# An Information-based Theory of Monopsony Power

Anton Cheremukhin*
Paulina Restrepo-Echavarria†

March 26, 2026

**Abstract**

We develop a tractable model of monopsony power based on information frictions in job search. Workers and firms choose probabilistic search strategies, with information costs limiting how precisely they can target matches. Firms post wages strategically, anticipating application behavior and exploiting a first-mover advantage. The model nests both directed and random search and yields a closed-form wage equation showing how monopsony power depends on the equilibrium sorting regime. We show that strong assortative matching can emerge even when wage incentives are compressed, if sorting is driven by firm screening rather than worker self-selection. Estimating the model on moments from the Swedish labor market, we find the market operates precisely in this firm-driven sorting regime. This regime, characterized by extreme monopsony and a wage-productivity disconnect, simultaneously resolves the empirical puzzles of strong sorting with flat wage profiles and large markdowns with small firm fixed effects.

*Federal Reserve Bank of Dallas, 2200 N Pearl St, Dallas TX 75201, chertosha@gmail.com
†Federal Reserve Bank of St. Louis, One Federal Reserve Bank Plaza, Broadway and Locust St., St. Louis MO 63166, paulinares@me.com
Declarations of interest: none. We thank David Berger, Olivier Coibion, Kyle Herkenhoff, Benjamin Lester, Guillaume Rocheteau, Esteban Rossi-Hansberg, Aysegul Sahin, Kjetil Storesletten, Gianluca Violante, as well as participants of the seminar at the Dallas Fed, St. Louis Fed, and SED meetings in Copenhagen for their insightful comments. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Reserve Banks of Dallas and St. Louis or the Federal Reserve System.

# 1 Introduction

A central insight of models of monopsony is that firms set wages below marginal product, either because workers face search frictions or because jobs are differentiated (Card, 2022). Yet canonical models typically treat these sources in isolation. In this paper, we develop an information-based theory of monopsony power that unifies the search and job differentiation perspectives in a single framework. Our model endogenizes both the targeting precision of job search and the elasticity of firm access to labor, yielding new insights into the determinants of wage-setting power and resolving several empirical puzzles in the literature on sorting and inequality.

Recent empirical work using matched employer-employee data has uncovered a central puzzle: labor markets exhibit strong positive assortative matching, yet wage profiles are surprisingly flat with respect to firm productivity (Bonhomme, Lamadon, and Manresa, 2019, henceforth BLM). So, if wage signals are too weak to incentivize workers to find better jobs, what drives the strong sorting patterns observed in the data? Standard models struggle to reconcile this. Directed search models generate sorting via steep wage gradients, while random search models generate flat wages but struggle to produce sorting. In this paper we develop a model that can rationalize this puzzle when sorting is driven by firm screening and not by worker self-selection.

We build on Cheremukhin, Restrepo-Echavarria, and Tutino (2020), who introduced a model of simultaneous targeted search under information constraints, more applicable in the context of marriage markets. Here, we adapt and extend that framework to the labor market, introducing sequential wage posting and strategic firm behavior. Workers and firms choose probabilistic search strategies, deciding how precisely to target a partner type. These decisions are subject to endogenous information frictions, modeled as relative entropy (Kullback-Leibler divergence) between a uniform prior and the chosen strategy. This rational inattention specification provides a microfoundation for the multinomial logit (MNL) discrete choice structure frequently assumed in the literature on job differentiation.

Our framework nests both directed and random search as limiting cases. When information costs are zero, workers can fully target their desired firms types, replicating the classic directed search environment. When costs are infinite, search is random and matches occur with uniform probability. For intermediate levels of costs, agents choose

imperfectly targeted strategies, and wages posted by firms reflect their expectations about the number and composition of applicants, as well as the costliness of screening. The model preserves the key features of directed search: sequential wage posting, strategic attraction of worker types, and matching via submarket-specific technologies, but introduces information frictions which span the continuum between directed and random search.

We derive a closed-form expression for the equilibrium wage that isolates the fundamental drivers of monopsony power. In our model, monopsony arises from five distinct sources: (i) worker-side search costs affect the precision with which wage compensation attracts workers and determine the elasticity of labor supply; (ii) firm-side screening costs shape firms' willingness to screen applicants and determine the elasticity of labor demand; (iii) labor market tightness amplifies both elasticities through numerical imbalances; (iv) equilibrium sorting, where positive assortative matching reduces wage competition across firms and increases firm surplus shares; and (v) the sequential nature of search, which gives firms a first-mover advantage by letting them strategically commit to wage menus. A key finding is that wage markdowns are not a simple function of labor supply elasticity; they are profoundly shaped by equilibrium sorting patterns.

To test this mechanism, we estimate the model to match a rich set of moments from the Swedish labor market, as documented by BLM. We find that the model can simultaneously replicate the seemingly contradictory facts of strong sorting, flat wage profiles, and the moments from Abowd, Kramarz, and Margolis (1999, henceforth AKM): high worker fixed effects, low firm fixed effects, high sorting correlation. The estimation reveals that the market operates in a firm-driven sorting regime: firms exercise extreme monopsony power, paying wages near workers' outside options. This flattens the wage gradient, leaving workers with weak incentives to direct their search. Sorting is instead enforced by firms, who use their screening technology to "cherry-pick" the most productive workers from the diffuse pool of applicants. This mechanism explains how the firm fixed effect on productivity can be hidden in profits rather than passed through to wages.

Our contributions are thus threefold. First, we propose a unified model of monopsony that endogenizes the precision of search on both sides of the market. Second, we identify a novel mechanism of firm-driven sorting that resolves a central puzzle in the empirical literature on sorting and wages. Third, by estimating the model, we provide

quantitative evidence of this mechanism in action and characterize the resulting misallocation, finding that the market systematically under-produces the highest-surplus matches. These contributions respond to recent calls for more structurally grounded models of monopsony that integrate strategic interactions, sorting, and information frictions (Card, 2022; Azar and Marinescu, 2024).

The paper proceeds as follows. Section 2 presents the model of sequential targeted search with endogenous information costs and characterizes the competitive equilibria and the planner's solution. Section 3 analyzes the mechanics of equilibrium wages and sorting, contrasting the worker-driven and firm-driven sorting regimes. Section 4 presents the empirical estimation and its implications for the monopsony puzzles. Section 5 concludes.

## Related literature

This paper relates to several strands of the economic literature. First, it contributes to the directed search literature, which studies markets where firms post wages and workers choose where to apply (e.g., Moen, 1997; Burdett, Shi, and Wright, 2001; Shi, 2001). Standard directed search models assume perfect information and costless targeting, leading to submarket-based equilibria that are often efficient and feature assortative matching (see Eeckhout and Kircher, 2010; Guerrieri, Julien, Kircher, and Wright, 2021). Our framework generalizes this class of models by introducing information costs that determine how precisely workers and firms can target one another. This friction generates a continuum between directed and random search and allows us to recover standard directed search as a limiting case. We explore how the forces of submarket diffusion, coordination, and bounded information processing interact to determine equilibrium properties. In doing so, we highlight a fundamental tension in frictional markets: while positive assortative matching is often technologically efficient, it systematically increases firms' monopsony power. Because highly sorted workers lack viable alternative matches, sorting eliminates effective wage competition and allows firms to extract larger markdowns. A broader perspective on these modeling choices and their implications for sorting, matching technologies, and equilibrium efficiency is provided in Chade, Eeckhout, and Smith (2017).

Second, we contribute to the monopsony literature. Recent empirical studies docu-

ment substantial wage markdowns across a wide range of labor markets (e.g., Staiger, Spetz, and Phibbs, 2010; Azar, Marinescu, and Steinbaum, 2019; Lamadon, Mogstad, and Setzler, 2022; Dube, Naidu and Reich 2022), often attributing them to limited labor supply elasticity. Theoretical models typically attribute monopsony power to search frictions (e.g., Burdett and Mortensen, 1998), job differentiation (Manning, 2003), or oligopsony (Robinson, 1933; Boal and Ransom, 1997), and have quantified the aggregate welfare losses due to labor market power as large (Berger, Herkenhoff, and Mongey, 2022). We provide a tractable framework that unifies these mechanisms and yields a closed-form wage equation decomposing the sources of monopsony power into supply-side frictions, demand-side frictions, labor market tightness, and equilibrium sorting.

Several recent papers attempt to integrate key elements of monopsony: search frictions, heterogeneity, and concentration — into unified models (e.g., Lamadon, Mogstad, and Setzler, 2022; Berger, Herkenhoff, Kostol, and Mongey, 2023). Our work is also related to Bils, Kaymak, and Wu (2025), who show that comparative advantage can be a source of monopsony power by reducing the set of relevant outside options for sorted workers. While these frameworks are powerful, they often treat the relationship between sorting, elasticities, and wage setting as largely mechanical. In contrast, our model captures the full strategic interaction between search precision, wage posting, and equilibrium sorting. This allows us to resolve the tension between strong sorting and flat wages by showing that sorting can be driven by firms' non-wage screening efforts when wage incentives are compressed by monopsony power. Sorting thus becomes a central determinant of wage markdowns, not just match efficiency.

Several papers have attempted to nest directed and random search to generate intermediate degrees of randomness in matching (e.g., Menzio, 2007; Lester, 2011; Lentz, Maibom and Moen, 2022) and study the efficiency consequences (Rabinovich and Wolthoff, 2022). Unlike these models, which impose hybrid structures, our framework endogenizes the degree of search precision via information costs. This yields a continuous spectrum between uniform random matching and directed search within a unified model. Relatedly, Wu (2024) studies a version of our targeted search model in a labor market setting, but does not characterize the structure of monopsony power or use it to resolve empirical puzzles.

Methodologically, our approach connects to recent work that convexifies discrete assignment or search problems using regularization techniques. Our use of information

costs mirrors entropic regularization in optimal transport (Cuturi, 2013) and convex approximations in spatial assignment models (Oberfield, Rossi-Hansberg, Sarte, and Trachter 2024). This allows us to efficiently compute equilibria in environments where standard directed search models exhibit indeterminacy or are difficult to solve.

Finally, we build on our earlier work in Cheremukhin, Restrepo-Echavarria, and Tutino (2020), which studied targeted search in a simultaneous-move setting with bargaining. The current paper shifts to sequential search with wage posting, focuses on the labor market rather than the marriage market, and introduces the firm-driven sorting mechanism as a key explanatory channel for observed labor market outcomes. In doing so, we reconcile the tractability of multinomial logit-type models with the equilibrium structure of directed search and monopsony.

# 2 Model

We develop a model of a frictional labor market in which heterogeneous workers and firms strategically search for each other. Our model labor market is characterized by two key features: information frictions and sequential strategic interaction. We first describe the environment, the sequential matching game, and the equilibrium concept. Then we characterize equilibrium properties.

### The Economic Environment

The economy contains a large, finite number of workers indexed by type $x \in \{1, ..., W\}$ and firms indexed by type $y \in \{1, ..., F\}$. The masses of agents are denoted by $\mu_x$ and $\mu_y$, respectively.

A match between a worker of type $x$ and a firm of type $y$ generates a non-negative surplus $f_{xy}$. If a match is formed, the worker receives a wage $\omega_{xy}$ and the firm retains profit $f_{xy} - \omega_{xy}$. Without loss of generality, we normalize the outside options of workers to zero.

**Assumption 1.** The payoffs are non-negative: $f_{xy} \geq \omega_{xy} \geq 0$.

The matching process unfolds in three stages. First, firms post wage menus to attract workers. Second, workers observe these wages and choose probabilistic application strategies. Third, firms observe their applicant queues and use probabilistic screening

strategies to select workers. We now describe the search strategies, the information frictions that govern these choices, and the technology that transforms choices into matches.

**Search Strategies and Information Costs.** Agents face an information friction: they cannot perfectly and costlessly target specific counterparts. Instead, they choose a search strategy (a probability distribution of whom to select) subject to an entropic cost.This cost reflects the effort required to deviate from a random search strategy (a uniform prior).

A worker of type $x$ chooses a search strategy $\bar{p}_x(y)$, representing the probability of applying to a firm of type $y$. The worker's "default" behavior is random application across all firms, represented by the uniform prior $\tilde{p}_x(y) = 1/\delta_x$, where $\delta_x = \sum_y \mu_y$ is the total number of firms. The cost of search is a function of the search effort, measured as the relative entropy (Kullback-Leibler divergence) between the chosen strategy and this prior.[1] Letting $p_x(y) = \bar{p}_x(y)/\tilde{p}_x(y)$ denote the search intensity, the workers' search effort then equals

$$\kappa_x = \sum_y \frac{\mu_y}{\delta_x} p_x(y) \ln p_x(y).$$

Symmetrically, a firm of type $y$ chooses a strategy $\bar{q}_y(x)$, representing the probability of screening/selecting a worker of type $x$ *from within its queue of applicants.* This is a crucial distinction: firms do not screen the general population; they screen only those who apply to them. The firm's prior is the empirical composition of its applicant queue, not the population distribution. The queue composition is defined by weights $a_{xy} = \frac{\mu_x p_x(y)}{\delta_y}$, where $\delta_y = \sum_x \mu_x p_x(y)$ is the effective (expected) queue length. The firm's screening effort is

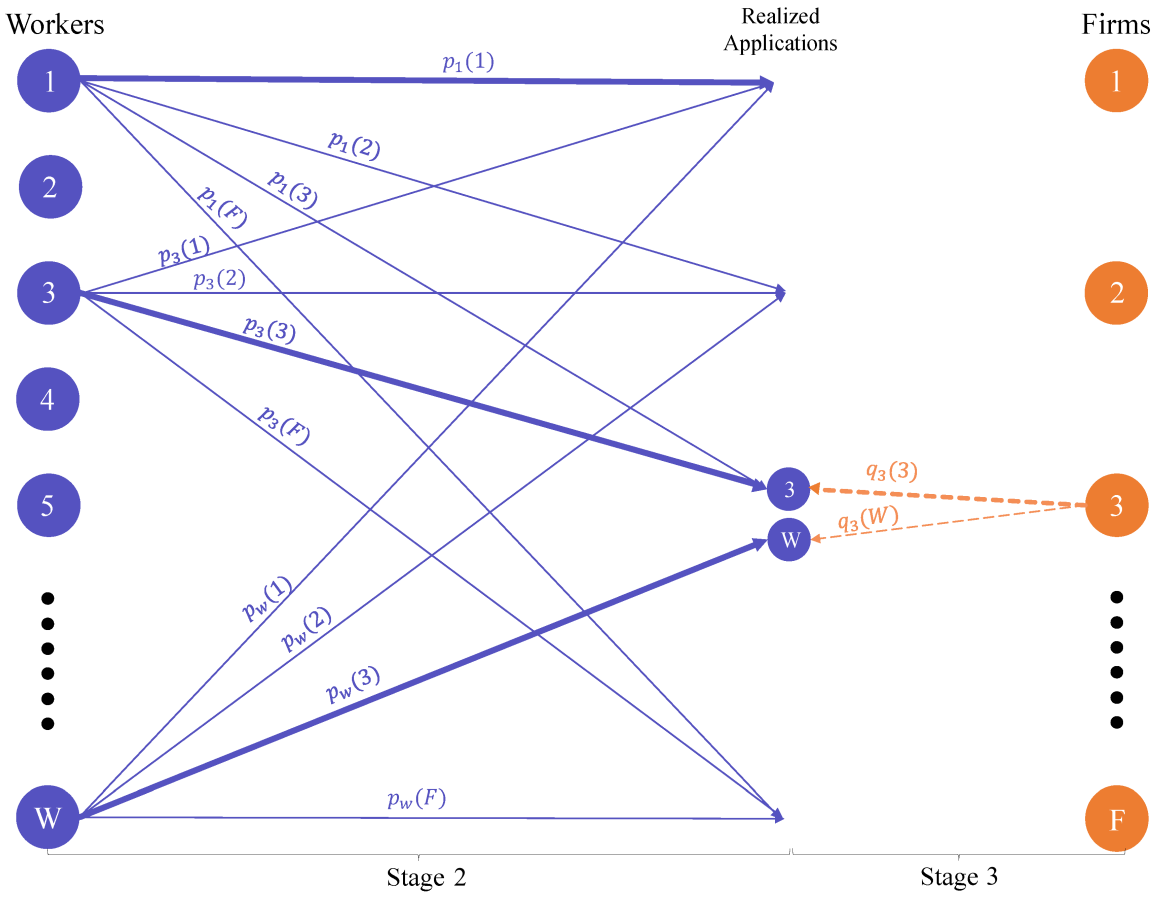
$$\kappa_y = \sum_x a_{xy} q_y(x) \ln q_y(x),$$

where $q_y(x)$ is the screening intensity relative to the prior.[2]

[1] In this case, the relative entropy measures the amount of information processed by the agent. This definition is a special case of Shannon's channel capacity, where information structure is the only choice variable (See Thomas and Cover (1991), Chapter 2). See also Cheremukhin, Popova, and Tutino (2015) and Matejka and McKay (2015) for applications to stochastic discrete choice with information costs.

