the firms offering the highest expected payoff, so the example strips away idiosyncratic choice noise and highlights the equilibrium consequences of productivity heterogeneity alone.

Proposition 6. *Under competitive search, w_i, θ_i, and MPL_i are homogeneous. Under oligopsony, higher z_i implies higher θ_i, m_i, l_i, but lower w_i. The wage markdown is*

$$\mu_i = \frac{w_i}{MPL_i} = \frac{1}{1 + \frac{\varepsilon_i}{(1-s_i)(1-\varepsilon_i)}}. \tag{3.11}$$

The proposition makes clear why productivity dispersion has different implications under competitive search and under oligopsony. In the competitive benchmark, more productive firms bid more aggressively for workers until wages and tightness are equalized across employers. With finitely many firms, that equalization fails because high-productivity firms also attract a disproportionate share of applicants. Their larger market share reduces the elasticity of labor supply they face, allowing them to expand employment and tightness while paying a lower wage relative to marginal product. This mechanism is central for the Beveridge-curve implication developed next.

3.6 Implications for the Beveridge Curve

This subsection links the firm-level distortion to an aggregate labor-market implication. The objective is to show that dispersion in firm-level tightness lowers aggregate matching efficiency for a given total vacancy stock. The key assumptions are the heterogeneous-firm environment above, a common concave matching elasticity across firms, and a fixed aggregate number of vacancies. The subsection contributes an observable implication that distinguishes the model from a representative-firm search environment: concentration can shift the Beveridge curve even without any change in aggregate matching technology.

Let total vacancies be $V = \sum_k v_k$. With $b = 0$, $N = 1$, and $m_i = e\theta_i^{\varepsilon}$,

$$\text{Employment} = e \sum_k v_k^{\varepsilon} l_k^{1-\varepsilon} \leq e\Big(\sum_k v_k\Big)^{\varepsilon} = eV^{\varepsilon}, \tag{3.12}$$

with equality iff $\theta_i = \theta$ for all i.

The inequality follows from the concavity induced by $0 < \varepsilon < 1$. Holding total vacancies fixed, aggregate employment is maximized when vacancies and applicants are allocated so that all firms operate at the same tightness. Any dispersion in θ_i lowers total matches because some vacancies are posted where they are relatively unproductive while other firms face excessive congestion.

Proposition 7. *Under heterogeneity and oligopsony, θ_i disperse, shifting the Beveridge curve outward relative to the competitive benchmark.*

This proposition is the aggregate counterpart of the heterogeneous-productivity example. Under oligopsony, productive firms post more vacancies and attract more applicants, but wages do not fully arbitrage away those differences. The resulting dispersion in firm-level tightness lowers aggregate matching efficiency, so a location with the same total vacancies can exhibit higher unemployment. In this sense, concentration changes the Beveridge curve through reallocation across firms, not through an exogenous deterioration in matching efficiency.

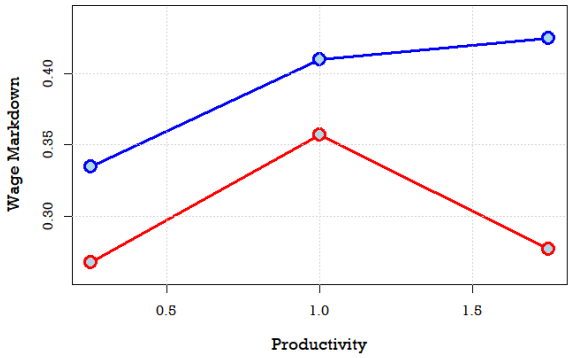
3.7 Illustrative Example

This subsection makes the preceding propositions visually concrete in the simplest nontrivial market; the full quantitative analysis in Section 4 calibrates these forces to data. Consider a single segmented location with three firms ($M = 3$), one labor input, and one competitive capital input (supply normalized to one). Matching is urn-ball, $m(v, l) = v[1 - \exp(-\xi l/v)]/\xi$, with the parameterization in the table below.

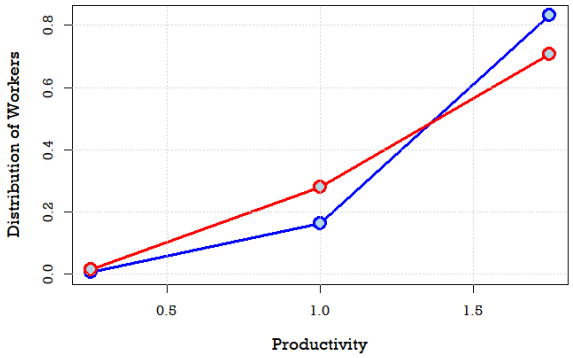
Parameter	Value	Interpretation
M	3	Number of firms
ξ	0.3	Search efficiency
α_k	0.25	Capital share
α_l	0.5	Labor share
β	10	Imperfect substitution among firms
b	0	Unemployment benefit
c	0.01	Vacancy posting cost
$\boldsymbol{z}$	(0.25, 1, 1.75)	Firm productivity

Under the baseline parameterization, oligopsony produces larger wage markdowns and a more concentrated worker allocation than competitive search: the high-productivity firm attracts a disproportionate share of applicants and pays a lower wage relative to marginal product (Figure 1). Markdowns shrink as β increases (workers substitute more readily across firms) and are amplified as β decreases; the quantitative section disciplines this parameter using the application elasticity.

Under oligopsony, employment and market tightness concentrate in high-productivity firms. Because wages do not fully arbitrage away productivity differences, firm-level θ_i disperse, lowering



(a) Wage markdowns across firms



(b) Worker distribution across firms

Figure 1: Oligopsonistic vs. competitive search (baseline). *Notes:* Red = oligopsony; Blue = competitive search.

aggregate matching efficiency for a given vacancy stock. Figure 2 confirms the Beveridge-curve implication of Section 3.6: the oligopsony curve lies outward relative to the competitive benchmark.

4 Quantitative Analysis

4.1 Calibration Strategy

We map the segmented-location model into a quantitative economy that disciplines concentration, firm size, applications, vacancies, and labor-market flows while keeping distinct the *long-run* reallocation of workers across locations from the *short-run* movements along a Beveridge curve. Production is

$$y_{i\ell} = z_{i\ell} k_{i\ell}^{\alpha_K} \left(m_{i\ell} l_{i\ell}\right)^{\alpha_L}, \qquad \alpha_L = \tfrac{2}{3}\alpha,\ \alpha_K = \tfrac{1}{3}\alpha,\ \alpha = 0.94,$$

Following Berger et al. (2022), we set the total production elasticity to $\alpha = 0.94$ and allocate two-thirds to labor and one-third to capital. Capital is hired in a competitive market at exogenous rental rate $R = 1$, and firm-level capital demand therefore satisfies

$$k_{i\ell} = \left(\frac{\alpha_K z_{i\ell} (m_{i\ell} l_{i\ell})^{\alpha_L}}{R} \right)^{\frac{1}{1-\alpha_K}}.$$

For matching we use a Cobb–Douglas technology with elasticity 0.5, a standard value in the search literature. Across locations, long-run labor supply is disciplined by the Berger–Herkenhoff–Mongey

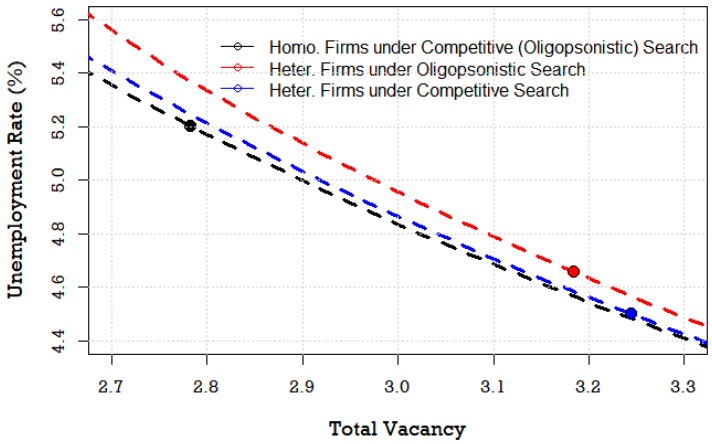


Figure 2: Beveridge curves: heterogeneous firms, oligopsony vs. competitive search.

estimate of the across-market elasticity, $\theta = 0.42$. Specifically, let

$$\mathcal{U}_\ell \equiv \frac{1}{\beta} \log \left(\sum_{i=1}^{M_\ell} e^{\beta u_{i\ell}} \right)$$

denote the ex ante promised value from searching in location ℓ, where $u_{i\ell}$ is the equilibrium promised utility from applying to firm i in location ℓ. We normalize the aggregate index to $\overline{\mathcal{U}} = 1$, so location labor supply satisfies

$$L_\ell = \bar{L} \mathcal{U}_\ell^\theta,$$

where $\bar{L}$ is a scale parameter chosen so that the aggregate economy matches the target mean firm size and unemployment rate. Thus, location labor supply is an isoelastic function of the local promised value from search rather than of the realized wage. The common outside option is set in each steady state to satisfy $b/\overline{MPL} = 0.5$, where $\overline{MPL}$ is the employment-weighted average marginal product of labor across locations. In the Beveridge counterfactuals below, location labor supply is held fixed at its steady-state value.