[2] We have chosen to write the model in terms of intensities rather than probabilities purely for mathematical and computational convenience.

Figure 2.1: Strategies of Workers and Firms under Sequential Targeted Search




**Assumption 2.** Search costs $c_i(\kappa)$ are strictly increasing, twice continuously differentiable, and convex in search effort. We assume linear costs $c_x = \theta_x \kappa_x$ and $c_y = \theta_y \kappa_y$.

Figure 2.1 illustrates the interactions and search strategies of workers and firms. The solid arrows show the intensity $p_x(y)$ that a worker of type $x$ assigns to targeting a firm of type $y$. Similarly, dashed arrows show the intensity $q_y(x)$ that a firm of type $y$ assigns to targeting a worker of type $x$ in its queue.

**The Matching Technology and Congestion.** How do probabilistic choice strategies translate into actual hires? If we simply multiplied probabilities, the number of matches would exhibit increasing returns to scale. To capture the physical constraints of the meeting process, we introduce a submarket-specific matching technology.

Consider a submarket for the pair $(x, y)$. The mass of workers of type $x$ applying

to firms of type $y$ is determined by the worker's strategy: $c_{xy} = \mu_x p_x(y) \frac{\mu_y}{\delta_x}$. The term $\mu_y/\delta_x$ converts individual targeting intensities into mass of applications, reflecting that workers choose among firm types rather than individual firms. Similarly, the mass of firms of type $y$ that has screened workers of type $x$ out of their queue is determined by the firm's screening strategy and the queue composition: $d_{xy} = \mu_y q_y(x) a_{xy}$. As this is a submarket with homogeneous firms and workers, it would be natural to expect imperfect coordination between workers and firms regarding who applied where, and who picked whom. As is common in the literature, we assume that the number of matches in a submarket is determined by a matching function $M(c_{xy}, d_{xy})$, where $c_{xy}$ and $d_{xy}$ are the effective masses of searchers from each side defined above. Because queue composition structurally links $c_{xy}$ and $d_{xy}$, we interpret the matching function $M(\cdot, \cdot)$ as a reduced-form aggregator of the submarket's queuing capacity, capturing the physical bottlenecks of the interview process rather than a mechanical urn-ball arrival process.

To facilitate the individual optimization problems, we rewrite this technology via a *congestion function* $\phi_{xy}$. This function acts as a market-clearing price for "meetings." It reconciles the individual search intensities ($p$ and $q$) chosen by the agents with the aggregate number of matches determined by matching technology $M$.[3] The relationship between the total number of matches and the congestion function is given by:

$$M_{x,y} = \mu_x \mu_y p_x(y) q_y(x) \phi(p_x(y), q_y(x), \mu_x, \mu_y).$$

Unlike standard models where atomistic agents take matching rates as given, we assume that workers and firms rationally internalize the physical congestion of the matching process. They explicitly account for how their own individual search and screening strategies marginally affect the submarket congestion rates $\phi_{xy}$.

**Assumption 3.** The matching function $M$ (and congestion function $\phi$) are twice continuously differentiable in strategies $p$ and $q$.

**Assumption 4.** Agents internalize the endogenous matching rates ($\phi$). Specifically, they account for the elasticities of the congestion function with respect to their chosen strategies, $\varepsilon_{\phi,p} \equiv \frac{\partial \phi_{xy}}{\partial p_x(y)} \frac{p_x(y)}{\phi_{xy}}$ and $\varepsilon_{\phi,q} \equiv \frac{\partial \phi_{xy}}{\partial q_y(x)} \frac{q_y(x)}{\phi_{xy}}$.

[3] We introduce this congestion function in the spirit of Mortensen (1982) and Shimer and Smith (2001), who assume a linear search technology.

From the properties of the matching function, agents internalize the elasticities of the total match probability with respect to applications and screening, denoted as $E_{px} \equiv 1 + \varepsilon_{\phi,p} = \varepsilon_{m,c} + (1 - a_{xy})\varepsilon_{m,d}$, and $E_{qy} \equiv 1 + \varepsilon_{\phi,q} = \varepsilon_{m,d}$. [4]

**Sequential Choices**

The matching process unfolds as a three-stage game that gives firms a first-mover advantage in wage setting. Firms initiate the process by posting job openings. Each firm of type $y$ commits to a contract consisting of a type-dependent wage menu, $\{\omega_{xy}\}_{x=1}^{W}$, which specifies the wage it will pay to a worker of type $x$ upon matching, and a verifiable, automated screening/testing policy, $\{q_y(x)\}_{x=1}^{W}$. Firms choose these contracts strategically, anticipating how they will influence application decisions.

In the second stage, after observing all posted wage menus and screening policies, workers decide where to send their applications. As described above, workers face an information friction that makes it costly to perfectly target their preferred firm type. Each worker chooses a probabilistic application strategy ($p$) introduced earlier.

In the third stage, applications are sunk and the firm's applicant queue has been formed. A firm of type $y$ knows the expected length and composition of its queue, but not the specific type of each individual applicant.

To test applicants, the firm executes its pre-committed screening policy ($q$), a probabilistic screening strategy that determines the probability of considering the application of a worker of type x from among those who applied. Screening resolves the information friction of the firms, revealing the specific type of the applicant prior to the final match. By committing to $q$ in the first stage alongside wages, firms overcome the time-inconsistency problem that would arise if they could re-optimize screening after applications are sunk.

The game is played in terms of expected payoffs. Agents choose probability distributions $p$ and $q$ to maximize expected utility, effectively smoothing over the discrete nature of final matches. At the end of the third stage, after the agents have sequentially chosen wages and probabilistic search strategies, the actual draws from the distributions are taken and matches are formed according to the matching functions. At this stage, the matched agents learn each others' types, receive the corresponding surplus

[4] An alternative assumption 4′ where workers and firms take congestion $\phi$ as given would be captured by assuming $E_{px} = E_{qy} = 1$.

and split it according to the pre-commited wage contracts. Unmatched agents get a zero payoff. This is a one-shot model: the game ends and no further interactions are considered. Let us now consider the problems of the workers and the firms.

**The Worker's Problem.** We denote by $S_x$ the set of admissible choice strategies for a worker of type $x$ capturing the constraint that probabilities must sum to one (intensities average to one):

$$S_x = \left\{ p_x(y) \in R_+^F : \sum_{y=1}^{F} \frac{\mu_y}{\delta_x} p_x(y) \leq 1 \right\},$$

where $\delta_x = \sum_{y=1}^{F} \mu_y$ is the total number of firms in the market (which is equal to the possible set of choices for worker $x$).

In the second stage, workers take posted wages $\{\omega_{xy}\}$ and the firm's screening intensities $q_y(x)$ as given. Crucially, they internalize how their own application intensity $p_x(y)$ affects the submarket congestion rate $\phi_{xy}$. They choose $p_x(y) \in S_x$ to maximize expected wage income net of search costs:

$$\max_{p_x \in S_x} \quad \sum_{y=1}^{F} \mu_y q_y(x) \phi_{xy} \cdot \omega_{xy} \cdot p_x(y) - \theta_x \sum_{y=1}^{F} \frac{\mu_y}{\delta_x} p_x(y) \ln p_x(y),$$

**Proposition 1.** *Suppose that search costs are strictly positive ($\theta_x > 0$) and the constraint set $S_x$ is non-empty. Assuming the matching technology exhibits standard diminishing marginal returns to applications, the worker's optimization problem is strictly concave, and there exists a unique interior solution $p_x^*(y)$ characterized by:*

$$p_x^*(y) = \frac{\exp\left(\frac{q_y(x)\phi_{xy}\omega_{xy}E_{px}}{\theta_x/\delta_x}\right)}{\sum_{y'=1}^{F} \frac{\mu_{y'}}{\delta_x} \exp\left(\frac{q_{y'}(x)\phi_{xy'}\omega_{xy'}E_{px'}}{\theta_x/\delta_x}\right)}. \tag{2.1}$$

*Proof.* The worker's objective function has two terms. The first term evaluates the expected wage. Because the worker internalizes congestion, this term depends on the expected match probability $p_x(y)\phi_{xy}(p_x)$, which is strictly increasing but strictly concave in $p_x(y)$ under standard matching technologies. The second term involves the function $f(p) = p \ln p$. Since $f''(p) = 1/p > 0$ for all $p > 0$, the entropic cost function is

strictly convex. Because $\theta_x > 0$, subtracting a strictly convex function from a concave function makes the total objective strictly concave. The constraint set $S_x$ is a simplex, which is a convex compact set. Maximizing a strictly concave function over a convex set guarantees a unique global maximum. Furthermore, since $\lim_{p\to 0} f'(p) = -\infty$, the marginal cost of search approaches $-\infty$ as probability mass vanishes, ensuring that the optimal solution is strictly interior ($p_x^*(y) > 0$ for all $y$).

We set up the Lagrangian with multiplier $\theta_x\lambda_x$ for the budget constraint and then the objective function of the worker is:

$$Y_x = \sum_{y=1}^{F} \mu_y p_x(y)\, q_y(x)\, \phi_{xy}\omega_{xy} - \theta_x \sum_{y=1}^{F} \frac{\mu_y}{\delta_x} p_x(y) \left(\ln p_x(y) + \lambda_x\right) + \theta_x\lambda_x,$$

The first-order (necessary and sufficient) condition with respect to $p_x(y)$ requires applying the product rule to account for the congestion elasticity:

$$\frac{\partial Y_x}{\partial p_x(y)} = \mu_y q_y(x)\omega_{xy}\left(\phi_{xy} + p_x(y)\frac{\partial \phi_{xy}}{\partial p_x(y)}\right) - \theta_x \frac{\mu_y}{\delta_x}(1 + \ln p_x(y) + \lambda_x) = 0.$$

Factoring out $\phi_{xy}$ and substituting the definition of the application congestion multiplier $E_{px} \equiv 1 + \frac{\partial \phi_{xy}}{\partial p_x(y)} \frac{p_x(y)}{\phi_{xy}}$, we get:

$$\mu_y q_y(x)\phi_{xy}\omega_{xy}E_{px} - \theta_x \frac{\mu_y}{\delta_x}(1 + \ln p_x(y) + \lambda_x) = 0.$$

Dividing by $\mu_y$ and rearranging terms, then exponentiating both sides, using the budget constraint to solve for the Lagrange multiplier term, and then substituting it back into the expression for $p_x(y)$ gives the result in (2.1). □

The worker's optimal strategy, characterized by equation (2.1), takes a Multinomial Logit (MNL) form. It shows that workers tilt their application strategy toward firms offering a higher marginal expected return. Because workers internalize market congestion, this marginal return, $q_y(x)\phi_{xy}\omega_{xy}E_{px}$, discounts the average expected value of an application ($q_y(x)\phi_{xy}\omega_{xy}$) by the congestion multiplier $E_{px} \leq 1$ to account for the crowding their own applications generate.

The parameter $\theta_x$ governs the baseline sensitivity of this response. To understand the mechanics of the worker's choice, we define the *simple* semi-elasticity of applications

as $z_{xy}^0 \equiv \frac{q_y(x)\phi_{xy}E_{px}}{\theta_x/\delta_x}\left(1 - \frac{\mu_y}{\delta_x}p_x(y)\right)$. This term captures how workers would respond to wage variations if they evaluated the physical matching rate $\phi_{xy}$ as a fixed constant. When $\theta_x \to \infty$, information costs dominate, the simple elasticity $z_{xy}^0$ approaches zero, and workers apply randomly across all firms.

However, when rational workers internalize that their collective application behavior alters the physical congestion of the submarket, the effective semi-elasticity, $\hat{z}_{xy} \equiv \frac{\partial \ln p_x(y)}{\partial \omega_{xy}}$, is structurally dampened. Differentiating the optimal application strategy yields:

$$\hat{z}_{xy} = \frac{z_{xy}^0}{1 - z_{xy}^0 \omega_{xy}(E_{px} - 1)}.$$

The term $(E_{px} - 1)$ represents the elasticity of the congestion rate with respect to applications ($\varepsilon_{\phi,p}$). Under standard matching technologies (CRS/DRS), this congestion elasticity is strictly negative, meaning the denominator is strictly greater than one. Economically, this captures a rational feedback loop: a higher wage attracts more applicants, which workers anticipate will physically crowd the submarket and lower their probability of being hired. This anticipated crowding acts as a natural brake on their labor supply elasticity.

As $\theta_x \to 0$, information frictions vanish and the simple elasticity $z_{xy}^0$ diverges to infinity. Yet, because workers internalize the congestion they cause, the effective elasticity $\hat{z}_{xy}$ converges to a finite bound, $\frac{-1}{\omega_{xy}(E_{px}-1)}$. By preventing the application response from becoming infinitely elastic, this congestion brake guarantees that firms retain some monopsony power even without information costs. This mechanism allows the model to smoothly bridge the gap between pure random search and the sorting and pricing behaviors characteristic of canonical directed search models.

**The Firm's Problem.** In the first stage, firms choose and commit to a contract package consisting of a type-dependent wage menu $\{\omega_{xy}\}$ and a verifiable screening policy $\{q_y(x)\}$. By committing to both instruments simultaneously, firms overcome the time-inconsistency problem that would arise if screening were optimized ex-post after applications were sunk. Firms possess a first-mover advantage: they fully internalize how their posted contracts affect the workers' application strategies (Eq. 2.1), the endogenous queue composition ($a_{xy}$), and the physical congestion of the submarket

$(\phi_{xy})$.[5]

As with workers, we define the set $S_y$ of admissible strategies for a firm of type $y$, which includes both the screening intensity and the wage menu:

$$S_y = \left\{ q_y(x), \omega_{xy} \in R_+^W : \sum_{x=1}^{W} a_{xy}^* q_y(x) \leq 1, \omega_{xy} \leq f_{xy} \right\}.$$

Like workers, firms choose admissible strategies that maximize their expected profits net of search costs:

$$\max_{\{q_y(x),\omega_{xy}\}\in S_y} \quad \sum_{x=1}^{W} \mu_x p_x^*(y) \phi_{xy} q_y(x) (f_{xy} - \omega_{xy}) - \theta_y \sum_{x=1}^{W} a_{xy}^* q_y(x) \ln q_y(x)$$

It is important to reiterate the interpretation of the firm's choice set and cost function. While the firm knows the aggregate population $\mu_x$, it can only form matches with workers in its queue. The term $a_{xy}$ in the admissibility constraint and in the cost function ensures the firm chooses among workers in their queue and pays to identify specific types *relative* to their prevalence in the application pool, not in the whole population.