Our quantitative unit is a segmented location-industry labor market. We simulate a continuum of such locations. In each location ℓ, the number of firms M_ℓ is finite, workers search only within location, and firms draw productivity from a common distribution $F(z; \bar{Z}, \sigma_z)$. The baseline calibration uses *all industries*. This choice prioritizes the empirical distribution of local employer counts in the full economy; the tradable-industry specification is retained as a robustness check rather than as the organizing baseline.

We discipline the cross-location distribution of firm counts using Berger et al. (2022). Specif-

ically, M_ℓ follows a mixture of a mass point at $M_\ell = 1$ and a Pareto tail on $[2, \infty)$. Under the all-industries baseline, the mass at monopoly locations is 0.09, the Pareto tail parameter is 0.71, and the scale parameter is 38.36. These parameters reproduce the mean, dispersion, and skewness of employer counts across locations in their data. This object is externally disciplined rather than estimated inside our model; it is the source of cross-location heterogeneity in the baseline economy.

In the baseline calibration, the productivity distribution is common across locations. That is, locations differ only in the number of firms they contain, while firms within each location draw from the same log-normal distribution. We choose the dispersion parameter σ_z to match the payroll-weighted wage-bill Herfindahl index and the level parameter $\bar{Z}$ to match mean firm size *across all locations*. Under the all-industries baseline, the corresponding targets are an HHI of 0.11 and a mean firm size of 22.8, following Table 3 of Berger et al. (2022). The crucial distinction relative to the earlier normalization is that we no longer force every location to have the same labor force. Instead, the long-run labor supply block reallocates workers toward high-wage locations, while $\bar{Z}$ scales the overall level of firm size in equilibrium.

We discipline the logit parameter β using the elasticity of applications with respect to wages. This is the natural moment in our environment because workers choose where to apply, not only whether to participate. In the model, application shares respond to wages directly through the logit choice probabilities and indirectly through equilibrium tightness. We compute the model-implied application elasticity at the calibrated equilibrium and target the evidence in Marinescu and Skandalis (2020), where a 10 percent wage increase raises applications by roughly 7.7 percent within a job title.

We treat matching technology and unemployment as joint discipline for labor-market flows. With heterogeneous firms, the aggregate matching function is endogenous and need not coincide with the firm-level technology. For that reason, we do not estimate a macro matching function directly from aggregate data. Instead, we set the matching elasticity to 0.5 and choose the matching-efficiency parameter to match a *steady-state unemployment rate* of roughly 6 percent. This is the appropriate target in the current static version of the model: there is no dynamic transition that would separately identify a quarterly job-finding rate. In the model, unemployment is the mass of workers who fail to match in a period; in the data, unemployment is usually restricted to non-employed workers who are actively searching. We therefore interpret the target as a survey-based stock concept rather than as a mechanical count of all non-employed workers.

We calibrate the outside option and vacancy costs jointly. The flow value of unemployment b is calibrated relative to the *average marginal* product of labor across locations rather than to average product. This distinction matters because production exhibits decreasing returns to scale. In the implementation we impose $b/\overline{MPL} = 0.5$, which is in the empirical range discussed by Chodorow-Reich and Karabarbounis (2016). The vacancy cost c is then chosen to discipline aggregate tightness,

defined as the economy-wide vacancy–applicant ratio

$$\theta^{\text{agg}} = \frac{\sum_{i,\ell} v_{i\ell}}{\sum_{i,\ell} l_{i\ell}}.$$

Under the constrained matching system, this target is informative and can be matched reasonably closely jointly with unemployment. Together, these parameters determine how much of the surplus from a local match is dissipated through search frictions and how much is shared with workers.

4.2 Baseline Fit

The all-industries baseline yields the following parameter values:

$$\sigma_z = 0.207, \qquad \bar{Z} = 84.4, \qquad \beta = 0.0091, \qquad c = 0.237, \qquad a_m = 0.115,$$

where a_m denotes the matching-efficiency shifter in the Cobb–Douglas matching function. The same run implies $b = 96.1$, so that $b/\overline{MPL} = 0.49$. Table 2 reports the baseline parameter values compactly. Under this calibration the model implies a payroll-weighted HHI of 0.114, mean firm size of 22.8, steady-state unemployment of 5.9 percent, an application elasticity of 0.756, and aggregate tightness of 69.0 against a target of 70.

The aggregate tightness target remains the vacancy–applicant ratio θ^{agg} defined above, which is the appropriate aggregation of the firm-level tightness object in the model. In addition, the constrained matching block imposes that realized matches cannot exceed either applicants or vacancies at the firm-location level. These two changes make the aggregate tightness target economically coherent in the static environment and materially improve fit relative to the earlier vacancy–unemployment ratio.

The Beveridge curve is not used as a primary calibration target. Instead, it serves as an out-of-sample validation exercise. After matching concentration, firm size, application elasticity, and steady-state unemployment, we ask whether the calibrated model reproduces the cross-sectional relationship between vacancies and unemployment across more and less concentrated locations. We then use the same calibrated economy for counterfactuals that change local market structure, such as increasing M_ℓ toward the competitive-search benchmark or reducing M_ℓ to mimic consolidation.

4.3 Aggregate Beveridge Curve and Counterfactuals

Within a location, higher-productivity firms attract larger applicant shares and therefore charge larger markdowns. Across locations, smaller-M_ℓ markets display more concentrated employment and a thicker upper tail of firm-specific markdowns. Relative to the competitive-search benchmark,

Table 1: Calibration strategy for the segmented-location economy

Object	**Role**	**Baseline target / value**	**Source / comment**
Externally disciplined objects			
$\alpha = \alpha_L + \alpha_K$	Production elasticity	$\alpha = 0.94$; $\alpha_L = \frac{2}{3}\alpha$, $\alpha_K = \frac{1}{3}\alpha$	Berger et al. (2022), all-industries calibration
M_ℓ distribution	Cross-location heterogeneity	Mass at 1: 0.09; Pareto tail: 0.71; scale: 38.36	Berger et al. (2022), Tables F1–F2, all industries
Matching elasticity	External discipline	Standard search elasticity	Taken from the matching literature because macro aggregation is endogenous
b/MPL range	External discipline	0.47–0.96	Chodorow-Reich and Karabarbounis (2016); mapped to MPL, not APL
Internally calibrated parameters			
σ_z	Productivity dispersion	Payroll-weighted HHI: 0.11	Berger et al. (2022), Table 3, all industries
$\bar{Z}$	Productivity level	Mean firm size: 22.8	Berger et al. (2022), Table 3, all industries
β	Application responsiveness	Application elasticity: ≈ 0.77	Marinescu and Skandalis (2020); wage-applicant elasticity within job title
Matching efficiency	Labor-market flows	Steady-state unemployment rate: ≈ 0.06	Static version targets the unemployment stock rather than quarterly flows
c	Vacancy posting	Aggregate tightness: 70	Vacancy-applicant ratio, $\sum v_{i\ell} / \sum l_{i\ell}$
Validation moments and exercises			
Beveridge curve slope and shift	Quantitative validation	Cross-location vacancy–unemployment relationship	Compared across more and less concentrated locations
Tradables specification	Robustness	Alternative baseline sample	Used to assess sensitivity to product-market concerns

finite-firm local markets lower the average wage, increase the aggregate markdown, and reallocate employment toward high-share firms inside each location. Because the vacancy condition is

Table 2: Baseline parameter values in the implemented calibration

Parameter	Value	Status	Interpretation
α_L	0.627	Imposed	Labor elasticity in production
α_K	0.313	Imposed	Capital elasticity in production
R	1.00	Imposed	Exogenous rental rate of capital
Matching elasticity	0.50	Imposed	Cobb–Douglas matching elasticity
θ	0.42	External	Across-location labor-supply elasticity
σ_z	0.207	Internally calibrated	Dispersion of firm productivity
$\bar{Z}$	84.4	Internally calibrated	Common productivity level
β	0.0091	Internally calibrated	Application responsiveness
a_m	0.115	Internally calibrated	Matching-efficiency shifter
c	0.237	Internally calibrated	Vacancy posting cost
b	96.1	Implied	Outside option, satisfying $b/\overline{MPL} = 0.49$

Table 3: Model fit: calibrated moments and data targets

Moment	Model	Data / target	Source / comment
Payroll-weighted HHI	0.114	0.110	Berger et al. (2022), Table 3, all industries
Mean firm size	22.82	22.8	Berger et al. (2022), Table 3, all industries
Application elasticity	0.756	0.77	Marinescu and Skandalis (2020)
Steady-state unemployment rate	0.059	0.06	Static calibration target
Aggregate tightness	68.96	70.0	Vacancy-applicant ratio, $\sum v_{i\ell} / \sum l_{i\ell}$

invariant, these wage effects are the central margin through which local labor market power operates in the model.