**Proposition 2.** *Suppose that screening costs are strictly positive ($\theta_y > 0$). The firm's optimization problem yields necessary first-order conditions that robustly characterize the optimal screening intensities and wages when agents internalize market congestion.*

$$q_y^*(x) = \frac{\exp\left(\frac{\phi_{xy}}{\theta_y/\delta_y}(f_{xy} - \omega_{xy}) E_{qy} - \frac{1-\hat{v}_{xy}G_y}{1+\omega_{xy}\hat{z}_{xy}E_{qy}}\right)}{\sum_{x'=1}^{W} a_{x'y}^* \exp\left(\frac{\phi_{x'y}}{\theta_y/\delta_y}(f_{x'y} - \omega_{x'y}) E_{qy} - \frac{1-\hat{v}_{x'y}G_y}{1+\omega_{x'y}\hat{z}_{x'y}E_{qy}}\right)}, \tag{2.2}$$

$$\omega_{xy}^* = \left[\hat{O}_{xy} + \left(1 - \frac{1}{q_y^*(x)}\right)\frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{z_{xy}^0}\right]_0^{f_{xy}}, \tag{2.3}$$

*where $z_{xy}^0 \equiv \frac{q_y(x)\phi_{xy}E_{px}}{\theta_x/\delta_x}\left(1 - \frac{\mu_y}{\delta_x}p_x(y)\right)$ is the simple semi-elasticity of applications, $\hat{z}_{xy} \equiv \frac{z_{xy}^0}{1-z_{xy}^0\omega_{xy}(E_{px}-1)}$ is the full semi-elasticity bounded by physical congestion, and $\hat{v}_{xy} \equiv \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}$ is the full screening semi-elasticity. The term $\hat{O}_{xy} \equiv f_{xy}\varepsilon_{m,c} + \varepsilon_{m,d}\sum_{x'\neq x} w_{x'|x}^y f_{x'y}$ is the firm's internalized expected outside option evaluated at the relative matching*

[5] This follows the literature on leader-follower problems in the spirit of Stackelberg competition.

*weights* $w^{y}_{x'|x} \equiv a^{*}_{x'y}\frac{q^{*}_{y}(x')\phi_{x'y}}{q^{*}_{y}(x)\phi_{xy}}$, *and* $G_y \equiv \sum_{x'=1}^{W} a^{*}_{x'y}q^{*}_{y}(x')\left(\ln q^{*}_{y}(x') + \lambda_y\right)$ *represents the information cost spillover of attracting a longer queue.*

*Proof.* The firm maximizes expected profit subject to the constraint $\sum_x a^{*}_{xy}q_y(x) \leq 1$. Let $\theta_y\lambda_y$ be the Lagrange multiplier. Using the aggregate queue length $\delta_y = \sum_x \mu_x p^{*}_{x}(y)$ and queue weights $a^{*}_{xy} = \mu_x p^{*}_{x}(y)/\delta_y$, the objective function is:

$$Y_y = \delta_y \sum_{x=1}^{W} a^{*}_{xy}q_y\left(x\right)\phi_{xy}\left(f_{xy} - \omega_{xy}\right) - \theta_y \sum_{x=1}^{W} a^{*}_{xy}q_y(x)(\ln q_y(x) + \lambda_y) + \theta_y\lambda_y.$$

The firm internalizes the worker's application response and the physical congestion. Differentiating $\ln p^{*}_{x}(y)$ with respect to screening $q_y(x)$ yields the full screening semi-elasticity $\hat{v}_{xy} = \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}$. Because changing $q_y(x)$ alters the mass of type $x$ applications, it alters the denominator ($\delta_y$) of every weight in the queue. By the quotient rule, the cross-type derivatives are:

$$\frac{\partial a^{*}_{xy}}{\partial q_y(x)} = a^{*}_{xy}\left(1 - a^{*}_{xy}\right)\hat{v}_{xy}, \quad \text{and} \quad \frac{\partial a^{*}_{x'y}}{\partial q_y(x)} = -a^{*}_{x'y}a^{*}_{xy}\hat{v}_{xy} \quad \text{for } x' \neq x.$$

Differentiating $Y_y$ with respect to $q_y(x)$, applying these cross-type derivatives, dividing by $\delta_y a^{*}_{xy}$, and substituting $\hat{v}_{xy} = \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}$ yields the marginal condition:

$$\frac{\theta_y}{\delta_y}(\ln q_y(x)+1+\lambda_y) = \phi_{xy}(f_{xy}-\omega_{xy})E_{qy}\left(1+\omega_{xy}\hat{z}_{xy}E_{px}\right) - \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}\frac{\theta_y}{\delta_y}\left(q_y(x)(\ln q_y(x) + \lambda_y) - G_y\right).$$

where we define $G_y \equiv \sum_{x'} a^{*}_{x'y}q_y(x')(\ln q_y(x') + \lambda_y)$. Algebraic rearrangement isolates the logarithmic screening intensity:

$$\ln q_y(x) + \lambda_y = \frac{\delta_y}{\theta_y}\phi_{xy}(f_{xy} - \omega_{xy})E_{qy} - \frac{1 - \hat{v}_{xy}G_y}{1 + \omega_{xy}\hat{z}_{xy}E_{qy}}.$$

Exponentiating and using the budget constraint $\sum_x a^{*}_{xy}q_y(x) = 1$ to eliminate $\lambda_y$ yields the robust closed-form screening strategy (2.2).

To derive the wage, we differentiate $Y_y$ with respect to $\omega_{xy}$. Using the full wage semi-elasticity $\hat{z}_{xy}$ and the identical quotient rule for queue weights, dividing by $\delta_y a^{*}_{xy}q_y(x)\hat{z}_{xy}E_{px}$

establishes the condition for the unconstrained interior wage:

$$\phi_{xy}(f_{xy} - \omega_{xy}) - \frac{\phi_{xy}}{\hat{z}_{xy}E_{px}} - \frac{\theta_y/\delta_y}{q_y(x)E_{px}}\left(q_y(x)(\ln q_y(x) + \lambda_y) - G_y\right) = 0.$$

To obtain the interpretable expression for the interior target wage, we solve the system of equations algebraically by substituting the unconstrained wage condition into the screening condition. Isolating and substituting $\frac{\theta_y}{\delta_y}(\ln q_y(x) + \lambda_y)$ from the combined FOCs yields:

$$\frac{\theta_y}{\delta_y}(\ln q_y(x) + \lambda_y) = \phi_{xy}f_{xy}E_{qy} - \frac{\theta_y}{\delta_y}.$$

We can use this interior equation to evaluate the queue-average cost $G_y$. Multiplying $G_y$ by $\frac{\theta_y}{\delta_y}$, substituting the identity for all $x'$ in the queue, and using the admissibility constraint $\sum_{x'} a^*_{x'y}q^*_y(x') = 1$, this evaluates to:

$$G_y = \frac{\delta_y}{\theta_y}\left(\sum_{x'=1}^{W} a^*_{x'y}q^*_y(x')\phi_{x'y}f_{x'y}E_{qy}\right) - 1 = \frac{\delta_y}{\theta_y}q^*_y(x)\phi_{xy}O_{xy}E_{qy} - 1,$$

where $O_{xy} \equiv \sum_{x'} a^*_{x'y}\frac{q^*_y(x')\phi_{x'y}}{q^*_y(x)\phi_{xy}}f_{x'y}$ is the firm's standard expected outside option.

Combining these simplified relations back into the wage FOC allows us to isolate $\omega_{xy}E_{px}$:

$$\omega_{xy}E_{px} = f_{xy}(E_{px} - E_{qy}) + O_{xy}E_{qy} + \left(1 - \frac{1}{q^*_y(x)}\right)\frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{\hat{z}_{xy}}.$$

Substituting the elasticity definitions $E_{px} - E_{qy} = \varepsilon_{m,c} - a_{xy}\varepsilon_{m,d}$ and $O_{xy}E_{qy} = a_{xy}f_{xy}\varepsilon_{m,d} + \varepsilon_{m,d}\sum_{x'\neq x} w^y_{x'|x}f_{x'y}$, the own-type terms $a_{xy}f_{xy}\varepsilon_{m,d}$ cancel out, yielding the internalized outside option $\hat{O}_{xy}$. Finally, applying the elasticity identity $\frac{1}{\hat{z}_{xy}} = \frac{1}{z^0_{xy}} - \omega_{xy}(E_{px} - 1)$ strips the remaining $E_{px}$ multiplier to yield the target wage:

$$\omega^*_{xy} = \hat{O}_{xy} + \left(1 - \frac{1}{q^*_y(x)}\right)\frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{z^0_{xy}}.$$

Applying the non-negativity and feasibility constraints enforces the bounds $[0, f_{xy}]$. □

The characterization in Proposition 2 reveals how firms balance direct profitability against physical congestion and informational externalities. The general screening

condition (2.2) takes a Multinomial Logit form where the term $\hat{v}_{xy}G_y$ captures the information cost spillover: attracting more applicants expands the queue, generating a marginal reduction in the firm's average entropy burden. Meanwhile, the target interior wage (2.3) anchors compensation to the worker's internalized replacement value ($\hat{O}_{xy}$), which combines the worker's direct technological contribution with the firm's cross-applicant threat point, adjusted for the worker's relative screening cost and a monopsony markdown. If a submarket's wage is strictly interior ($0 < \omega_{xy} < f_{xy}$), the wage condition holds with equality, and the informational spillovers perfectly offset the wage pass-through. In this specific case, the wage and relative cost terms cancel completely, and the optimal screening intensity collapses to a simple formulation $q_y^*(x) \propto \exp\left(\frac{\delta_y}{\theta_y}\phi_{xy}f_{xy}E_{qy}\right)$ driven entirely by the physical surplus scaled by the firm's matching elasticity. However, for boundary solutions where the wage condition binds as an inequality, this exact cancellation fails, and the robust general formulas (2.2) and (2.3) govern the equilibrium choices.

### Equilibrium

We can now define the equilibrium of the model. An equilibrium requires that agents optimize while rationally internalizing the physical congestion of the market, firms correctly anticipate worker responses, and the aggregate matching conditions are consistent with individual behaviors.

**Definition.** A *matching equilibrium* is a set of strategies for workers $\{p_x^*\}_{x=1}^W$, strategies for firms $\{q_y^*, \omega_{xy}^*\}_{y=1}^F$, and a set of congestion rates $\{\phi_{xy}^*\}$ such that:

1. **Worker Optimization:** For every type $x$, $p_x^*$ solves the worker's problem given the firms' posted contracts $\{q_y^*(x), \omega_{xy}^*\}$, fully internalizing how their own application intensity alters the submarket congestion rate $\phi_{xy}$ (via the elasticity $E_{px}$).

2. **Firm Optimization:** For every type $y$, $\{q_y^*, \omega_{xy}^*\}$ solves the firm's problem. Firms explicitly internalize both the worker's full best response function $p_x^*(\omega, q)$ and the endogenous congestion function $\phi_{xy}(p, q)$ (via the elasticities $E_{qy}$ and $E_{px}$).

3. **Consistency:** The congestion rates $\phi_{xy}^*$ and their local elasticities are consistent

with the aggregate strategies according to the matching technology:

$$\phi^*_{xy} = \frac{M\left(c^*_{xy}, d^*_{xy}\right)}{\mu_x \mu_y p^*_x(y) q^*_y(x)},$$

where $c^*_{xy} = \mu_x p^*_x(y) \frac{\mu_y}{\delta_x}$ and $d^*_{xy} = \mu_y q^*_y(x) a^*_{xy}$.

To operationalize the model and derive the properties of the wage equation, we specify the matching technology $M(\cdot, \cdot)$ as a Constant Elasticity of Substitution (CES) function:

$$M(c, d) = \left(\zeta c^\rho + (1 - \zeta)\, d^\rho\right)^{\frac{\sigma}{\rho}}, \quad \sigma > 0, \quad \rho \leq 1.$$

This specification nests the Cobb-Douglas case ($\rho \to 0$) and the Leontief case ($\rho \to -\infty$). Under this specification, the congestion function $\phi_{xy}$ is bounded and twice continuously differentiable, satisfying Assumption 3.

Finally, we address the existence and uniqueness of the equilibrium. While the worker's problem is strictly concave, the firm's problem involves the non-linear reaction function of the worker, which can introduce non-convexities in the firm's objective. However, sufficiently high information costs, or appropriate restrictions on the matching technology, convexify the problem.

**Proposition 3.** *Under Assumptions 1-4, for any strictly positive search costs ($\theta_x, \theta_y > 0$), a matching equilibrium exists. The uniqueness of this equilibrium depends on the properties of the matching technology and the magnitude of information frictions:*

1. ***General Uniqueness Threshold:*** *For any general matching function, there exists a threshold $\underline{\theta}$ such that for sufficiently high search costs $\left(\frac{\theta_x}{\delta_x}, \frac{\theta_y}{\delta_y}\right) > \underline{\theta}$, the matching equilibrium is unique.*

2. ***Global Uniqueness under CRS/DRS:*** *If the congestion elasticity internalized by the workers $E_{px} < 1$, for instance, when the matching technology $M(c, d)$ exhibits Constant Returns to Scale (CRS) or strictly decreasing returns, the matching equilibrium is globally unique for all $\theta_x, \theta_y > 0$.*

*Proof. Existence:* The existence of an equilibrium follows from Brouwer's Fixed Point Theorem (see also Tarski, 1955). The strategy space for each agent is the product of their search intensities and their wage menus. By Assumption 1, wages are bounded

$\omega_{xy} \in [0, f_{xy}]$. By the definition of $S_x$ and $S_y$ (simplexes), the search intensities $p$ and $q$ are bounded. The total strategy space is thus convex and compact. By Assumption 2 (strictly convex costs), the best-response functions $p_x^*(\cdot)$ and $q_y^*(\cdot)$ derived in Propositions 1 and 2 are continuous. By Assumption 3, the congestion function $\phi$ is continuously differentiable. Since the best-response mapping is a continuous function from a compact, convex set to itself, at least one fixed point exists for any $\theta > 0$.

*The Jacobian and Hessian Elements:* Uniqueness requires the Jacobian of the system of first-order conditions to be non-singular globally. Because individual firms are atomistic, they do not directly interact with one another's submarket congestion rates, rendering the aggregate Hessian block-diagonal. We must evaluate the definiteness of the Hessian block $H_{xy}$ for a single firm's choice of $\{q_y(x), \omega_{xy}\}$:

$$H = \begin{bmatrix} H_{qq} & H_{q\omega} \\ H_{q\omega} & H_{\omega\omega} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 Y_y}{\partial q_y^2} & \frac{\partial^2 Y_y}{\partial q_y \partial \omega_{xy}} \\ \frac{\partial^2 Y_y}{\partial \omega_{xy} \partial q_y} & \frac{\partial^2 Y_y}{\partial \omega_{xy}^2} \end{bmatrix}.$$

From the first-order condition for screening, $\frac{\partial Y_y}{\partial q_y(x)} = \hat{K}_{xy} \cdot \hat{B}_{xy} = 0$, where $\hat{K}_{xy} = \delta_y a_{xy}^* (1 + \omega_{xy} \hat{z}_{xy} E_{qy}) > 0$, and the interior root is characterized by $\hat{B}_{xy} = \phi_{xy} f_{xy} E_{qy} - \frac{\theta_y}{\delta_y}(\ln q_y(x) + 1 + \lambda_y)$. Evaluating the first row of the Hessian at the optimum ($\hat{B}_{xy} = 0$) yields:

$$H_{qq} = \hat{K}_{xy} \frac{\partial \hat{B}_{xy}}{\partial q_y(x)} = -\theta_y a_{xy}^* \frac{1 + \omega_{xy} \hat{z}_{xy} E_{qy}}{q_y(x)}$$

$$H_{q\omega} = \hat{K}_{xy} \frac{\partial \hat{B}_{xy}}{\partial \omega_{xy}} = \delta_y a_{xy}^* \left(1 + \omega_{xy} \hat{z}_{xy} E_{qy}\right) \frac{\partial (\phi_{xy} E_{qy})}{\partial \omega_{xy}} f_{xy}$$

The second derivative with respect to the wage, $H_{\omega\omega}$, captures the curvature of the firm's revenue. Differentiating the marginal profit condition $\frac{\partial Y_y}{\partial \omega_{xy}} = 0$ yields a leading term structurally proportional to the derivative of the effective application elasticity $\hat{z}_{xy}$:

$$H_{\omega\omega} \approx \delta_y a_{xy}^* q_y(x) \phi_{xy} E_{px} \left[ \frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}} (f_{xy} - \omega_{xy}) - 2\hat{z}_{xy} \right].$$

To ensure a globally unique maximum, the Hessian must be negative definite, which requires $H_{qq} < 0$, $H_{\omega\omega} < 0$, and a positive determinant: $\text{Det}(H) = H_{qq} H_{\omega\omega} - (H_{q\omega})^2 > 0$. The strict convexity of the screening cost ensures $H_{qq}$ is unconditionally negative. However, the sign of $H_{\omega\omega}$ depends on $\frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}}$. Because applications follow a Logistic

distribution, the unadjusted elasticity $z^0_{xy}$ grows with the wage, which can generate a sufficiently large, positive $\frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}}$, causing regions of convexity ($H_{\omega\omega} > 0$).