We trace the aggregate Beveridge curve by fixing per-location labor supply $\{L_\ell\}$ and b at their baseline steady-state values and varying the common productivity level $\bar{Z}$ across a grid. Each $\bar{Z}$ point solves the short-run equilibrium within every location type and aggregates the resulting vacancies and unemployment. Because labor supply is held constant, movements along the curve reflect only how aggregate productivity shifts the short-run vacancy–unemployment trade-off for a given market structure. Counterfactuals change either productivity dispersion (σ_z) or the cross-location distribution of firm counts while maintaining the same per-location labor supply and b, isolating the short-run shift from any migration response.

Figure 3 plots the baseline short-run curve (calibrated $\sigma_z = 0.21$) together with two counterfactuals. All three curves hold per-location labor supply and b fixed; only the source of variation—productivity dispersion or market structure—changes. Vacancies are normalized by the total number of workers.

The first counterfactual doubles productivity dispersion ($\sigma_z = 0.41$). Higher σ_z shifts the curve outward at a given unemployment rate: greater dispersion in firm-level productivity increases dispersion in tightness and lowers aggregate matching efficiency for a given vacancy stock.

The second counterfactual increases concentration through both margins of the firm-count distribution: $P(M_\ell = 1)$ is doubled from 0.09 to 0.18, and the Pareto scale governing $M_\ell \mid M_\ell > 1$ is compressed so that $E[M_\ell]$ falls by half (from 8.4 to 4.2). The compressed distribution is projected onto the same set of location types, so per-location labor supply and b remain exactly the same; only the weights that each location type carries in the aggregate change. The payroll-weighted HHI rises from 0.114 to 0.423. This concentration counterfactual shifts the Beveridge curve outward only modestly: at the baseline unemployment rate, vacancies per worker edge down from 69.0 to 68.7. The quantitative effect is small relative to the productivity-dispersion channel, suggesting that changes in local firm counts shift the curve primarily through their effect on the equilibrium allocation of workers across locations—a margin shut down by construction in the fixed-labor-supply exercise.

4.4 Beveridge Curves by Location Size

Figure 4 sorts locations into four groups by the number of firms—monopoly ($M_\ell = 1$), small oligopoly ($M_\ell \in \{2, 3, 4\}$), medium oligopoly ($M_\ell \in \{5, \ldots, 9\}$), and competitive ($M_\ell \geq 10$)—and traces a separate Beveridge curve for each group. All four curves hold per-location labor supply and b at the baseline long-run steady state; the only source of variation is $\bar{Z}$.

The monopoly curve does not appear in the figure because a single-firm location drives the wage to b and posts enough vacancies to match nearly all available workers: the within-group unemployment rate is below 0.1 percent for every $\bar{Z}$ on the grid. Among the remaining groups, the Beveridge curve shifts outward and steepens as concentration rises. As the number of firms increases, the curve flattens and shifts inward: competition reduces markdowns, compresses the tightness distribution, and improves local matching efficiency. The ordering is monotone, confirming that local employer concentration is a quantitatively meaningful shifter of the vacancy–unemployment relationship.

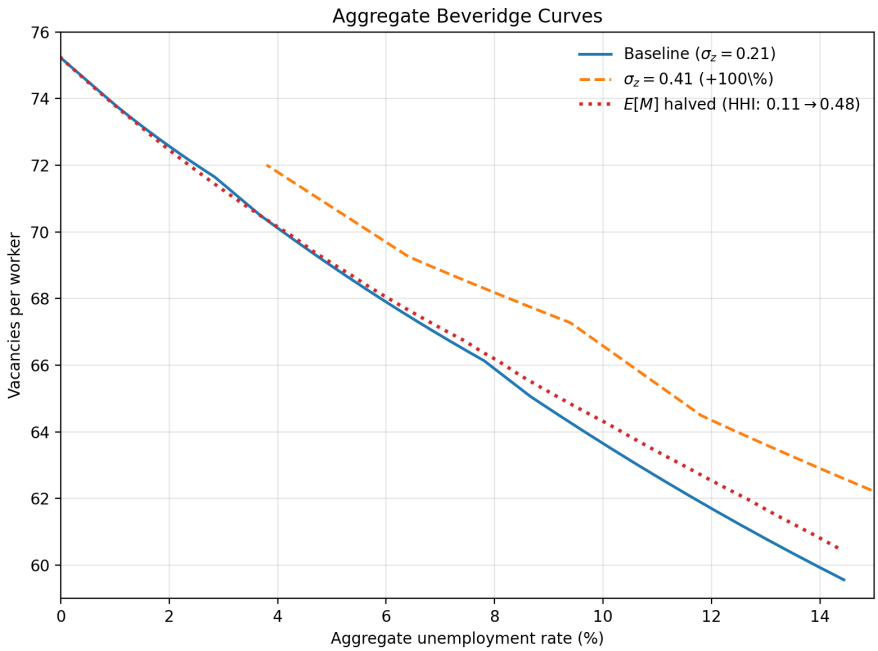


Figure 3: Aggregate Beveridge curves under fixed per-location labor supply. The horizontal axis reports the aggregate unemployment rate (restricted to below 15 percent); the vertical axis reports vacancies per worker. All curves hold per-location labor supply and b at the baseline long-run steady state; the curve is traced by varying $\bar{Z}$. The solid curve is the baseline ($\sigma_z = 0.21$). The dashed curve doubles σ_z to 0.41. The dotted curve halves $E[M_\ell]$ by doubling $P(M_\ell = 1)$ from 0.09 to 0.18 and compressing the Pareto tail of $M_\ell \mid M_\ell > 1$, reweighting onto the same location types (HHI rises from 0.114 to 0.423).

4.5 Cyclical Productivity Dispersion

In the baseline, σ_z is held constant across productivity levels. We now consider an alternative economy in which σ_z varies linearly with $\bar{Z}$: as $\bar{Z}$ ranges from 5 percent below to 5 percent above the level associated with 5.5 percent unemployment, σ_z decreases from twice the calibrated value to half the calibrated value. This specification captures the empirical regularity that firm-level productivity dispersion rises in recessions.

Figure 5 compares the two curves. The constant-σ_z curve is the same baseline as in Figure 3. The cyclical-σ_z curve rotates: in recessions, higher dispersion pushes the curve outward because greater within-location tightness heterogeneity worsens matching efficiency; in expansions, lower dispersion pulls the curve inward. The net effect is a steeper Beveridge curve that amplifies the business cycle relative to the constant-dispersion benchmark.

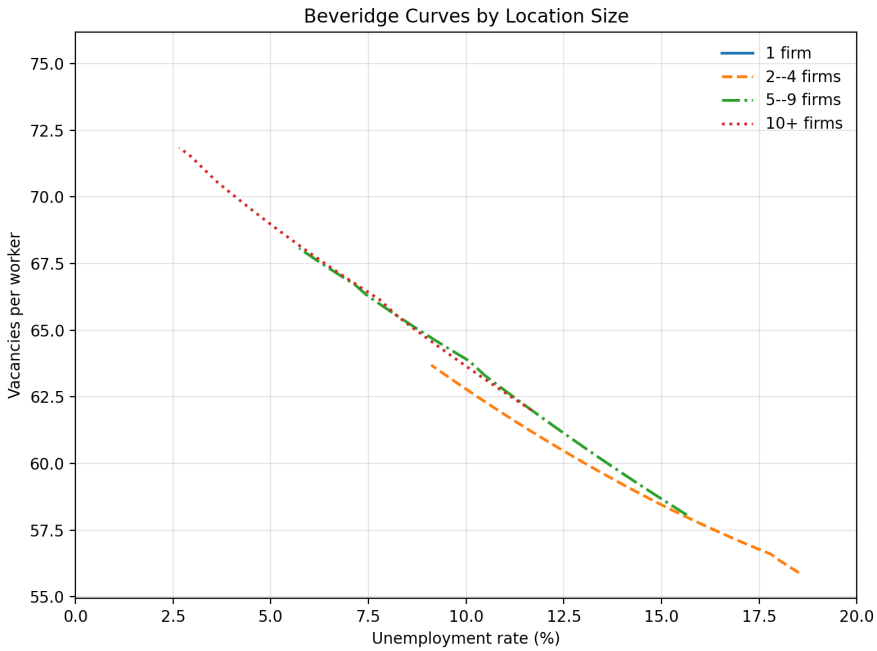


Figure 4: Beveridge curves by location size group. Each curve traces vacancies per worker against the group-specific unemployment rate as $\bar{Z}$ varies, holding per-location labor supply and b fixed. Dashed: $M_\ell \in \{2, 3, 4\}$. Dash-dot: $M_\ell \in \{5, \ldots, 9\}$. Dotted: $M_\ell \geq 10$. The monopoly curve ($M_\ell = 1$) is omitted: the monopsonist's extreme markdown drives matching rates to one, so within-group unemployment is near zero for all $\bar{Z}$.