*General Uniqueness Threshold ($\underline{\theta}$).* We analyze the matrix as search costs become large, assuming symmetric scaling $\theta_x \approx \theta_y = \theta \to \infty$. The unadjusted semi-elasticity of applications scales inversely with the worker's search cost: $z^0_{xy} \propto 1/\theta$. Consequently, the effective semi-elasticity $\hat{z}_{xy}$ and its cross-derivatives are dominated by this inverse relationship. Substituting these asymptotic orders into the Hessian elements yields:

1. $H_{qq} \sim \mathcal{O}(\theta)$, dominated by the strict convexity of the firm's entropy cost.
2. $H_{\omega\omega} \sim \mathcal{O}(1/\theta)$, representing the dampened logit revenue concavity.
3. $H_{q\omega} \sim \mathcal{O}(1/\theta)$, since changes in congestion are proportional to $\hat{z}_{xy}$.

Multiplying the asymptotic orders, the product of the diagonals is $H_{qq}H_{\omega\omega} \sim \mathcal{O}(\theta) \times \mathcal{O}(1/\theta) = \mathcal{O}(1)$, while the squared off-diagonal is $(H_{q\omega})^2 \sim \mathcal{O}(1/\theta^2)$. For sufficiently large $\theta$, the $\mathcal{O}(1)$ diagonal product strictly dominates the $\mathcal{O}(1/\theta^2)$ cross-derivative term. By continuity, there exists a threshold $\underline{\theta}$ above which the strict convexity of the screening cost overpowers any local revenue non-convexities, ensuring the Hessian is globally negative definite.

*Global Uniqueness under CRS/DRS:* The global concavity of the revenue function ($H_{\omega\omega} < 0$) relies on the structural dampening of the application elasticity. Differentiating the marginal profit condition yields the exact structural form for the revenue curvature:

$$H_{\omega\omega} \propto \hat{z}_{xy} \left\{ \hat{z}_{xy}(f_{xy} - \omega_{xy}) \left[ 2E_{px} - 1 + \frac{(E_{px} - 1) - \frac{\tilde{p}_x}{1-\tilde{p}_x}}{1 - z^0_{xy}\omega_{xy}(E_{px} - 1)} \right] - 2 \right\},$$

where $\tilde{p}_x = \frac{\mu_y}{\delta_x} p_x(y)$ is the absolute application probability. In a general model when $E_{px} \geq 1$, e.g. when agents take congestion as given ($E_{px} = 1$), the bracketed term reduces to $\left[1 - \frac{\tilde{p}_x}{1-\tilde{p}_x}\right]$, which is strictly positive for small $\tilde{p}_x$. Because the unadjusted elasticity $z^0_{xy}$ explodes to infinity as $\theta \to 0$, this positive $\hat{z}^2_{xy}$ term quadratically overpowers the negative $-2\hat{z}_{xy}$ term, causing severe non-convexities and allowing for multiple equilibria.

To guarantee global uniqueness, the physical matching technology must actively penalize congestion, which requires $(E_{px} - 1) < 0$. By definition, $(E_{px} - 1) = \varepsilon_{m,c} +$

$(1 - a_{xy})\varepsilon_{m,d} - 1$. For example, this term is strictly negative under two conditions:

1. **Strictly Decreasing Returns to Scale (DRS):** If the matching function exhibits DRS such that $\varepsilon_{m,c} + \varepsilon_{m,d} = \nu < 1$, then $(E_{px} - 1) \leq \nu - 1 < 0$ regardless of the queue composition.

2. **Constant Returns to Scale (CRS) with Firm Capacity Constraints:** If the matching function exhibits CRS ($\varepsilon_{m,c} + \varepsilon_{m,d} = 1$), then $(E_{px} - 1) = -a_{xy}\varepsilon_{m,d}$. For this to be strictly negative, the firm's screening input must impose a binding physical constraint on match formation ($\varepsilon_{m,d} > 0$). If $\varepsilon_{m,d} = 0$ (e.g., $M(c, d) = c$), firms have infinite physical capacity, the congestion brake fails, and the model can produce Bertrand-like competition in the limit.

When the matching technology ensures $(E_{px} - 1) < 0$, the denominator of the full elasticity, $1 - z^0_{xy}\omega_{xy}(E_{px} - 1)$, is strictly greater than one and grows monotonically with the unadjusted elasticity $z^0_{xy}$. This congestion feedback provides a speed limit on the labor supply elasticity: as $\theta \to 0$, the elasticity $\hat{z}_{xy}$ converges to a finite bound $\frac{-1}{\omega_{xy}(E_{px}-1)}$.

Because $\hat{z}_{xy}$ remains finite, the potentially positive $\hat{z}^2_{xy}$ factor is capped. Furthermore, the strictly negative $(E_{px} - 1)$ term injects a negative anchor into both the numerator of the fraction and the leading $2E_{px} - 1$ term. The negative components overpower the bounded quadratic term, guaranteeing that the expression inside the braces remains strictly negative for any valid wage. Consequently, the firm faces uniformly diminishing marginal returns to wage increases ($H_{\omega\omega} < 0$). Coupled with the strictly convex entropy cost ($H_{qq} < 0$), the determinant of the Hessian is strictly positive. The objective function maps to a globally strictly concave maximization problem on a convex compact set, ensuring the equilibrium is globally unique for all $\theta_x, \theta_y > 0$.

*Boundary Solutions:* This uniqueness property extends to boundary solutions. If the optimal wage for a given submarket binds at a corner ($\omega^*_{xy} \in \{0, f_{xy}\}$), the wage ceases to be an unconstrained choice variable. The local Hessian matrix for that submarket drops the wage dimension and collapses to the scalar $H_{qq}$. Because the entropy cost of screening ensures $H_{qq} < 0$ globally, the firm's objective is unconditionally strictly concave with respect to the remaining screening intensity choice, further reinforcing the uniqueness of the global mapping. □

**Social Planner's Solution**

We solve the social planner's problem for the sequential model assuming a utilitarian welfare function. The planner internalizes all benefits and costs of the matching process, subject to the same informational constraints on search and screening intensities as individual agents. Because wages are pure transfers between agents, they cancel out of the aggregate welfare function. The planner chooses the optimal allocations of search effort ($p$) and screening intensity ($q$) to maximize:

$$\Omega = \sum_{x=1}^{W} \mu_x \theta_x \lambda_x + \sum_{y=1}^{F} \mu_y \theta_y \lambda_y$$
$$+ \sum_{y=1}^{F} \sum_{x=1}^{W} \mu_x \mu_y p_x(y) \left( q_y(x) \phi_{xy} f_{xy} - \tfrac{\theta_x}{\delta_x} \left( \ln p_x(y) + \lambda_x \right) - \tfrac{\theta_y}{\delta_y} q_y(x) \left( \ln q_y(x) + \lambda_y \right) \right).$$

We begin by taking the first-order condition with respect to the screening intensity $q_y(x)$. Let $\varepsilon_{m,d} = 1 + \varepsilon_{\phi,q}$ denote the elasticity of the matching function with respect to the mass of screened workers. Because screening is targeted from within the queue, changing $q_y(x)$ only affects the submarket for type $x$. Differentiating the welfare function yields:

$$\frac{\partial \Omega}{\partial q_y(x)} = \mu_x \mu_y p_x(y) \left[ \phi_{xy} f_{xy} \varepsilon_{m,d} - \frac{\theta_y}{\delta_y} \left( \ln q_y(x) + 1 + \lambda_y \right) \right] = 0.$$

Rearranging this equation provides the planner's optimal screening condition:

$$\frac{\theta_y / \delta_y}{\phi_{xy}} \left( \ln q_y(x) + 1 + \lambda_y \right) = f_{xy} \varepsilon_{m,d}. \tag{2.4}$$

Notice that the planner equates the marginal entropy cost of screening to the social return of a match, which is the total surplus scaled by the firm's technological contribution ($\varepsilon_{m,d}$).

Next, we take the first-order condition with respect to the worker's application strategy $p_x(y)$. This is significantly more complex because an application increases the aggregate queue length ($\delta_y = \sum_{x'} \mu_{x'} p_{x'}(y)$). By the quotient rule, increasing $p_x(y)$ reduces the effective queue weight $a_{x'y}$ for all other applicants, generating two cross-submarket externalities:

1. **Congestion/Cannibalization:** It physically dilutes the firm's screening capacity ($d_{x'y}$) for all other applicant types $x'$, reducing their matching rates with

elasticity $-a_{xy}\varepsilon_{m,d}$.

2. **Information Spillovers:** It lowers the firm's average entropy cost of screening by providing a larger base over which to spread information processing.

To formalize the congestion externality, we define the planner's expected outside option of the queue, $O_{xy}^{PO}$. Using the relative matching weights $w_{x'|x}^{y} \equiv a_{x'y}\frac{q_y(x')\phi_{x'y}}{q_y(x)\phi_{xy}}$ defined in Proposition 2, the planner evaluates the outside option at the screening elasticity share:

$$O_{xy}^{PO} \equiv \sum_{x'=1}^{W} w_{x'|x}^{y} \left(f_{x'y}\varepsilon_{m,d}\right).$$

Taking the derivative of the total generated surplus with respect to $p_x(y)$ yields the Marginal Social Benefit (MSB). It consists of the direct submarket contribution ($\varepsilon_{m,c} + \varepsilon_{m,d}$) minus the cannibalization of the rest of the queue:

$$\text{MSB}_{p_x(y)} = \mu_x\mu_y q_y(x)\phi_{xy}\left[f_{xy}(\varepsilon_{m,c} + \varepsilon_{m,d}) - O_{xy}^{PO}\right].$$

Taking the derivative of the aggregate entropy costs with respect to $p_x(y)$ yields the Marginal Social Cost (MSC). It includes the worker's direct search cost and the firm's cross-submarket information spillover:

$$\text{MSC}_{p_x(y)} = \mu_x\mu_y\frac{\theta_x}{\delta_x}\left(\ln p_x(y) + 1 + \lambda_x\right) + \mu_x\mu_y\frac{\theta_y}{\delta_y}q_y(x)\left[\ln q_y(x) + \lambda_y - \sum_{x'=1}^{W} w_{x'|x}^{y}\frac{\phi_{xy}}{\phi_{x'y}}\left(\ln q_y(x') + \lambda_y\right)\right].$$

Equating MSB and MSC seems to imply that the planner's allocation depends heavily on the composition of the entire queue. However, we can substitute the planner's screening FOC (2.4) directly into the firm's information spillover term. Because $\frac{\theta_y}{\delta_y}(\ln q_y(x') + \lambda_y) = \phi_{x'y}f_{x'y}\varepsilon_{m,d} - \frac{\theta_y}{\delta_y}$, the firm's cost derivative perfectly absorbs the cannibalization term:

$$\text{Firm's Cost Derivative} = \mu_x\mu_y q_y(x)\phi_{xy}\left[f_{xy}\varepsilon_{m,d} - O_{xy}^{PO}\right] - \mu_x\mu_y q_y(x)\frac{\theta_y}{\delta_y} + \mu_x\mu_y\frac{\theta_y}{\delta_y}.$$

When we set $\text{MSB}_{p_x(y)} = \text{MSC}_{p_x(y)}$, the terms $f_{xy}\varepsilon_{m,d}$ and the entire queue outside option $O_{xy}^{PO}$ *perfectly cancel out* on both sides. The negative social cost of diluting the queue is exactly offset by the positive social benefit of lowering the firm's average

entropy cost. Dividing the remaining terms by $\mu_x \mu_y q_y(x) \phi_{xy}$ isolates the planner's optimal application condition:

$$\frac{\theta_x/\delta_x}{q_y(x)\phi_{xy}} \left(\ln p_x(y) + 1 + \lambda_x\right) = f_{xy}\varepsilon_{m,c} + \left(1 - \frac{1}{q_y(x)}\right) \frac{\theta_y/\delta_y}{\phi_{xy}}. \tag{2.5}$$

To benchmark the efficiency of the competitive equilibrium, we derive the shadow prices that would decentralize the social planner's allocation. We do this by equating the marginal private returns in the competitive equilibrium to the marginal social returns in (2.5).

Recall from Proposition 1 that the worker's private first-order condition for applications in the competitive market is:

$$\frac{\theta_x/\delta_x}{q_y(x)\phi_{xy}} \left(\ln p_x(y) + 1 + \lambda_x\right) = \omega_{xy}.$$

Because the left-hand side of the worker's private FOC is identical to the left-hand side of the planner's condition (2.5), the right-hand side of (2.5) exactly defines the shadow wage $\omega_{xy}^{PO,W}$ required to induce socially optimal applications. The firm's residual shadow profit is simply the remaining surplus: $\pi_{xy}^{PO,F} = f_{xy} - \omega_{xy}^{PO,W}$.

**Proposition 4.** *Under Assumptions 1-4, the shadow wages and residual profits required to induce the socially optimal search and screening strategies are given by:*

$$\omega_{xy}^{PO,W} = \underbrace{f_{xy}\varepsilon_{m,c}}_{\textit{Hosios Share}} + \underbrace{\left(1 - \frac{1}{q_y^{PO}(x)}\right) \frac{\theta_y/\delta_y}{\phi_{xy}}}_{\textit{Screening Cost Premium}}.$$

$$\pi_{xy}^{PO,F} = \underbrace{f_{xy}\varepsilon_{m,d}}_{\textit{Hosios Share}} - \underbrace{\left(1 - \frac{1}{q_y^{PO}(x)}\right) \frac{\theta_y/\delta_y}{\phi_{xy}}}_{\textit{Screening Cost Premium}}.$$

The standard efficiency requirement is that an agent receives a share of the physical surplus equal to the elasticity of the matching function with respect to their input ($\varepsilon_{m,\cdot}$), mirroring the classic Hosios condition. The second term in both equations is a zero-sum transfer that acts as a screening cost premium. Because $q_y(x)$ is an intensity that averages to 1 across the queue, a heavily targeted worker ($q_y(x) > 1$) is systematically

"cheaper" for the firm to identify relative to the baseline. The planner passes this information-cost savings onto the worker via a higher shadow wage.

Assuming the matching technology exhibits constant returns to scale ($\varepsilon_{m,c} + \varepsilon_{m,d} = 1$), summing the planner's shadow prices perfectly exhausts the physical surplus ($f_{xy}$).

**Proposition 5.** *Under Assumptions 1-4, for finite positive costs of search, the competitive equilibrium is generically constrained inefficient.*

*Proof.* Comparing the social planner's shadow wage with the competitive equilibrium wage derived under internalized congestion reveals a much closer alignment in pricing, though a fundamental divergence remains for finite search costs. Substituting the interior screening condition into the wage equation, the competitive wage can be expressed as:

$$\omega_{xy}^{CE} = \underbrace{f_{xy}\varepsilon_{m,c}}_{\text{Direct Hosios Share}} + \underbrace{\varepsilon_{m,d} \sum_{x' \neq x} w_{x'|x}^{y} f_{x'y}}_{\text{Cross-Applicant Outside Option}} + \left(1 - \frac{1}{q_y^*(x)}\right) \frac{\theta_y/\delta_y}{\phi_{xy}} - \underbrace{\frac{1}{z_{xy}^0}}_{\text{Monopsony Markdown}}.$$

Because the firm internalizes physical congestion, it correctly anchors the base of the worker's compensation to their direct technological contribution ($f_{xy}\varepsilon_{m,c}$), perfectly aligning with the social planner. However, the competitive equilibrium still deviates from the planner's allocation through two distinct mechanisms. First, the private market allows the firm to leverage the heterogeneous queue composition against the worker. The firm uses the expected surplus of *other* applicant types as a bargaining threat point, introducing the cross-applicant outside option. The social planner, who values total surplus creation, does not penalize workers for the presence of others. Second, the competitive wage is distorted by the monopsony markdown $1/z_{xy}^0$. Because workers face information frictions, their search is noisy, and firms strategically depress wages to extract rents.

For the competitive market to achieve the socially efficient allocation, the cross-applicant outside option must exactly offset the monopsony markdown:

$$\varepsilon_{m,d} \sum_{x' \neq x} w_{x'|x}^{y} f_{x'y} = \frac{1}{z_{xy}^0}.$$

Because the left side is determined by endogenous strategic queue composition and the

right side is driven by the exogenous information cost $\theta_x$, this equality generically fails to hold for any finite $\theta_x > 0$. Thus, the equilibrium is generically constrained inefficient.

However, this formulation nests the efficiency conditions of the canonical directed search literature in the frictionless limit. If information costs vanish ($\theta_x \to 0$), the simple elasticity explodes ($z^0_{xy} \to \infty$) and the monopsony markdown approaches zero. If the market simultaneously perfectly sorts such that submarkets become homogeneous ($a_{xy} \to 1$), the cross-applicant threat point becomes an empty sum and vanishes. In this frictionless, fully sorted limit, the distortions disappear, and the competitive market delivers the efficient Hosios share ($\omega^{CE}_{xy} \to f_{xy}\varepsilon_{m,c}$). □

For any strictly positive information costs, however, the $1/z^0_{xy}$ term serves as a "monopsony tax." The firm's strategic restriction of wages to extract this rent represents a private benefit but dissipates social surplus. However, it is important to note that not all wage distortions are allocatively relevant; some are merely transfers that do not lead to misallocation. In our environment, where workers face a constraint that application intensities must average to one, they cannot adjust their total search supply; they can only substitute search effort between firms. Therefore, allocative inefficiency arises precisely when the market distorts the *relative* payoffs across different firms. While the absolute magnitude of the monopsony markdown determines the aggregate distribution of surplus, welfare loss is ultimately driven by the variance of the net wage distortions ($\varepsilon_{m,d} \sum_{x' \neq x} w^y_{x'|x} f_{x'y} - 1/z^0_{xy}$) across potential matches. It is this relative variance that distorts the price signals guiding worker sorting.

# 3 Mechanics of Equilibrium Wages and Sorting

In this section, we analyze the properties of the matching equilibrium. We decompose the closed-form wage equation to isolate the structural determinants of monopsony power. Finally, we demonstrate how the interaction of these forces generates distinct regimes of sorting. A key theoretical finding is that strong assortative matching can be sustained even when wage signals are compressed, provided that sorting is driven by firm screening rather than worker self-selection.

## The Anatomy of the Wage

To understand wage determination along the continuum between random and directed search, we examine the closed-form solution for the equilibrium wage (see Equation 2.3). For a symmetric Cobb-Douglas specification of the matching technology the wage can be decomposed into three structural components:

$$\omega_{xy}^{*} = \left[ \underbrace{\frac{1}{2} f_{xy}}_{\text{Hosios Share}} + \underbrace{\frac{1}{2} \sum_{x' \neq x} w_{x'|x}^{y} f_{x'y}}_{\text{Cross-Applicant OO}} - \underbrace{\frac{\theta_x}{\sqrt{q_y(x)\frac{\delta_x}{\delta_y}}} \frac{1/E_{px}}{\left(1 - \frac{\mu_y}{\delta_x} p_x(y)\right)}}_{\text{Monopsony Discount, } 1/z_{xy}^{0}} + \underbrace{\left(1 - \frac{1}{q_y(x)}\right) \frac{\theta_y}{\sqrt{\frac{\delta_y}{q_y(x)\delta_x}}}}_{\text{Screening Cost Premium}} \right]_{0}^{f_{xy}}$$

**Hosios Share and Cross-Applicant Outside Option.** The base components of the wage separate the worker's direct technological contribution from the firm's bargaining threat. Because the firm internalizes physical congestion, it anchors the wage to the worker's Hosios Share ($\frac{1}{2} f_{xy}$ under the symmetric Cobb-Douglas matching function where $\varepsilon_{m,c} = \varepsilon_{m,d} = \frac{1}{2}$). Rather than evaluating the worker against the full expected surplus of the queue, the firm isolates this direct contribution from the cross-applicant outside option. This alternative threat point is defined as a weighted average of the surpluses of all *other* potential applicants:

$$\text{Cross-Applicant Option} \equiv 0.5 \sum_{x' \neq x} \underbrace{a_{x'y}^{*} \frac{q_y^{*}(x')\,\phi_{x'y}}{q_y^{*}(x)\,\phi_{xy}}}_{w_{x'|x}^{y}} f_{x'y}.$$

The weights $w_{x'|x}^{y}$ represent the relative likelihood that the firm forms an alternative match with a worker of type $x'$ instead of type $x$. They are formed by taking the empirical frequency of type $x'$ in the queue $a_{x'y}^{*}$ and adjusting it by the relative matching rate $\frac{q_y^{*}(x')\phi_{x'y}}{q_y^{*}(x)\phi_{xy}}$. This adjustment is economically important: the firm's threat point during wage setting is not merely the average applicant, but the expected alternative conditional on the firm's own screening behavior $q$ and physical submarket congestion $\phi$. This base is determined by the endogenous sorting pattern. In highly assortative equilibria where workers target firms precisely, the queue becomes homogeneous ($a_{xy}^{*} \to$

1). Consequently, the cross-applicant sum is empty, and the outside option vanishes entirely, leaving the base wage at the pure Hosios share. However, as we show next, this does not imply the worker retains this full share, as it is offset by the monopsony discount.

**Monopsony Discount.** The third term represents the firm's ability to mark down wages due to an upward-sloping labor supply curve. It is driven by the worker's search cost $\theta_x$ (the inverse semi-elasticity of labor supply) and the precision of targeting. The denominator of this markdown reveals two distinct mechanisms through which sorting suppresses wages: perceived self-congestion and a lack of viable outside options.

First, under the symmetric Cobb-Douglas matching function ($\varepsilon_{m,c} = \varepsilon_{m,d} = 0.5$), the congestion multiplier simplifies to $E_{px} = 1 - a_{xy}/2$. This term captures the worker's perceived self-congestion. When a worker's type constitutes a small fraction of the queue ($a_{xy} \to 0$), an additional application mostly crowds out other applicant types, yielding a high marginal return to search ($E_{px} \approx 1$). However, as the market sorts and the queue becomes homogeneous ($a_{xy} \to 1$), workers increasingly crowd out their own peers, causing the marginal return to drop to 0.5. Because this self-congestion term sits in the denominator, strong sorting amplifies the monopsony markdown. When workers know that applying more heavily will largely result in self-cannibalization, their application strategy becomes highly inelastic to wage increases.

Second, the term $(1 - \frac{\mu_y}{\delta_x} p_x(y))$ ensures that this discount can persist even as search costs vanish ($\theta_x \to 0$) in assortative equilibria. When workers perfectly target a single firm type, the probability of applying to a second-best firm approaches zero. Because this deviation probability shrinks exponentially fast under the entropic cost, the overall application elasticity collapses. Economically, this means that in a highly assortative equilibrium, firms realize they have a "captive audience." With no effective wage competition because workers are cornered into a specific match, the firm can extract a large share of the surplus via the markdown, compounding the penalty of the self-congestion effect.

The first two and the third term of our wage equation reveal a direct structural parallel to the canonical theory of monopsonistic competition (e.g., Manning, 2003). In standard oligopsony models, a profit-maximizing firm facing an upward-sloping labor

supply curve $L_y(\omega)$ sets the wage according to the inverse elasticity rule:

$$\omega = MRP - \frac{L_y(\omega)}{L_y'(\omega)} = MRP - \frac{1}{\partial \ln L_y(\omega)/\partial \omega}.$$

Our wage equation closely recovers this structure while adapting it to a frictional matching environment. The combination of the direct Hosios share and the cross-applicant outside option ($f_{xy}\varepsilon_{m,c} + \varepsilon_{m,d}\sum_{x' \neq x} w^y_{x'|x} f_{x'y}$) represents the effective baseline value (the marginal revenue product net of capacity constraints) facing the firm. The term $1/z^0_{xy}$ represents the inverse semi-elasticity of the firm's labor supply. However, while standard oligopsony models typically assume this elasticity arises from exogenous preference heterogeneity (e.g., Type I Extreme Value shocks in a Logit model), our framework provides a microfoundation for it. The elasticity is endogenously determined by the information friction $\theta_x$ and the strategic queue composition. As search costs rise, worker targeting becomes less precise, labor supply becomes less elastic, and the monopsony markdown increases. Thus, our model nests the "preference-based" monopsony mechanism within an information-theoretic framework.

**Screening Cost Premium.** The third term is a wage adjustment reflecting the firm's relative informational burden. It is proportional to the firm's screening cost $\theta_y$. Because $q_y(x)$ is an intensity that averages to 1 across the queue, a heavily targeted worker ($q_y(x) > 1$) is systematically "cheaper" for the firm to identify relative to the baseline applicant. The firm passes a portion of these information-cost savings to the worker via a wage premium. Conversely, if a worker is avoided ($q_y(x) < 1$), the wage is penalized. However, in assortative equilibria where the queue is perfectly homogeneous, the firm does not need to selectively screen ($q_y(x) \to 1$) and this premium disappears.

The equilibrium wage is the outcome of a tension between the monopsony discount (pulling wages down) and the screening cost premium (pushing wages up). The balance is determined by relative search costs ($\theta_x$ vs $\theta_y$), labor market tightness (relative scarcity of workers vs. firms $\delta_x/\delta_y$), and the sorting pattern. Strong assortative matching tends to eliminate the screening premium while maximizing the monopsony discount, creating a powerful force for wage compression.

The interaction of these forces reveals that sorting can be generated by two distinct mechanisms with opposite implications for wage inequality. This distinction is central

to resolving the empirical puzzle of "flat wage profiles" in sorted markets.

**Competitive (Worker-Driven) Sorting.** In environments with low information costs or weak monopsony power (e.g., mixing equilibria), sorting is driven by the worker's application strategy $p_x$. To induce high-productivity workers to target high-productivity firms, the market must generate steep wage gradients. High wages serve as the necessary signal to overcome search costs and direct the worker's attention. Consequently, in this regime, strong sorting (a high correlation between worker and firm types) is always accompanied by high wage dispersion and low markdowns. This aligns with the intuition of standard directed search models, where wages play the primary allocative role.

**Monopsony (Firm-Driven) Sorting.** A different regime emerges when monopsony power is strong or workers are effectively captive. Here, the monopsony discount dominates, driving wages down toward the worker's outside option. This flattens the wage profile across firms, weakening the price signal. Under these conditions, workers have little incentive to target specific firms, and their application strategies $p_x$ become diffuse (random). However, sorting does not disappear. Instead, it is enforced by the firm's screening strategy $q_y$. If productive complementarities are strong, firms find it profitable to exert effort to "cherry-pick" the best matches from their random applicant pools. This regime can generate strong positive assortative matching coexisting with flat wage profiles and high markdowns. This decoupling of sorting from wage signals provides a theoretical foundation for the empirical patterns documented in the recent literature (e.g., Bonhomme et al., 2019), suggesting that observed sorting may be driven by firm selection rather than worker self-selection.

# 4 Resolving the Monopsony Puzzles

Labor market data presents a constellation of stylized facts that are difficult to reconcile within a single theoretical framework. Any successful structural model of the labor market must be able to jointly rationalize three robust empirical regularities that appear, on the surface, to be contradictory.

First, a large body of empirical work, culminating in the dynamic estimation of Bonhomme, Lamadon, and Manresa (2019) (BLM), documents strong positive assortative matching. High-skill workers are disproportionately employed by, and transition to, high-wage firms. This suggests powerful complementarities in production: the match surplus generated by pairing high-skill workers with high-productivity firms must be significantly larger than that of random assignment. In canonical models, such sorting is sustained by price signals: high-productivity firms must post high wages to attract high-skill applicants.

However, the same data reveals a second fact that conflicts with this intuition: wage profiles are surprisingly flat with respect to firm productivity. While wages do increase as workers move to better firms, the gradient is shallow. The "pass-through" of firm productivity into wages is low, meaning that workers capture only a small fraction of the additional surplus generated by moving up the job ladder. This creates a tension with the first fact. If the wage premium for moving to a top firm is small, why do high-skill workers sort there so effectively? Standard search models typically require steep wage gradients to incentivize the costly search effort needed to generate strong sorting. The coexistence of strong sorting and weak price signals is a central puzzle.

Finally, there is a conflict regarding the source of wage variation. The literature following Abowd, Kramarz, and Margolis (1999) consistently finds that firm fixed effects explain a relatively small share of the variance in wages, with the majority explained by worker fixed effects. This suggests that the firm's identity is not a primary driver of wage levels. Yet, a parallel literature on monopsony and concentration finds evidence of substantial markdowns, implying that firms possess significant market power. The puzzle is how firms can simultaneously exercise large monopsony power, setting wages far below marginal product, while generating wage data that appears to be dominated by worker characteristics rather than firm identity.

Standard frameworks struggle to resolve these tensions simultaneously. Directed search models typically generate efficient sorting but predict competitive pricing where wages align with marginal products, failing to explain markdowns or flat wage profiles. Conversely, random search models naturally generate monopsony power and can produce flat wages, but they struggle to generate strong two-sided sorting without assuming it directly through the production technology. Similarly, bargaining models can support sorting if the surplus is supermodular, but they generally imply that workers

receive a constant share of the surplus, linking wages mechanically to productivity and contradicting the observation of low pass-through.

We aim to show that our information-based model resolves these tensions by identifying a new regime of the labor market: one where sorting is driven by firm screening rather than worker self-selection.

## Data and Estimation Strategy

To investigate these puzzles quantitatively, we fit our model to moments from the Swedish labor market, using the dynamic structural estimates of Bonhomme, Lamadon, and Manresa (2019) (hereafter BLM) as our data generating process.

BLM characterize the labor market with 6 worker types and 10 firm types. However, because their model focuses on job-to-job transitions (excluding unemployment), a worker's state is defined not just by their permanent type but also by their current employer. To map this dynamic environment into our framework, we expand the state space by defining a "searcher type" $s = (x, y_{prev})$ as a worker of permanent type $x \in \{1, \ldots, 6\}$ currently employed at a firm of type $y_{prev} \in \{1, \ldots, 10\}$. This yields 60 distinct worker types searching for jobs at 10 distinct firm types. The inputs to our model: the masses of searching workers $\mu_s$ and firms $\mu_y$, are calibrated directly from the steady-state distribution of workers across firms reported by BLM.