4.6 Implications for the Wage Phillips Curve

A structural Phillips curve through the Beveridge-curve channel. The same comparative-statics exercise that traces the Beveridge curve also delivers a structural wage Phillips curve. As the common productivity level $\bar{Z}$ rises, firms post more vacancies, unemployment falls, and the employment-weighted average wage rises, generating a downward-sloping wage–unemployment relationship that is the real analogue of the relationship documented by Phillips (1958).

Because the average wage can be decomposed as $\bar{w} = \bar{\mu} \cdot \overline{MPL}$, where $\bar{\mu} \equiv w/MPL$ is the employment-weighted average wage–marginal-product ratio and $\overline{MPL}$ is the employment-weighted average marginal product, the wage response to a productivity shift has two components:

$$\frac{d\bar{w}}{d\bar{Z}} = \underbrace{\bar{\mu}\,\frac{d\overline{MPL}}{d\bar{Z}}}_{\text{marginal-product channel}} + \underbrace{\overline{MPL}\,\frac{d\bar{\mu}}{d\bar{Z}}}_{\text{markdown channel}}. \tag{4.1}$$

Under competitive search, $\bar{\mu}$ is close to one and relatively stable, so productivity shocks pass through to wages almost one-for-one and the Phillips curve is steep. Under oligopsony, $\bar{\mu}$ is substantially below one and *rises* when $\bar{Z}$ falls (markdowns shrink when the economy weakens), so the markdown

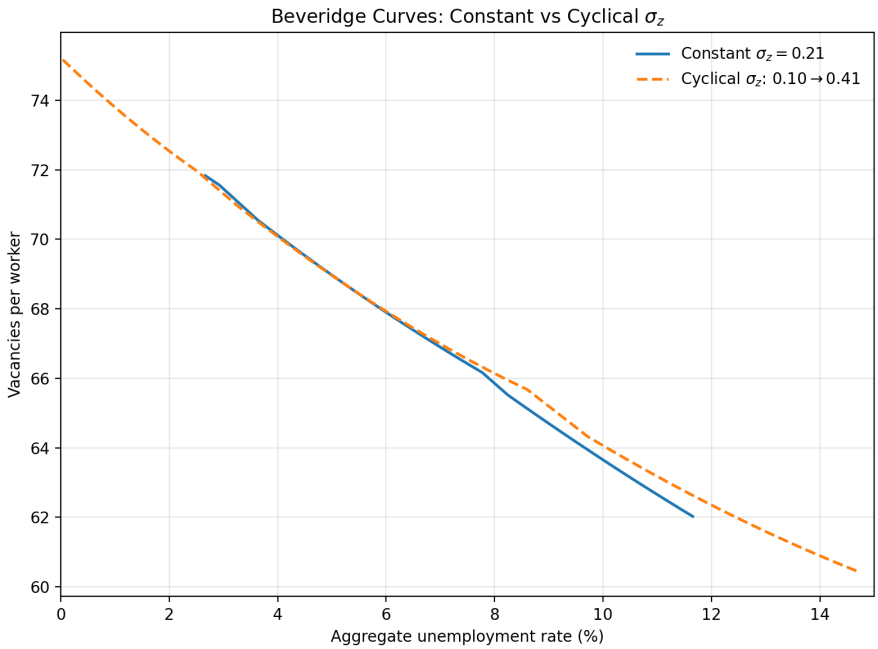


Figure 5: Aggregate Beveridge curves under constant versus cyclical σ_z. Both curves hold per-location labor supply and b fixed at the baseline long-run steady state. Solid: constant $\sigma_z = 0.21$. Dashed: σ_z varies linearly from $2\times$ baseline (in recessions) to $0.5\times$ baseline (in expansions) as $\bar{Z}$ moves $\pm 5\%$ around the level associated with 5.5% unemployment.

channel partially offsets the marginal-product channel. In the calibrated economy, this offset is nearly complete: the average wage barely moves despite large changes in unemployment.

Concentration flattens the wage Phillips curve. Figure 6 panel (a) plots the wage Phillips curve under three market structures, all holding per-location labor supply and b at the baseline steady state. Under competitive search, the average wage declines by roughly 22 percent of its steady-state level per unit increase in the unemployment rate—a steep Phillips curve in which wages track marginal product closely. The baseline oligopsony curve is essentially flat: wages remain within 0.5 percent of the outside option b across the entire range of unemployment rates considered. The $E[M]$-halved counterfactual is also very flat relative to competitive search, with a slope attenuated by more than an order of magnitude.

Panel (b) confirms the mechanism. The markdown ratio w/MPL is nearly constant and close to one under competitive search. Under oligopsony, w/MPL varies from 0.45 in recessions to 0.35 in expansions: the markdown margin absorbs the productivity shock that would otherwise translate into wage movements. Higher concentration amplifies this absorption.

The flatness of the oligopsony Phillips curve is consistent with the cross-state evidence in Hazell et al. (2022), who estimate a structural slope $\kappa \approx 0.006$ for the regional Phillips curve using

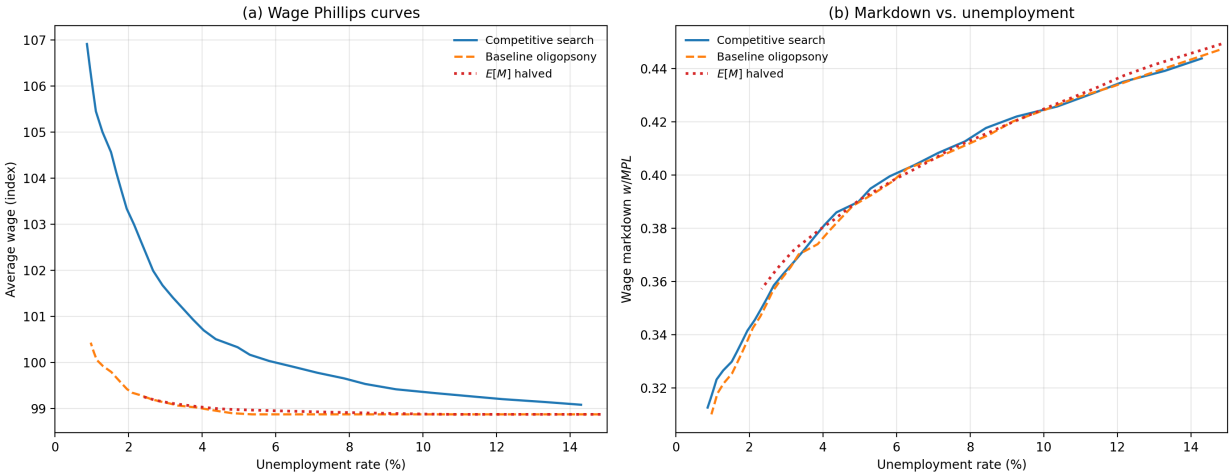


Figure 6: Wage Phillips curves and markdown decomposition. Panel (a): employment-weighted average wage (indexed) against the aggregate unemployment rate under competitive search (solid), baseline oligopsony (dashed), and $E[M]$ halved (dotted). Panel (b): employment-weighted markdown ratio w/MPL against unemployment under the same three market structures. All curves hold per-location labor supply and b at the baseline long-run steady state; variation is generated by changing $\bar{Z}$.

state-level nontradeable inflation and unemployment data from 1978 to 2018. They interpret this small slope as requiring "a large amount of real rigidity" relative to a textbook New Keynesian model. In our framework, the markdown mechanism provides exactly this real rigidity: even with fully flexible wages and no nominal frictions, oligopsonistic wage-setting makes the equilibrium wage nearly unresponsive to demand conditions. The markdown channel is therefore a structural microfoundation for the large real rigidity that Hazell et al. and earlier structural estimates (e.g., Nakamura and Steinsson, 2014) need to rationalize the small empirical slope of the Phillips curve.

Heterogeneity and the aggregate Phillips curve. Disaggregating by location size reveals that the flatness of the aggregate wage Phillips curve is driven by the most concentrated markets. Figure 7 plots group-specific wage Phillips curves for locations with 2–4 firms, 5–9 firms, and 10 or more firms. Locations with fewer employers display flatter curves: the markdown margin is wider and absorbs a larger share of the productivity shock. In competitive locations ($M_\ell \geq 10$), wages respond meaningfully to unemployment movements. The aggregate Phillips curve is a weighted average of these heterogeneous local curves, so changes in the composition of market structures—for example, a secular increase in employer concentration—can flatten the aggregate curve even without any change in nominal rigidities or inflation expectations. This cross-location pattern maps naturally to the cross-state estimates of Hazell et al. (2022), who identify the slope of the Phillips curve from regional variation in unemployment after differencing out aggregate inflation expectations via time

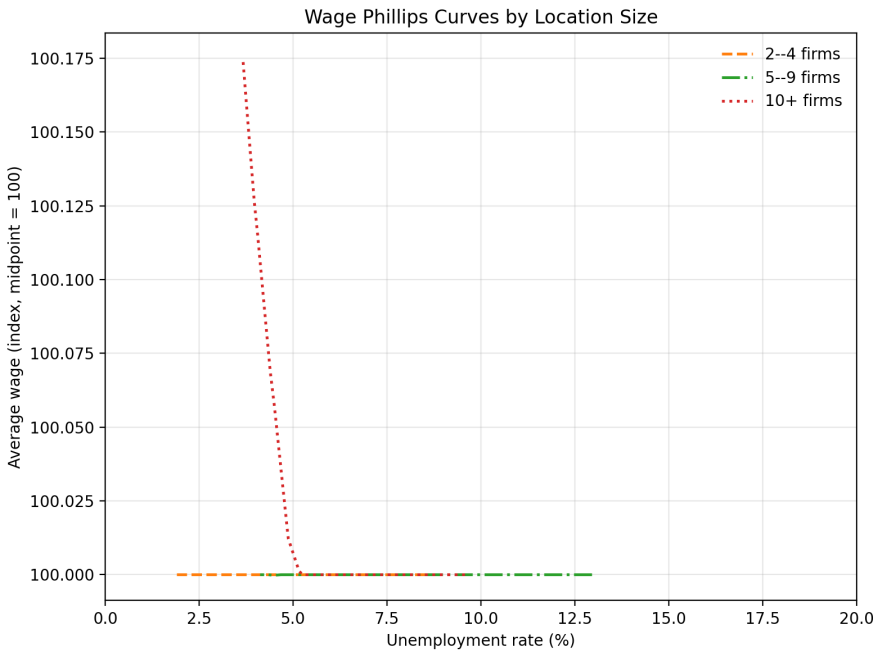


Figure 7: Wage Phillips curves by location size. Each curve traces the group-specific employment-weighted average wage (indexed) against the group-specific unemployment rate as $\bar{Z}$ varies, holding per-location labor supply and b fixed. Dashed: $M_\ell \in \{2, 3, 4\}$. Dash-dot: $M_\ell \in \{5, \ldots, 9\}$. Dotted: $M_\ell \geq 10$.

fixed effects. Our model predicts that their small estimated slope ($\kappa \approx 0.006$) reflects an average over heterogeneous local slopes that are systematically flatter in more concentrated markets—a prediction that can be tested by interacting their state-level unemployment measure with state-level employer concentration.

Cyclical markdowns as a structural interpretation of the time-varying Phillips curve. When productivity dispersion σ_z varies cyclically—rising in recessions and falling in expansions, as in the Beveridge-curve exercise above—the markdown distribution also becomes state-dependent. Higher dispersion increases the variance of firm-level tightness, which raises the average markdown and widens the gap between $\overline{MPL}$ and $\bar{w}$. Figure 8 compares the constant-σ_z and cyclical-σ_z wage Phillips curves. The cyclical-dispersion curve rotates: wages respond even less in recessions (when markdowns are large and dispersed) and somewhat more in expansions (when markdowns are compressed). This pattern provides a structural channel through which the slope of the wage Phillips curve can vary over the business cycle without invoking changes in expectations anchoring or monetary policy credibility.

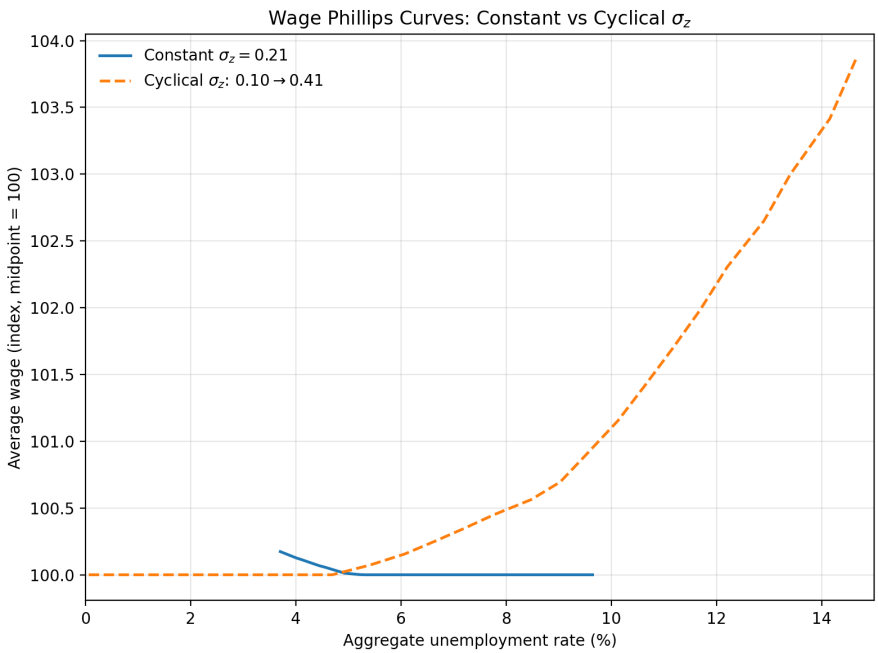


Figure 8: Wage Phillips curves under constant versus cyclical σ_z. Both curves hold per-location labor supply and b fixed. Solid: constant $\sigma_z = 0.21$. Dashed: σ_z varies linearly from $2\times$ baseline (in recessions) to $0.5\times$ baseline (in expansions) as $\bar{Z}$ moves $\pm 5\%$ around the level associated with 5.5% unemployment.

5 Discussion and Empirical Implications

5.1 Concentration and Wages

Firm level. At the firm level, Equation (3.9) implies that markdowns rise with a firm's accessibility-adjusted market share and fall with the elasticity of matching. Empirically, this links wage markdowns to application shares, vacancy yields, and measures of local concentration constructed from vacancy or payroll data.

The theory also predicts that markdowns are heterogeneous across firms within the same market: higher-productivity firms attract larger application shares, command larger market shares, and consequently extract larger markdowns. A growing body of evidence supports this prediction. Yeh et al. (2022) compute wage markdowns for U.S. manufacturing firms and document that markdowns are increasing in both firm size and productivity. Seegmiller (2021) infer markdowns from firm-level labor supply elasticities using matched employer–employee LEHD data, instrumenting for labor demand with stock returns; they find that high-labor-productivity firms face significantly lower elasticities, consistent with greater monopsony power. Complementary evidence comes from Mertens and Mottironi (2025), who show that larger firms exhibit lower product-market markups but more than offset this through substantially higher wage markdowns, suggesting that firm size

confers market power primarily on the buyer side of the labor market.

Local labor market level. At the location level, employer concentration is informative about wage-setting power because it changes the strategic importance of each firm's wage choice. By contrast, vacancy levels are not sufficient statistics for market power in the segmented-location economy. Conditional on fundamentals, the vacancy first-order condition is identical to the competitive benchmark.

The theory predicts that higher local labor market concentration leads to larger wage markdowns and lower wages. This prediction finds robust support across data sources and identification strategies. Azar, Marinescu, and Steinbaum (2022) measure concentration at the geographic–occupation level using CareerBuilder vacancy postings and find that higher concentration significantly depresses posted wages. Azar, Marinescu, Steinbaum, and Taska (2020) reach a similar conclusion using Burning Glass data at the commuting-zone–occupation level, suggesting the relationship is not an artifact of a particular platform or geographic definition. Benmelech, Bergman, and Kim (2022) move from posted to realized wages, documenting that local employer concentration lowers actual wages in plant-level U.S. Census data; they establish causality by exploiting variation generated by merger activity.

Number of employers and aggregate markdowns. Rather than relying on concentration indices, one can directly examine the number of employers. Difference-in-differences designs exploiting firm entry, exit, or market consolidation events—hospital mergers, plant closures in manufacturing towns—provide a natural testing ground for whether wages fall and markdowns rise following reductions in M.