We estimate the model via Generalized Method of Moments (GMM), targeting 722 moments derived from the BLM estimates.[6] These moments capture the full structure of worker flows and wages. Specifically, we target the transition probabilities at which each of the 60 worker types moves to each of the 10 destination firm types (600 moments), capturing the detailed sorting patterns. We also target the aggregate number of new matches formed between each permanent worker type and firm type (60 moments). Crucially, we match the estimated mean wage profiles for each firm-worker combination (60 moments), which captures the gradient of wages with respect to productivity. Finally, to ensure our model replicates the aggregate structure of wage inequality, we target two key moments from the AKM analysis of the same data: the share of wage

[6]For the current GMM estimation, we utilize an earlier model specification where firms act myopically by taking physical congestion as given and ignoring cross-applicant information spillovers, yielding slightly different equilibrium wage dynamics that are detailed in the Appendix. We are still working on estimation of the more general model.

variance explained by firm fixed effects (4.2%) and the correlation between worker and firm fixed effects (41%).

We estimate the deep parameters of the model to match these moments. An important extension for this estimation is that workers differ in their outside option based on their current employment. We denote the outside option of a worker of searcher type $s$ as $V_s$. We assume that each match between a permanent worker type $x$ and a firm type $y$ produces a productivity $f_{xy}$, which does not depend on the worker's previous employment. This determines the surplus of a match, $f_{sy} - V_s$, which is itself a $60 \times 10$ matrix derived from the $6 \times 10$ productivity parameters and 60 outside option parameters. We do not restrict the productivities or the outside options in our estimation.

In addition, we assume a flexible parameter structure for the type-pair specific matching technology. We specify the matching functions as Cobb-Douglas with Constant Returns to Scale (CRS), $M(c, d) = \psi_{xy} c^{\alpha_{xy}} d^{1-\alpha_{xy}}$, and estimate the matrices of matching efficiencies $\psi_{xy}$ and elasticities $\alpha_{xy}$. Finally, we estimate the vectors of search costs. Identification tests suggested that type-specific costs for workers were not separately identified, so we restrict the worker search cost to a single scalar $\theta_w$, while allowing firm screening costs $\theta_y$ to vary by firm type.

This parameterization results in 251 estimated parameters to explain 722 target moments. Given that the system is massively over-identified, the model cannot fit every moment perfectly. We therefore prioritize the moments we deem most important for the economic mechanism: the 60 wage moments, the 60 transition moments, and the 2 AKM moments, placing lower weight in the GMM objective function on the remaining fine-grained flow moments.

## Estimation Results

The estimated model achieves a remarkable fit to the data, simultaneously matching the detailed wage profiles, the complex matrix of worker transitions, and the aggregate AKM moments. Figure 4.1 illustrates this fit. The left panel plots the mean log-wages for each of the 6 worker types across the 10 firm types. The model (dashed lines) tracks the data (solid lines) almost perfectly. Notably, while wages generally rise with firm type, the profiles are relatively flat compared to the estimated productivities (dotted lines), indicating substantial wage compression. The right panels illustrate

the sorting patterns in stocks (top) and flows (bottom). High-wage workers (Types 5-6) are disproportionately employed by and move to high-wage firms (Types 8-10), demonstrating strong positive assortative matching. The model replicates these flow compositions exactly. This combination (flat wage profiles alongside strong sorting) is the central empirical puzzle. The model's ability to fit these moments simultaneously, along with the high AKM sorting correlation (41%) and low firm fixed effects (4.2%), suggests it successfully captures the underlying structure of the Swedish labor market. We now turn to the estimated parameters to understand the mechanism driving these results.

Figure 4.2 displays our estimated structural parameters. There is significant heterogeneity in productivity, search costs, matching efficiency, and elasticities. Two results stand out. First, the outside options $V_s$, although completely unrestricted in the estimation, are estimated to be nearly equal to the current wages of the workers, with only a few exceptions (as illustrated in the bottom-right panel of Figure 4.2). This implies that firms yield huge monopsony power and that markdowns are large. Firms capture most of the surplus generated by the match. This result explains the AKM puzzle. Wages are determined primarily by the worker's history (their outside option $V_s$) rather than the productivity of the new match. This explains why worker fixed effects are high (wages track the worker's history) while firm fixed effects are small: a high-productivity firm does not pay significantly more than a low-productivity firm; it simply captures more profit.

Second, the most productive firm-worker combinations are not the ones that pay the highest wages. The strong wage-productivity disconnect is shown separately in Figure 4.3. The correlation between estimated productivity $f_{xy}$ and wages $w_{xy}$ is close to zero. This confirms the theoretical prediction of our firm-driven sorting regime. Because high-productivity firms face inelastic labor supply (or effectively capture the queue), they exert stronger monopsony power, flattening the wage gradient. Consequently, wages cease to be a reliable signal of productivity. This finding provides structural support for the hypothesis of Eeckhout and Kircher (2011) that wage data alone may be insufficient, even misleading, for identifying sorting patterns in the presence of frictions.

Our estimation results resolve the apparent contradiction between strong sorting and flat wage profiles. The resolution lies in the distinct roles played by workers and firms, driven by the distribution of the match surplus. Figure 4.4 decomposes the value

Figure 4.1: Matched BLM Moments and Model Fit.

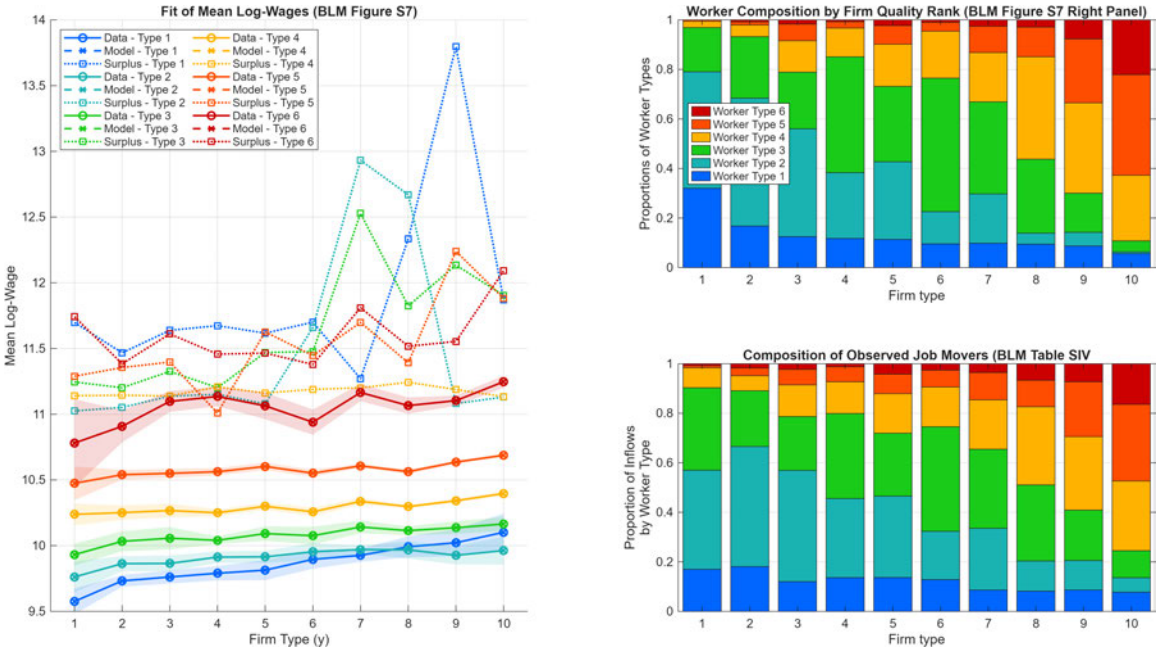



Figure 4.2: Estimated Structural Parameters.

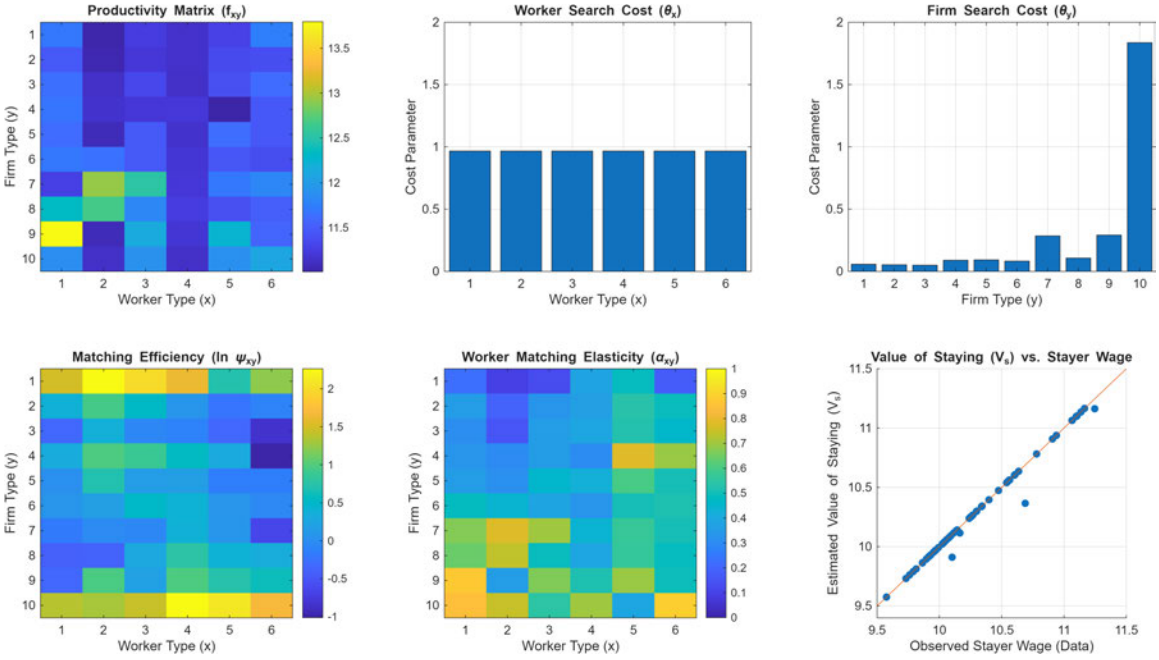
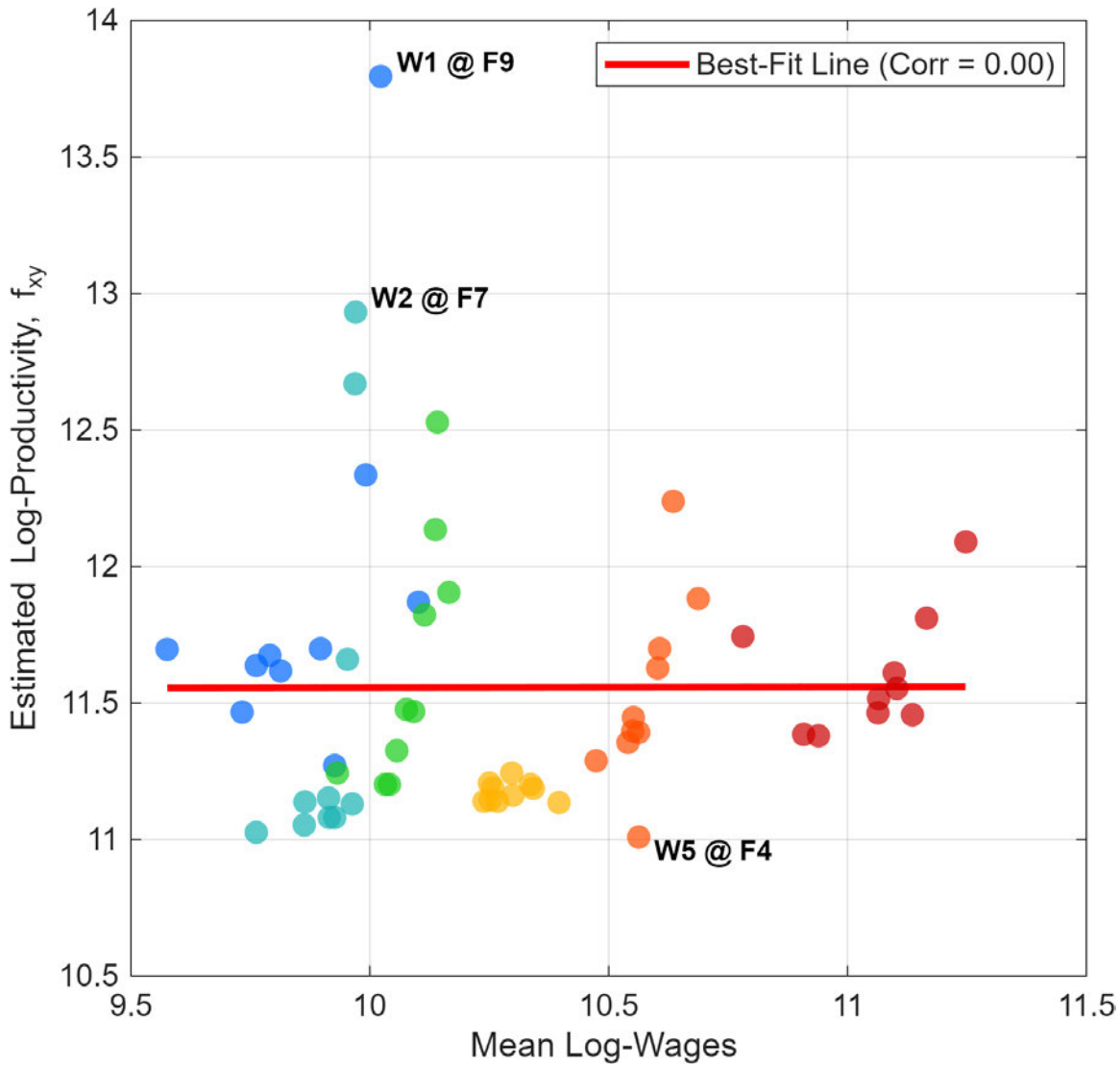


Figure 4.3: The Productivity-Wage Disconnect.




of a match for each firm type (panels) and worker type (x-axis). The match surplus is defined as the gap between the estimated match productivity $f_{xy}$ (red dotted line) and the worker's outside option $V_s$ (magenta line). The solid blue line shows the equilibrium wage, which splits this surplus between the worker and the firm. Crucially, for the vast majority of matches, the equilibrium wage is nearly identical to the outside option. This confirms that firms exercise extreme monopsony power, capturing almost the entire surplus as profit. This structure explains the AKM Puzzle: wages are pinned to the

Figure 4.4: Firm Payoffs and Surplus Decomposition.