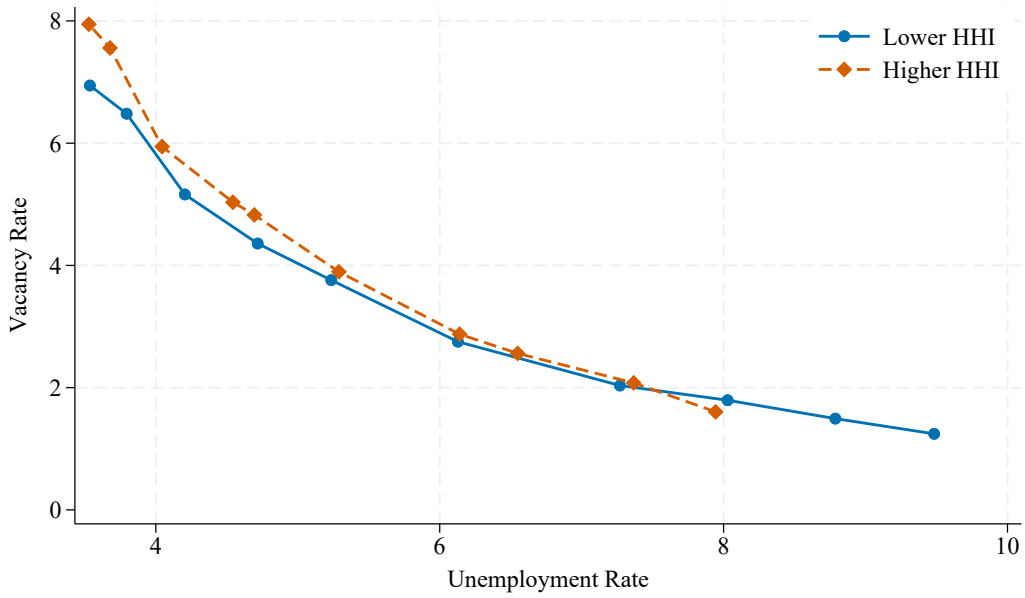
Prager and Schmitt (2021) offer the canonical application of this approach, exploiting well-defined hospital merger events; they document reduced wage growth among workers at merging hospitals while finding no change in wage trajectories following out-of-market mergers, lending credibility to a monopsony interpretation. Arnold (2019) scale this logic to the full U.S. economy using matched employer–employee LEHD data and find that mergers and acquisitions broadly increase local labor market concentration and depress worker earnings. Evidence from outside the U.S. reinforces these findings: Dodini et al. (2024) exploit mass layoffs and establishment closures in Norway as exogenous shocks to the local number of employers and show that separated workers experience substantially worse subsequent labor market outcomes when their local market is more concentrated, consistent with the model's prediction that markdowns are convex in concentration.

5.2 Beveridge curve shifts and market structure

For interpreting the Beveridge curve, the model implies that an outward shift need not be read solely as a deterioration in matching efficiency. Changes in local market structure or in the dispersion of firm outcomes within a location can generate similar patterns. This is the main empirical reason to treat the Beveridge curve as a validation exercise after the core calibration has already been fixed.

The baseline quantitative exercise has important limitations. It abstracts from cross-location worker mobility and from location-specific productivity shifters. Those omissions are useful for identification because they isolate the role of the firm-count distribution, but they also limit external validity. Any empirical implementation must therefore be explicit about the unit of observation, the unemployment concept, and the possibility of spillovers across nearby labor markets.

Figure 9: Beveridge Curves by Employer Concentration



Notes: This figure plots the Beveridge curve separately for states with above-median (Higher HHI) and below-median (Lower HHI) employer concentration. Vacancy and unemployment rates are from the Bureau of Labor Statistics' Job Openings and Labor Turnover Survey (JOLTS) at the state level. Employer concentration is aggregated from the Herfindahl–Hirschman Index (HHI) constructed by Choi and Marinescu (2024).

Figure 9 plots Beveridge curves separately for states with above- and below-median employer concentration. We aggregate the Herfindahl–Hirschman Index (HHI) constructed from Choi and Marinescu (2024) at the six-digit SOC–commuting zone level to the state level, and merge with JOLTS state-level vacancy and unemployment data. The Beveridge curve for more concentrated markets lies above that for less concentrated markets, indicating that concentrated markets sustain a higher vacancy rate at every level of unemployment. This pattern is consistent with the model's prediction that monopsonistic firms internalize their wage-setting power and post more vacancies at

lower wages, generating additional unfilled positions in equilibrium. The figure provides suggestive evidence that employer concentration shifts the Beveridge curve outward.

5.3 The Flattening Phillips Curve and Labor Market Power

A large macroeconomic literature documents that the slope of the Phillips curve has declined over the past several decades. Standard explanations emphasize better-anchored inflation expectations (Blanchard, 2016), improved monetary policy credibility, and globalization. McLeay and Tenreyro (2020) show that identification of the Phillips curve slope requires careful treatment of endogenous monetary policy, and that the reduced-form slope conflates structural parameters with policy responses.

Hazell et al. (2022) make the most direct empirical contribution to this debate. They construct new state-level price indices for nontradeable goods back to 1978 and exploit cross-state variation in unemployment to estimate the slope of the Phillips curve using a panel specification with state and time fixed effects. The time fixed effects absorb variation in long-run inflation expectations—the confound that drives the apparent steepness of the aggregate Phillips curve in the pre-1990 sample. Their preferred estimate of the structural New Keynesian slope is $\kappa \approx 0.006$, an order of magnitude smaller than the textbook New Keynesian calibration. Crucially, they find that the Phillips curve was *always* flat: when time fixed effects are included, the pre-1990 slope ($\kappa = 0.011$) is only twice as large as the post-1990 slope ($\kappa = 0.005$), not 50–100 times larger as the aggregate time series suggests. They conclude that the apparent flattening of the Phillips curve is largely an artifact of inflation expectations becoming more firmly anchored.

The markdown channel as a microfoundation for real rigidity. Hazell et al.'s small κ is consistent with New Keynesian models that incorporate substantial real rigidity. In their Table III comparison, structural estimates that embed heterogeneous labor markets and real rigidity (e.g., Nakamura and Steinsson, 2014) yield κ values close to the empirical estimate, while simpler models yield values an order of magnitude too large. Our framework provides a new structural microfoundation for this real rigidity that operates through the labor market rather than through the goods market. Under oligopsony, the markdown margin w/MPL is procyclical—it rises in expansions as firms capture more of the surplus—so that a given change in marginal product translates into a much smaller change in wages. This mechanism dampens the pass-through from marginal costs to wages that ultimately drives the New Keynesian Phillips curve, even when prices are fully flexible on the product side.

Cross-state variation and cross-location heterogeneity. The cross-state identification strategy of Hazell et al. has a natural analogue in our model's cross-location structure. In the model, different

locations have different numbers of employers, and Figure 7 shows that locations with fewer firms display flatter wage Phillips curves. This cross-location heterogeneity generates a testable prediction that can be taken directly to the Hazell et al. data: augmenting their specification with an interaction between state-level unemployment and state-level employer concentration should yield a coefficient indicating that the Phillips curve is flatter in more concentrated states. Formally, one could estimate

$$\pi_{it}^{N} = \alpha_i + \gamma_t - \psi\, u_{i,t-4} - \psi_H\, u_{i,t-4} \times \text{HHI}_i - \delta\, \hat{p}_{i,t-4}^{N} + \varepsilon_{it},$$

where HHI_i is a time-invariant measure of employer concentration in state i constructed from Choi and Marinescu (2024) or Azar, Marinescu, Steinbaum, and Taska (2020), and the model predicts $\psi_H > 0$ (the negative wage–unemployment relationship is weaker in concentrated markets). Beraja et al. (2019) estimate regional wage Phillips curves using scanner wage data and find cross-region heterogeneity in slopes; our model provides a structural explanation for that heterogeneity through the differential markdown margins.

Scope and limitations. Two caveats limit the scope of this implication. First, the model is real: it speaks to the *wage* Phillips curve but is silent on the price Phillips curve and on inflation expectations. The contribution is therefore to the structural relationship between real wages and unemployment, not to the reduced-form relationship between inflation and unemployment that dominates the policy debate. To the extent that wage costs pass through to prices, a flatter wage Phillips curve implies a flatter price Phillips curve, but the mapping depends on product-market structure that is outside the model. Second, the model is static, so it cannot address the dynamic interaction between expectations, policy, and wage-setting power that may be important for understanding the time-varying slope of the Phillips curve at business-cycle frequency. Nevertheless, the combination of the model's prediction that concentration flattens the wage Phillips curve with Hazell et al.'s cross-state identification strategy offers a concrete empirical test of whether labor market power is a quantitatively important source of real rigidity in the U.S. economy.

6 Conclusion

We study labor market power in a directed-search environment with a continuum of segmented locations, each containing a finite number of firms that post wages and vacancies. The central theoretical result is that wages, but not vacancy conditions, internalize local oligopsony power: markdowns depend on accessibility-adjusted market share and matching elasticity, while the vacancy first-order condition is identical to the competitive-search benchmark. Heterogeneity within a location then generates an outward shift of the Beveridge curve even without changes in matching

efficiency.