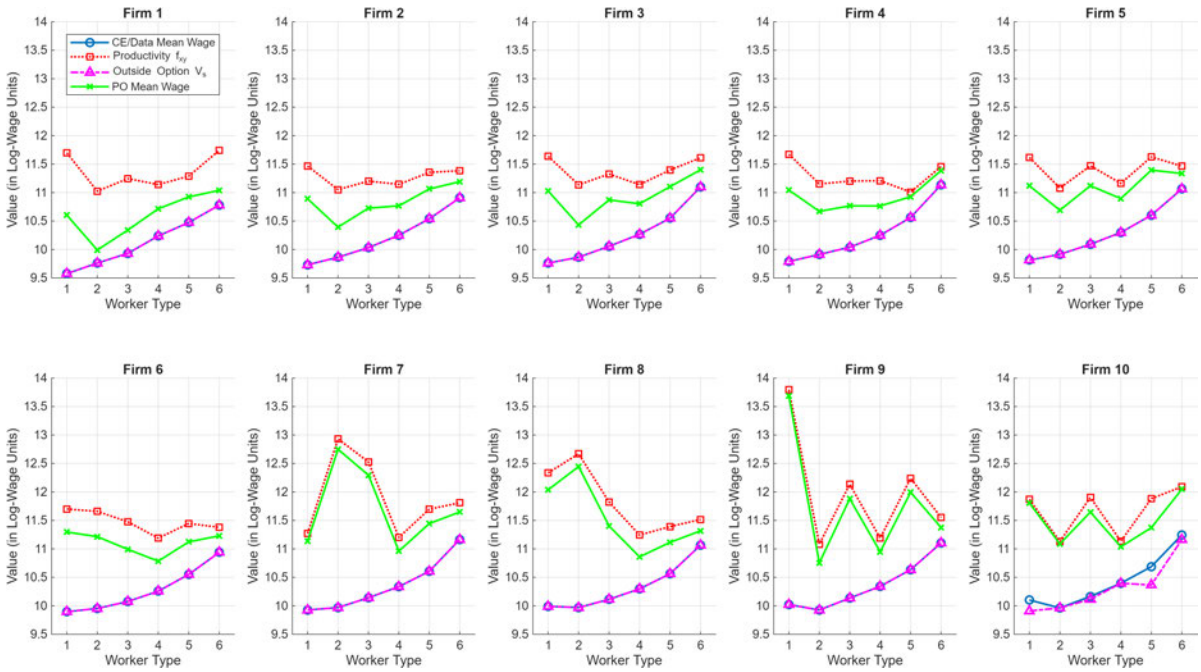
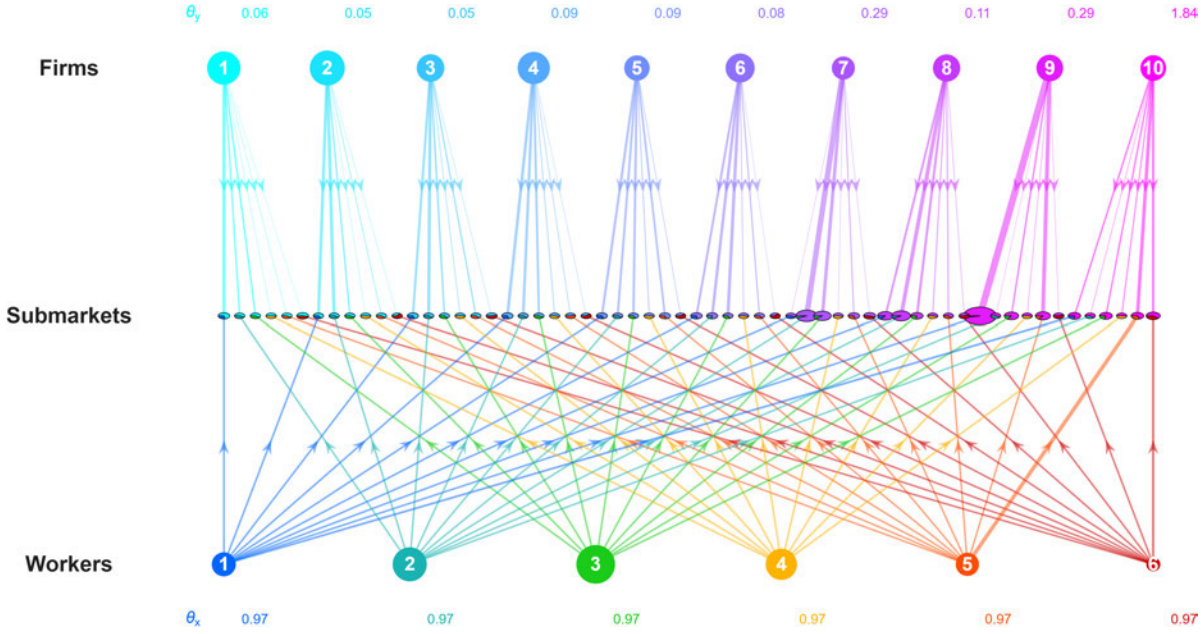



Figure 4.5: Equilibrium Search and Screening Strategies.



worker's history-dependent outside option $V_s$ rather than the current firm's productivity $f_{xy}$. Consequently, wage variation is dominated by worker fixed effects (history), while firm fixed effects are compressed, even though firms differ significantly in productivity. The only notable exception occurs at the highest-paying firms (Types 9-10), where a spike in screening costs (see $\theta_y$ in Figure 4.2) forces them to pay a small competitive premium, raising wages slightly above the outside option.

**Firm-Driven Sorting.** If wages are compressed and offer little signal of productivity, how does the market achieve strong sorting? Figure 4.5 visualizes the equilibrium strategies and outcomes. The circles in the middle layer represent the submarkets for each worker-firm pair. The size of each circle is proportional to the total match surplus ($f_{xy} - V_s$), and the pie charts within them indicate the surplus split: the vast majority is captured by firms (top), with a thin sliver for workers (bottom). The largest surpluses are generated by matches between Worker Type 1 and Firm Type 9, and Worker Type 2 and Firm Type 7. These are the most productive combinations available in the economy. The lines connecting agents to submarkets illustrate the search intensities. Worker strategies (bottom) are largely diffuse or uniform, reflecting the weak wage incentives. However, firm strategies (top) are highly targeted. The thickest purple lines emanate from firms targeting precisely those high-surplus matches (e.g., Firm 9 targeting Worker 1). Because firms capture nearly the entire surplus, they have a powerful incentive to screen for these specific high-value matches.

This confirms that the market operates in a firm-driven sorting regime: the strong sorting observed in the data is engineered by firms' screening efforts $q_y$, which direct the flow of matches toward high-surplus pairs, despite the lack of direction from workers.

**Tension Between Surplus and Matching Technology.** If firms aggressively screen for high-surplus pairs like (Worker 1, Firm 9) and (Worker 2, Firm 7), why do these matches not dominate the aggregate flows shown in Figure 4.1? Why do we instead observe a prevalence of diagonal sorting (e.g., Worker 6 to Firm 10)? The answer lies in the estimated matching technology, shown in Figure 4.2. While the incentive to form off-diagonal, "ladder-climbing" matches is high (massive surplus), the feasibility of these matches is constrained by low matching efficiency ($\psi_{xy}$). The estimation reveals that it is structurally difficult for high-productivity firms to successfully match

with low-history workers, perhaps due to credentialing barriers or network segregation. Conversely, "diagonal" matches (High-High and Low-Low) are technologically easier to form (higher $\psi_{xy}$), even if the marginal surplus generated ($f_{sy} - V_s$) is smaller because the worker already commands a high outside option.

This creates an inefficient but persistent dynamic in the labor market: Firms expend significant screening effort trying to find high-surplus "diamonds in the rough" (the thick screening lines in Figure 4.5 targeting submarkets 1-9 and 2-7), but the matching friction severely limits the realized volume of these transactions. The market settles for a high volume of "lower-net-value" assortative matches because they are the path of least resistance, not because they are the peaks of value creation.

## Welfare and Misallocation

The decoupling of wages from productivity has significant welfare implications. The inefficiency is not merely distributional; the monopsony distortion actively prevents the formation of the economy's most productive potential matches.

The green line in Figure 4.4 plots the planner-prescribed wage $\omega^{PO,W}$: the shadow price required to induce efficient worker application strategies. In almost all submarkets, and especially for the high-productivity pairs, the equilibrium wage is substantially below this efficient level. The gap between the green and blue lines represents the Monopsony Markdown and acts as a tax on match formation. Because firms depress wages to capture rents, workers have insufficient incentive to direct their search effort toward the high-surplus (but difficult-to-match) opportunities like the (1-9) pair.

The consequence is a specific form of misallocation. The economy's first-best matches, those that generate the largest net surplus by moving low-wage workers to high-productivity firms, are systematically under-formed. Although firms exert screening effort to find these matches, their effort is unilateral. Without the corresponding wage signal to induce worker search effort, the matching rate remains inefficiently low, constrained by the low technology parameters.

The planner (taking into account the matching technology) would prescribe a radically different wage structure: significantly higher wages for these specific high-surplus matches to "pull" workers toward them, overcoming the matching friction. Instead, the competitive equilibrium leaves these gains on the table, settling for a high volume

of lower-surplus assortative matches. This suggests that the observed sorting pattern, while strong, is fundamentally constrained by the inability of the price mechanism to guide workers toward the highest-value opportunities.

# 5 Conclusion

This paper develops a unified theory of monopsony power based on information frictions in job search. We model both firms and workers as choosing probabilistic targeting strategies, subject to endogenous information costs. This framework nests both random and directed search as limiting cases, and delivers a tractable wage-posting equilibrium with a closed-form wage equation that isolates the structural determinants of monopsony power. Crucially, our model reveals a novel firm-driven sorting mechanism: firms can exercise substantial market power, compressing wages and flattening worker incentives, while still achieving strong assortative matching through their own costly screening efforts.

We show how wage markdowns emerge endogenously from the interaction of five structural channels: information frictions on both sides of the market, labor market tightness, equilibrium sorting, and the sequential nature of wage setting. Our theoretical analysis predicts that strong sorting can coexist with extremely flat wage-productivity profiles if firms drive the sorting process. This theoretical prediction resolves a central puzzle in the empirical literature, reconciling observed strong sorting with low pass-through of firm productivity to wages.

By estimating our model on detailed moments from the Swedish labor market (Bonhomme, Lamadon, and Manresa, 2019), we provide quantitative evidence that the market operates precisely in this firm-driven sorting regime. The estimation reveals extreme monopsony power, with wages pinned near workers' outside options, generating large markdowns and a striking disconnect between wages and productivity. This simultaneously resolves the empirical puzzles of how strong sorting can coexist with flat wage profiles and how large markdowns can coexist with the small firm fixed effects found in wage data.

However, this alternative sorting regime comes at a significant social cost. The severe wage compression means workers are under-incentivized to direct their search towards the economy's highest-surplus matches. The market thus systematically under-

produces the most productive (and often off-diagonal) pairings, leading to substantial misallocation. While firms diligently screen for high-value workers, their unilateral efforts are insufficient to overcome the dampened worker incentives and technological frictions. A constrained social planner would prescribe significantly higher wages for these highest-value matches to induce efficient worker search, implying that optimal policy could raise wages and welfare without reducing the number of matches.

Our framework offers a tractable and empirically-validated alternative to models that impose wage-setting rules or estimate reduced-form elasticities. By deriving wages, applications, and sorting jointly from primitives, we clarify the mechanisms behind monopsony and provide a powerful diagnostic tool for analyzing the efficiency and equity of market outcomes. This model can serve as a go-to tool for evaluating policy interventions aimed at correcting labor market inefficiencies. Such policies could include modifying the game structure (e.g., promoting simultaneous search or ex-post bargaining), enhancing wage transparency to better align workers' incentives with productivity signals, or implementing a minimum-wage schedule with a gradient. A uniform minimum wage could inadvertently compress the incentive structure further, but a minimum wage that varies by, for example, expected productivity or firm type could strategically re-steepen the wage gradient for high-value matches, thereby incentivizing worker search for these misallocated opportunities. These interventions could shift the market away from the inefficient firm-driven sorting regime, fostering greater dynamism and welfare.

# References


[1] **Abowd, J., Kramarz, F., and D. Margolis (1999)** "High Wage Workers and High Wage Firms." *Econometrica*, 67(2), pp. 251-333. https://www.jstor.org/stable/2999586

[2] **Azar, J. and I. Marinescu (2024)** "Monopsony Power in the Labor Market: From Theory to Policy" *Annual Review of Economics*, 16, pp. 491-518. https://doi.org/10.1146/annurev-economics-072823-030431

[3] **Azar, J., Marinescu, I., and M. Steinbaum (2019)** "Measuring Labor Market Power Two Ways" *AEA Papers and Proceedings*, 109, pp. 317-321. https://doi.org/10.1257/pandp.20191068

[4] **Berger, D., Herkenhoff, K. and S. Mongey (2022)** "Labor Market Power" *American Economic Review*, 112(5), pp. 1147-1193. https://doi.org/10.1257/aer.20191521

[5] **Berger, D., Herkenhoff, K., Kostol, A. and S. Mongey (2023)** "An Anatomy of Monopsony: Search Frictions, Amenities and Bargaining in Concentrated Markets." NBER Working Paper 31149.

[6] **Bils, M., Kaymak, B. and K-J. Wu (2025)** "Robinson Meets Roy: Monopsony Power and Comparative Advantage." NBER Working Paper 34165.

[7] **Boal, W. M., and M. R. Ransom (1997)** "Monopsony in the Labor Market." *Journal of Economic Literature*, 35(1), pp. 86-112. https://www.jstor.org/stable/2729694

[8] **Bonhomme, S., Lamadon, T., and E. Manresa (2019)** "A distributional framework for matched employer employee data." *Econometrica*, 87(3), pp. 699-739. https://doi.org/10.3982/ECTA15722

[9] **Burdett, K. and D.T. Mortensen (1998)** "Wage differentials, employer size, and unemployment." *International Economic Review*, 39(2), pp. 257-273. https://doi.org/10.2307/2527292

[10] **Burdett, K., Shi, S., and R. Wright (2001)** "Pricing and Matching with Frictions." *Journal of Political Economy*, 109(5), pp. 1060-1085. https://doi.org/10.1086/322835

[11] **Card, D. (2022)** "Who Set Your Wage?" *American Economic Review*, 112(4), pp. 1075-1090. https://doi.org/10.1257/aer.112.4.1075

[12] **Chade, H., Eeckhout, J. and L. Smith (2017).** "Sorting Through Search and Matching Models in Economics." *Journal of Economic Literature*, 55(2), pp. 493-544. https://doi.org/10.1257/jel.20150777

[13] **Cheremukhin, A., Popova A. and A. Tutino (2015).** "A Theory of Discrete Choice with Information Costs." *Journal of Economic Behavior and Organization*, 113, pp. 34-50. https://doi.org/10.1016/j.jebo.2015.02.022

[14] **Cheremukhin, A., Restrepo-Echavarria P., and A. Tutino (2020).** "Targeted Search in Matching Markets." *Journal of Economic Theory*, 185, January. https://doi.org/10.1016/j.jet.2019.104956

[15] **Cuturi, M. (2013)** "Sinkhorn distances: Lightspeed computation of optimal transport." In Advances in Neural Information Processing Systems, pages 2292-2300.

[16] **Diamond, P. (1971)** "A Model of Price Adjustment." *Journal of Economic Theory* 3(2), pp. 156-168. https://doi.org/10.1016/0022-0531(71)90013-5

[17] **Dube, A., Naidu S., and A.D. Reich (2022)** "Power and dignity in the low-wage labor market: theory and evidence from Wal-Mart workers." NBER Working Paper 30441.

[18] **Eeckhout, J. and P. Kircher (2010).** "Sorting and Decentralized Price Competition." *Econometrica*, 78(2), pp. 539-574. https://doi.org/10.3982/ECTA7953

[19] **Guerrieri, V., Julien, B., Kircher, P. and R. Wright (2021)** "Directed Search and Competitive Search Equilibrium: A Guided Tour." *Journal of Economic Literature*, 59(1), pp. 90-148. https://doi.org/10.1257/jel.20191505

[20] **Lamadon, T., Mogstad, M., and B. Setzler (2022)** "Imperfect competition, compensating differentials, and rent sharing in the US labor market." *American Economic Review* 112 (1), 169-212. https://doi.org/10.1257/aer.20190790

[21] **Lentz, R., Maibom, J., and E. Moen (2022)** "Competitive or random search?" unpublished manuscript.

[22] **Lester, B. (2011).** "Information and Prices with Capacity Constraints," *American Economic Review*, 101(4), pp. 1591-1600. https://doi.org/10.1257/aer.101.4.1591

[23] **Matejka F. and A. McKay (2015).** "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review*, 105(1), pp. 272-298. https://doi.org/10.1257/aer.20130047

[24] **Manning, A., 2003.** "Monopsony in Motion: Imperfect Competition in Labor Markets." Princeton Univ. Press, Princeton, NJ

[25] **Menzio, G. (2007).** "A Theory of Partially Directed Search," *Journal of Political Economy*, 115(5), pp. 748-769. https://doi.org/10.1086/525544

[26] **Moen, E.R. (1997)** "Competitive search equilibrium." *Journal of Political Economy*, 105 (2), pp. 385-411. https://doi.org/10.1086/262077

[27] **Mortensen, D. T. (1982).** "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, 72(5), pp. 968-979.