The quantitative framework is designed to discipline that mechanism rather than to add reduced-form facts mechanically. The all-industries baseline maps each model market to a segmented location, matches the observed distribution of employer counts, concentration, firm size, application elasticities, and labor-market flows, and then uses the Beveridge curve as an out-of-sample validation exercise. Within this environment, counterfactual changes in local firm counts provide a transparent measure of how much wages and unemployment respond to employer concentration. The same markdown mechanism that shifts the Beveridge curve also flattens the wage Phillips curve: the calibrated oligopsony economy exhibits a nearly flat wage–unemployment trade-off because productivity shocks are absorbed through the markdown margin rather than through wage adjustments. Empirical patterns are consistent with the model's predictions: firm-level wage markdowns rise with size and productivity, local employer concentration depresses wages across multiple identification strategies, and states with above-median concentration display outward-shifted Beveridge curves relative to less concentrated states.

More broadly, the framework is not specific to labor markets. The same logic applies to environments in which a finite number of intermediaries or platforms post prices and capacity while one side of the market directs search under congestion. In a deposit-market application, banks would post deposit rates and branch or processing capacity; in a platform application, firms would post prices and seller capacity. In each case, the algebra is unchanged: the price wedge depends on an accessibility-adjusted residual share, while the capacity condition depends only on the elasticity of matching. Extensions that introduce cross-location mobility, richer productivity heterogeneity, policy interventions, or applications to other directed-search markets remain important next steps, but the baseline already clarifies which moments are required to take the theory to the data.

References

Arnold, D. (2019). Mergers and acquisitions, local labor market concentration, and worker outcomes. *Local Labor Market Concentration, and Worker Outcomes (October 27, 2019)*.

Azar, J., I. Marinescu, and M. Steinbaum (2022). Labor market concentration. *Journal of Human Resources 57*(S), S167–S199.

Azar, J., I. Marinescu, M. Steinbaum, and B. Taska (2020). Concentration in us labor markets: Evidence from online vacancy data. *Labour Economics 66*, 101886.

Benmelech, E., N. K. Bergman, and H. Kim (2022). Strong employers and weak employees: How does employer concentration affect wages? *Journal of Human Resources 57*(S), S200–S250.

Beraja, M., E. Hurst, and J. Ospina (2019). The aggregate implications of regional business cycles. *Econometrica 87*(6), 1789–1833.

Berger, D., K. Herkenhoff, S. Mongey, and S. Posner (2022). Labor market power. *American Economic Review 112*(7), 2014–2054.

Blanchard, O. (2016). The Phillips Curve: Back to the '60s? *American Economic Review 106*(5), 31–34.

Burdett, K., S. Shi, and R. Wright (2001). Pricing and matching with frictions. *Journal of Political Economy 109*(5), 1060–1085.

Chodorow-Reich, G. and L. Karabarbounis (2016). The cyclicality of the opportunity cost of employment. *Journal of Political Economy 124*(6), 1563–1618.

Choi, H. and I. Marinescu (2024). Data for labor market concentration using Lightcast (formerly Burning Glass Technologies). *Data in Brief 55*, 110647.

Dodini, S., M. Lovenheim, K. Salvanes, and A. Willén (2024). Monopsony, job tasks and labour market concentration. *The Economic Journal 134*(661), 1914–1949.

Gottfries, A. and G. Jarosch (2025). Monopsony in the labor market. *Quarterly Journal of Economics 140*(1), 1–52.

Hazell, J., J. Herreno, E. Nakamura, and J. Steinsson (2022). The slope of the Phillips Curve: Evidence from U.S. states. *Quarterly Journal of Economics 137*(3), 1299–1344.

Hurst, E., P. Kehoe, E. Pastorino, and T. Winberry (2025). The macroeconomic dynamics of labor market policies. Working paper, October 2025.

Jarosch, G., J. Nimczik, and I. Sorkin (2023). Granular search, market power, and wage inequality. *Quarterly Journal of Economics 138*(4), 2027–2080.

Kaas, L. and P. Kircher (2015). Efficient firm dynamics in a frictional labor market. *American Economic Review 105*(10), 3030–3060.

Lester, B. (2010). Directed search with multi-vacancy firms. *Journal of Economic Dynamics and Control 34*(4), 593–604.

Marinescu, I. and D. Skandalis (2020). Opening the black box of the application process. *Working paper*.

McLeay, M. and S. Tenreyro (2020). Optimal inflation and the identification of the Phillips Curve. *NBER Macroeconomics Annual 34*, 199–255.

Mertens, M. and B. Mottironi (2025). Do larger firms exert more market power? markups and markdowns along the size distribution. *Economics Letters*, 112500.

Moen, E. R. (1997). Competitive search equilibrium. *Journal of Political Economy 105*(2), 385–411.

Nakamura, E. and J. Steinsson (2014). Fiscal stimulus in a monetary union: Evidence from US regions. *American Economic Review 104*(3), 753–792.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica 25*(100), 283–299.

Prager, E. and M. Schmitt (2021). Employer consolidation and wages: Evidence from hospitals. *American Economic Review 111*(2), 397–427.

Seegmiller, B. (2021). Valuing labor market power: The role of productivity advantages. Working paper.

Wright, R., P. Kircher, and B. Julien (2017). Directed search: A structural approach. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Monetary Economics*, Volume 3, pp. 487–554. Elsevier.

Yeh, C., C. Macaluso, B. Hershbein, and L. B. Kahn (2022). Concentration in u.s. labor markets: Evidence from online vacancy data. *Journal of Labor Economics 40*(S1), S205–S247.

Online Appendix

A.1 Existence and Uniqueness in the Homogeneous Example

Economic objective. This subsection proves that the symmetric benchmark in Section 2 is well defined: there is a unique symmetric pure-strategy equilibrium and every ex ante identical firm is active in that equilibrium.

Key assumptions. Firms are ex ante identical, workers direct applications toward the highest expected income, $m(\theta, 1)$ is continuously differentiable, strictly increasing, and strictly concave, $m(0, 1) = 0$, and $\varepsilon(\theta) \in (0, 1)$.

Contribution to the paper. The result justifies using the homogeneous example as a benchmark for the rest of the paper: the markdown and vacancy formulas in the text are not local artifacts but characterize a unique symmetric outcome.

Consider a candidate symmetric profile with exactly $K \leq M$ active firms. Symmetry implies that each active firm receives

$$l = \frac{1}{K}, \qquad v = \frac{\theta}{K}, \tag{.1}$$

and promises the same expected income

$$U = w\, m(\theta, 1). \tag{.2}$$

The profit of an active firm can therefore be written as

$$\pi(K, \theta, U) = \frac{1}{K} \left[z\, m(\theta, 1) - U - c\theta \right]. \tag{.3}$$

For a given promised income U, the firm's problem is strictly concave in θ because $m(\theta, 1)$ is strictly concave. The first-order condition is

$$z\, m_\theta(\theta, 1) = c. \tag{.4}$$

By the maintained boundary conditions, this equation has a unique interior solution $\theta^* > 0$.

Let $\varepsilon^* \equiv \varepsilon(\theta^*)$. In a symmetric profile with K active firms, the wage first-order condition from Section 2 becomes

$$w(K) = \frac{(1 - \varepsilon^*)\left(1 - \frac{1}{K}\right)}{\varepsilon^* + (1 - \varepsilon^*)\left(1 - \frac{1}{K}\right)} z. \tag{.5}$$

Hence, for any given number of active firms K, there is at most one symmetric active-firm profile:

θ^* is unique, $w(K)$ is unique, and $v(K) = \theta^*/K$.

It remains to show that all firms are active. Use the identities

$$c = z\, m_\theta(\theta^*, 1), \qquad \varepsilon^* = \frac{\theta^* m_\theta(\theta^*, 1)}{m(\theta^*, 1)}.$$

Then the profit of an active firm in the symmetric K-firm profile is

$$\pi(K) = \frac{1}{K}\left[(z - w(K))m(\theta^*, 1) - c\theta^*\right] \tag{.6}$$

$$= \frac{z\, m_\theta(\theta^*, 1)\theta^*}{K}\left[\frac{1}{\varepsilon^* + (1-\varepsilon^*)\left(1 - \frac{1}{K}\right)} - 1\right]. \tag{.7}$$

Because $\varepsilon^* \in (0, 1)$ and $K \geq 1$, the bracketed term is strictly positive, so $\pi(K) > 0$.

Now suppose by contradiction that $K < M$, so at least one firm is inactive. An inactive firm earns zero because $m(0, v) = 0$ implies that posting vacancies without applicants generates no hires. If that inactive firm deviates by mimicking the common posting of the active firms, then $K + 1$ firms offer the same expected income and each attracts $1/(K + 1)$ applicants. The deviating firm's profit is therefore $\pi(K + 1) > 0$, a contradiction. Hence $K = M$.

Therefore all firms are active in equilibrium. Setting $K = M$ yields the unique symmetric pure-strategy equilibrium stated in Proposition 1.

A.2 Markdown Formula and Vacancy Invariance

Economic objective. This subsection derives the sufficient-statistics markdown formula in the finite-firm economy and explains why the vacancy-posting first-order condition is unchanged relative to competitive search.

Key assumptions. Workers choose among firms through the logit application rule, firms take rival postings as given, matching is constant returns to scale within each firm, and the equilibrium is interior.

Contribution to the paper. The derivation shows exactly where imperfect competition enters the equilibrium system: it distorts the wage-setting margin through firms' effect on workers' outside options, but it does not alter the surplus condition governing vacancies.

Denote

$$S_{ij} \equiv \beta(w_{ij} - b_j)m_{ij}\varepsilon_{ij}, \qquad A_{ij} \equiv \frac{\sum_{k \neq i} \frac{l_{kj}}{1+\beta(w_{kj}-b_j)m_{kj}\varepsilon_{kj}}}{\sum_k \frac{l_{kj}}{1+\beta(w_{kj}-b_j)m_{kj}\varepsilon_{kj}}}. \tag{.8}$$

The object A_{ij} is the accessibility-adjusted residual market share left to rival firms. It is the finite-firm term that disappears under competitive search.

Using the logit application system and differentiating workers' indirect utility with respect to firm i's own vacancy and wage choices yields the tightness elasticities

$$1 - \varepsilon_{\theta_{ij},v_{ij}} = \frac{S_{ij}}{1+S_{ij}} A_{ij}, \qquad \varepsilon_{\theta_{ij},w_{ij}} = -\frac{\beta w_{ij} m_{ij}}{1+S_{ij}} A_{ij}. \tag{.9}$$

These expressions are the direct analogues of the intermediate derivations in the longer 'v6' draft. The vacancy elasticity depends on $(w_{ij} - b_j)$ because vacancy posting changes worker utility only through matching probabilities, whereas the wage elasticity depends on the full wage bill because a wage change affects utility directly.

Let

$$MPL_{ij} \equiv \frac{\alpha_j z_i y_i}{m_{ij} l_{ij}}. \tag{.10}$$

Firm i's first-order conditions with respect to w_{ij} and v_{ij} can be written as

$$MPL_{ij} - w_{ij} = \frac{w_{ij}}{(1-\varepsilon_{ij})\frac{\beta w_{ij} m_{ij}}{1+S_{ij}} A_{ij}}, \tag{.11}$$

and

$$MPL_{ij} - w_{ij} = \frac{c/q_{ij}}{\varepsilon_{ij} + (1-\varepsilon_{ij})\frac{S_{ij}}{1+S_{ij}} A_{ij}}, \qquad q_{ij} \equiv \frac{m_{ij}}{\theta_{ij}}. \tag{.12}$$

The first expression comes from the wage first-order condition and isolates the market-power wedge; the second comes from the vacancy condition and uses the CRS identity linking hires to vacancy yields.

Rearranging the wage equation gives

$$\frac{w_{ij}}{MPL_{ij}} = \frac{(1-\varepsilon_{ij})\beta w_{ij} m_{ij} A_{ij}}{1 + \beta(w_{ij} - b_j) m_{ij} \varepsilon_{ij} + (1-\varepsilon_{ij})\beta w_{ij} m_{ij} A_{ij}}, \tag{.13}$$

which is Equation (3.9). Relative to competitive search, the only new sufficient statistic is A_{ij}.

To recover the vacancy equation, combine the same two first-order conditions and use $q_{ij} v_{ij} = m_{ij} l_{ij}$. The accessibility term A_{ij} cancels, leaving

$$c v_{ij} = \varepsilon_{ij} \Big(\alpha_j z_i y_i - m_{ij} l_{ij} b_j \Big). \tag{.14}$$

This is the vacancy-invariance result in Section 3.4. Intuitively, a wage change alters the value of search for all workers who compare firm i to rival firms, whereas an additional vacancy affects hiring only through the within-firm matching technology. With CRS matching, that congestion

effect is summarized entirely by ε_{ij}, so the strategic term that appears in the wage equation vanishes from the vacancy equation.

A.3 No-Search-Friction Limit

Economic objective. This subsection shows how the finite-firm directed-search model nests a frictionless oligopsony wedge.

Key assumptions. Take the matching elasticity $\varepsilon_{ij} \to 0$ and consider the limiting case in which matching frictions vanish so that $m_{ij} \to 1$.

Contribution to the paper. The limit clarifies the connection between the present framework and reduced-form oligopsony environments without directed-search frictions.

Letting $\varepsilon_{ij} \to 0$ in the markdown formula yields

$$\frac{w_{ij}}{MPL_{ij}} = \frac{\beta w_{ij} m_{ij} A_{ij}}{1 + \beta w_{ij} m_{ij} A_{ij}}. \tag{.15}$$

If matching is frictionless, so $m_{ij} = 1$, this simplifies to

$$\frac{w_{ij}}{MPL_{ij}} = \frac{\beta w_{ij} A_{ij}}{1 + \beta w_{ij} A_{ij}}, \qquad MPL_{ij} - w_{ij} = \frac{1}{\beta A_{ij}}. \tag{.16}$$

The wage wedge is then determined solely by the elasticity of substitution encoded by β and by the firm's accessibility-adjusted residual share A_{ij}. Under competitive search, $A_{ij} = 1$ and the firm takes workers' outside option as given. With finitely many firms, $A_{ij} < 1$, so each firm internalizes that a wage change shifts the value of search faced by inframarginal applicants. This is the sense in which the model nests the finite-employer wedge emphasized by Berger et al. (2022).

A.4 Beveridge-Curve Inequality

Economic objective. This subsection provides the inequality underlying the outward Beveridge-curve shift in the heterogeneous-firm economy.

Key assumptions. Set $N = 1$, $b = 0$, and $m_i = e\theta_i^{\varepsilon}$ with common matching elasticity $\varepsilon \in (0, 1)$.

Contribution to the paper. The argument shows that dispersion in firm-level tightness is sufficient to lower aggregate employment for a given vacancy stock.

Using $\theta_i = v_i/l_i$, employment equals

$$H = e \sum_i v_i^{\varepsilon} l_i^{1-\varepsilon}. \tag{.17}$$

Because $\sum_i l_i = 1$, Hölder's inequality implies

$$H = e \sum_i v_i^{\varepsilon} l_i^{1-\varepsilon} \leq e\Big(\sum_i v_i\Big)^{\varepsilon} \Big(\sum_i l_i\Big)^{1-\varepsilon} = eV^{\varepsilon}, \qquad V \equiv \sum_i v_i. \tag{.18}$$

Equality holds if and only if v_i/l_i is constant across firms, that is, if tightness is equalized. Under competitive search, identical firms within a location choose the same tightness, so the upper bound is attained. Under oligopsony with heterogeneous productivity, high-productivity firms attract more applicants and post more vacancies, but not in exact proportion, so tightness disperses across firms. The inequality is then strict, implying lower employment and therefore higher unemployment for a given aggregate vacancy stock. Aggregating across segmented locations yields the outward Beveridge-curve shift emphasized in the text.

A.5 Welfare Benchmark

Economic objective. This subsection states the planner's problem that serves as the natural welfare benchmark for the decentralized equilibrium.

Key assumptions. The planner chooses firm-level wage–vacancy menus within each segmented location, takes the same matching technology and outside options as given, and respects the logit application structure induced by idiosyncratic utility shocks.

Contribution to the paper. The planner problem clarifies which distortions are created by finite-firm wage setting and which objects are merely transfers.

Within a location, the planner maximizes total output net of vacancy costs plus the expected utility of workers:

$$\mathcal{W} = \sum_i z_i f(\{m_{ij}(\theta_{ij}, 1) l_{ij}\}_j) - c \sum_{i,j} v_{ij} + \sum_j \left[\frac{\gamma}{\beta} + \frac{1}{\beta} \log\left(\sum_i e^{\beta u_{ij}} \right) \right], \tag{.19}$$

where

$$u_{ij} = m_{ij}(\theta_{ij}, 1) w_{ij} + \big(1 - m_{ij}(\theta_{ij}, 1)\big) b_j. \tag{.20}$$

The wage bill does not appear separately in the welfare criterion because wages are transfers

between firms and workers; what matters socially is how the posted wage affects worker utility and hence the allocation of applicants across firms.

The planner's first-order conditions are the analogues of the equilibrium conditions in the 'v6' draft. They imply two broad conclusions. First, the planner values an additional vacancy by its contribution to total surplus rather than by the firm's private surplus alone. Second, the planner internalizes that a wage change at one firm reallocates applicants across all firms in the location. The decentralized equilibrium does not internalize this cross-firm effect, so positive markdowns and excessive concentration of applicants can arise even when the vacancy condition is unchanged. This benchmark is the appropriate object for future welfare calculations of mergers, entry, or changes in unemployment insurance.