[28] **Oberfield, Ezra, Rossi-Hansberg, Esteban, Sarte, Pierre-Daniel, and Nicholas Trachter (2024)** "Plants in Space." *Journal of Political Economy*, 132 (3), pp. 867-909.

[29] **Rabinovich, S., and R. Wolthoff (2022)** "Misallocation inefficiency in partially directed search." *Journal of Economic Theory*, 206(C).

[30] **Robinson, J. (1933)** "Imperfect competition." In: Robinson, J. (1966), An Essay on Marxian Economics. Palgrave MacMillan, UK, pp. 73-81. https://doi.org/10.1007/978-1-349-15228-5_9.

[31] **Shi, S. (2001).** "Frictional Assignment. I. Efficiency." *Journal of Economic Theory*, 98, pp. 232-260. https://doi.org/10.1006/jeth.2000.2713

[32] **Shimer, R. and L. Smith (2001).** "Matching, Search and Heterogeneity." The B.E. Journal of Macroeconomics, 1(1), pp. 1-18. https://doi.org/10.2202/1534-6013.1010

[33] **Staiger, D. O., J. Spetz, and C. S. Phibbs (2010)** "Is there monopsony in the labor market? evidence from a natural experiment." *Journal of Labor Economics* 28(2), pp. 211-236.

[34] **Tarski, A. (1955).** "A Lattice-theoretical Fixpoint Theorem and Its Applications," *Pacific Journal of Mathematics*, 5, pp. 285-309. https://doi.org/10.2140/pjm.1955.5.285

[35] **Thomas, C. and J. Cover (1991)**. "Elements of Information Theory." John Wiley & Sons, Inc.

[36] **Wu L. (2024)** "Partially directed search in the labor market." *Journal of Economic Theory* 220(C).

# Appendix: Submarket Myopia and Equilibrium Multiplicity

In this section, we analyze a behavioral variant of the firm's optimization problem characterized by submarket myopia (or narrow bracketing). In the fully rational model, firms internalize that drawing more applicants of type $x$ expands the aggregate queue $\delta_y$, which compositionally dilutes the frequency of all other applicant types $x' \neq x$ (generating cross-submarket information spillovers). Under submarket myopia, the firm's hiring division optimizes the contract for type $x$ while treating the composition of the rest of the queue as a fixed parameter.

Mathematically, this assumption imposes that cross-type derivatives are zero: $\frac{\partial a_{x'y}}{\partial \omega_{xy}} = 0$ and $\frac{\partial a_{x'y}}{\partial q_y(x)} = 0$ for all $x' \neq x$. The own-type derivatives remain intact: $\frac{\partial a_{xy}}{\partial \omega_{xy}} = a_{xy}(1 - a_{xy})\hat{z}_{xy}$. We allow the congestion elasticities $E_{px}$ and $E_{qy}$ to remain general, accommodating both internalized matching elasticities and the case where congestion is taken as given ($E_{px} = E_{qy} = 1$).

## A1. Firm's Optimization under Myopia

The worker's problem is unaffected by the firm's cognitive assumption, leaving the effective application semi-elasticity as $\hat{z}_{xy} = \frac{z^0_{xy}}{1 - z^0_{xy}\omega_{xy}(E_{px}-1)}$.

The myopic firm maximizes expected profit $Y_y$. Differentiating with respect to $\omega_{xy}$ and applying the myopic derivative $\frac{\partial a_{xy}}{\partial \omega_{xy}}$ yields:

$$\frac{\partial Y_y}{\partial \omega_{xy}} = \delta_y a^*_{xy} q_y(x) \phi_{xy} \left[\hat{z}_{xy} E_{px}(f_{xy} - \omega_{xy}) - 1\right] - \theta_y a^*_{xy}(1 - a_{xy})\hat{z}_{xy} C_x = 0,$$

where $C_x = q_y(x)(\ln q_y(x) + \lambda_y)$. Dividing by $\delta_y a^*_{xy} \hat{z}_{xy} q_y(x) \phi_{xy} E_{px}$ provides the general wage condition under myopia:

$$\omega^*_{xy} = f_{xy}\left[1 - \frac{(1 - a_{xy})\, E_{qy}}{E_{px}}\right] + \frac{1 - a_{xy}}{E_{px}} \frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{\hat{z}_{xy} E_{px}} \tag{5.1}$$

Similarly, differentiating with respect to $q_y(x)$ utilizing the myopic cross-elasticity

$\hat{v}_{xy} = \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}$ yields:

$$\frac{\partial Y_y}{\partial q_y(x)} = \delta_y a^*_{xy}\phi_{xy}(f_{xy}-\omega_{xy})\left[E_{qy} + \hat{v}_{xy}q_y(x)E_{px}\right] - \theta_y a^*_{xy}\left[\hat{v}_{xy}(1-a_{xy})C_x + (\ln q_y(x) + 1 + \lambda_y)\right] = 0.$$

Defining the myopic information spillover as $G^M_{xy} \equiv \frac{\omega_{xy}\hat{z}_{xy}E_{qy}}{q_y(x)}(1-a_{xy})C_x$, we can isolate the screening intensity:

$$\ln q_y(x) + \lambda_y = \frac{\delta_y}{\theta_y}\phi_{xy}(f_{xy}-\omega_{xy})E_{qy} - \frac{1-G^M_{xy}}{1+\omega_{xy}\hat{z}_{xy}E_{qy}}. \tag{5.2}$$

## A2. The Myopic Congestion-as-Given Model ($E_{px} = E_{qy} = 1$)

We now evaluate the specific case where firms are myopic and agents treat the congestion rate as a fixed constant ($E_{px} = E_{qy} = 1$). Under this assumption, the wage and screening spillovers perfectly cancel out, allowing us to recover a highly tractable closed-form solution.

Substituting $E_{px} = E_{qy} = 1$ into the wage condition (5.1), we have:

$$\hat{z}_{xy}\phi_{xy}(f_{xy}-\omega_{xy}) - 1 = \frac{\theta_y/\delta_y}{q_y(x)}(1-a_{xy})C_x\hat{z}_{xy}.$$

Multiplying by $\omega_{xy}$ reveals that the right side is exactly equal to the myopic spillover $G^M_{xy}$. Thus, $G^M_{xy} = \omega_{xy}\hat{z}_{xy}\phi_{xy}(f_{xy}-\omega_{xy}) - \omega_{xy}$. Substituting this identity into the numerator of the screening FOC (5.2) yields:

$$\ln q_y(x) + 1 + \lambda_y = \frac{\delta_y}{\theta_y}\left[\phi_{xy}(f_{xy}-\omega_{xy}) + \omega_{xy}\hat{z}_{xy}\phi_{xy}(f_{xy}-\omega_{xy})\right] - 1 + \frac{\delta_y}{\theta_y}G^M_{xy}$$

$$\ln q_y(x) + \lambda_y = \frac{\delta_y}{\theta_y}\left[\phi_{xy}(f_{xy}-\omega_{xy}) + \omega_{xy}\right] = \frac{\delta_y}{\theta_y}\phi_{xy}f_{xy}.$$

This confirms that the interior shadow cost is exactly $C_x = q_y(x)\left(\frac{\delta_y}{\theta_y}\phi_{xy}f_{xy} - 1\right)$. Substituting this simplified $C_x$ directly back into the wage equation (5.1) yields:

$$\omega^*_{xy} = f_{xy} - \frac{1}{\hat{z}_{xy}} - \frac{\theta_y/\delta_y}{\phi_{xy}q_y(x)}(1-a_{xy})q_y(x)\left(\frac{\delta_y}{\theta_y}\phi_{xy}f_{xy} - 1\right)$$

$$\omega_{xy}^{*} = f_{xy} - \frac{1}{\hat{z}_{xy}} - (1 - a_{xy}) f_{xy} + (1 - a_{xy}) \frac{\theta_y / \delta_y}{\phi_{xy}}.$$

Rearranging terms, we recover the wage equation where the worker is evaluated strictly against their demographic share rather than a cross-applicant outside option:

$$\omega_{xy}^{*} = a_{xy} f_{xy} + (1 - a_{xy}) \frac{\theta_y / \delta_y}{\phi_{xy}} - \frac{1}{\hat{z}_{xy}}. \tag{5.3}$$

## A3. Hessian Curvature and Equilibrium Multiplicity

The interaction between the congestion assumption ($E_{px}$) and the Multinomial Logit application response dictates whether the matching market exhibits a unique equilibrium or multiple local coordination equilibria. This property is governed by the curvature of the firm's revenue function, $H_{\omega\omega} = \frac{\partial^2 Y_y}{\partial \omega_{xy}^2}$.

Taking the second derivative of the expected revenue component yields a structural bracket dependent on the application elasticity:

$$H_{\omega\omega} \propto \left[ \left( \hat{z}_{xy}^2 E_{px} + \frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}} \right) (f_{xy} - \omega_{xy}) - 2\hat{z}_{xy} \right].$$

Let $\tilde{p}_x = \frac{\mu_y}{\delta_x} p_x(y)$ denote the absolute probability of application. Differentiating the effective elasticity $\hat{z}_{xy}$ and factoring the expression yields:

$$\hat{z}_{xy}^2 E_{px} + \frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}} = \hat{z}_{xy}^2 \left[ 2E_{px} - 1 + \frac{(E_{px} - 1) - \frac{\tilde{p}_x}{1 - \tilde{p}_x}}{1 - z_{xy}^0 \omega_{xy} (E_{px} - 1)} \right].$$

Substituting this relation into the structural bracket governing the revenue curvature, $\mathcal{B} \equiv \left[ \left( \hat{z}_{xy}^2 E_{px} + \frac{\partial \hat{z}_{xy}}{\partial \omega_{xy}} \right) (f_{xy} - \omega_{xy}) - 2\hat{z}_{xy} \right]$, and factoring out $\hat{z}_{xy}$ provides the exact condition for local concavity:

$$\mathcal{B} = \hat{z}_{xy} \left\{ \hat{z}_{xy} (f_{xy} - \omega_{xy}) \left[ 2E_{px} - 1 + \frac{(E_{px} - 1) - \frac{\tilde{p}_x}{1 - \tilde{p}_x}}{1 - z_{xy}^0 \omega_{xy} (E_{px} - 1)} \right] - 2 \right\}.$$

To establish conditions for global uniqueness, we evaluate this curvature in the frictionless limit ($\theta \to 0$). As information costs vanish, the unadjusted elasticity diverges ($z_{xy}^0 \to \infty$). Provided the matching technology strictly penalizes congestion ($E_{px} < 1$), the denominator $1 - z_{xy}^0 \omega_{xy} (E_{px} - 1)$ diverges to infinity, driving the fractional term

inside the inner brackets to zero. Simultaneously, the effective elasticity converges to a finite bound: $\lim_{\theta \to 0} \hat{z}_{xy} = \frac{1}{\omega_{xy}(1-E_{px})}$.

The asymptotic sign of the firm's revenue curvature ($H_{\omega\omega}$) is therefore determined by the limit of the inner expression. For the objective function to be strictly concave (ensuring a unique equilibrium), this limit must be strictly negative:

$$\lim_{\theta \to 0} \frac{\mathcal{B}}{\hat{z}_{xy}} = \left( \frac{f_{xy} - \omega_{xy}}{\omega_{xy}(1 - E_{px})} \right) (2E_{px} - 1) - 2 < 0. \tag{5.4}$$

To evaluate this threshold condition explicitly, we substitute the interior myopic wage equation. By isolating the shadow screening cost from the unconstrained first-order conditions, the general target wage can be algebraically simplified to:

$$\omega^*_{xy} = f_{xy} \left[ 1 - \frac{(1 - a_{xy})\, E_{qy}}{E_{px}} \right] + \frac{1 - a_{xy}}{E_{px}} \frac{\theta_y / \delta_y}{\phi_{xy}} - \frac{1}{\hat{z}_{xy} E_{px}}.$$

In the frictionless limit, the direct screening cost term vanishes ($\theta_y \to 0$). Substituting the limiting bound $\frac{1}{\hat{z}_{xy}} \to \omega^*_{xy}(1 - E_{px})$ yields:

$$\lim_{\theta \to 0} \omega^*_{xy} = f_{xy} \left[ 1 - \frac{(1 - a_{xy})\, E_{qy}}{E_{px}} \right] - \frac{\omega^*_{xy}(1 - E_{px})}{E_{px}}.$$

Multiplying by $E_{px}$ and solving for $\omega^*_{xy}$ isolates the fundamental base wage:

$$\lim_{\theta \to 0} \omega^*_{xy} = f_{xy} \left[ E_{px} - (1 - a_{xy}) E_{qy} \right].$$

Substituting the firm's markup ratio, $\frac{f_{xy} - \omega^*_{xy}}{\omega^*_{xy}}$, back into the uniqueness condition (5.4) yields the exact requirement for a globally strictly concave revenue function in terms of the general matching elasticities:

$$\left( \frac{1 - E_{px} + (1 - a_{xy}) E_{qy}}{E_{px} - (1 - a_{xy}) E_{qy}} \right) \frac{2E_{px} - 1}{1 - E_{px}} < 2. \tag{5.5}$$

This complex inequality can be solved algebraically to reveal a clean threshold for global concavity. Let $X = (1 - a_{xy}) E_{qy}$. Expanding the condition $(1 - E_{px} + X)(2E_{px} - 1) < 2(1 - E_{px})(E_{px} - X)$ and canceling the common quadratic and cross-terms reduces

the requirement precisely to:

$$1 - E_{px} > (1 - a_{xy})E_{qy}.$$

To guarantee that the firm's objective function is globally strictly concave, this condition must hold for any endogenous queue composition, meaning it must be satisfied for all $a_{xy} \in [0, 1]$. Rearranging the inequality with respect to the demographic share $a_{xy}$ yields:

$$a_{xy} > 1 - \frac{1 - E_{px}}{E_{qy}}.$$

For this inequality to hold universally, even in the worst-case scenario of a highly diffuse queue where a worker's type constitutes a negligible fraction of the applicant pool ($a_{xy} \to 0$), the right-hand side must be strictly less than zero. This imposes the following restriction on the congestion multipliers:

$$\frac{1 - E_{px}}{E_{qy}} > 1 \implies E_{px} + E_{qy} < 1.$$

This makes it clear that under myopia, the assumption of workers and firms taking congestion as given ($E_{px} + E_{qy} = 2$) always leads to multiplicity in the directed search limit. We can further express this requirement in terms of the underlying physical congestion elasticities. Since $E_{px} = 1 + \varepsilon_{\phi,p}$ and $E_{qy} = 1 + \varepsilon_{\phi,q}$, the condition requires $\varepsilon_{\phi,p} + \varepsilon_{\phi,q} < -1$.

Alternatively, substituting the structural definitions of the matching elasticities ($E_{px} = \varepsilon_{m,c} + (1 - a_{xy})\varepsilon_{m,d}$ and $E_{qy} = \varepsilon_{m,d}$), the global uniqueness condition becomes:

$$\varepsilon_{m,c} + (2 - a_{xy})\varepsilon_{m,d} < 1.$$

Because this must hold for all $a_{xy} \in [0, 1]$, evaluating at the supremum ($a_{xy} = 0$) provides the necessary and sufficient technological condition for unconditional global uniqueness in the frictionless limit:

$$\varepsilon_{m,c} + 2\varepsilon_{m,d} < 1. \tag{5.6}$$

This inequality reveals that Constant Returns to Scale ($\varepsilon_{m,c}+\varepsilon_{m,d} = 1$) is insufficient

to guarantee global uniqueness across all market structures as information costs vanish. Under CRS, diffuse submarkets ($a_{xy} \to 0$) fail the concavity condition, allowing the application elasticity to diverge and reintroducing multiple local coordination equilibria. Therefore, unconditional global uniqueness requires either strong Decreasing Returns to Scale in the matching technology or strictly positive information frictions ($\theta > \underline{\theta}$) to structurally bound the labor supply elasticity.