An Information-based Theory of Monopsony Power

Anton Cheremukhin*

Paulina Restrepo-Echavarria†

July 18, 2026

Abstract

We develop an information-based theory of monopsony power that unifies search frictions and job differentiation by endogenizing worker search and firm screening through rational inattention. The model yields a closed-form wage equation that decomposes the sources of monopsony power and nests directed and random search as limiting cases. We show that information frictions generate localized worker captivity, allowing firms to sustain substantial wage markdowns without relying on assumptions about market concentration or exogenous job differentiation. The framework resolves three empirical puzzles by reconciling strong assortative matching with flat wage-productivity profiles, large wage markdowns with small firm wage effects, and high estimated labor supply elasticities with substantial monopsony power. Estimating the model using moments from Swedish matched employer-employee data, we find that worker information frictions are considerably larger than firms' screening costs, implying that firms drive sorting through selective screening while compressing wages toward workers' outside options. We further show that equilibrium is inefficient because firms suppress wages and distort search incentives, and we quantify the resulting welfare losses, demonstrating that the primary cost of monopsony arises from inefficiently low match formation.

*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, Daniel Haanwinckel, Kyle Herkenhoff, Philipp Kircher, Rasmus Lentz, Benjamin Lester, Guido Menzio, Giuseppe Moscarini, Guillaume Rocheteau, Esteban Rossi-Hansberg, Aysegul Sahin, Kjetil Storesletten, Antonella Tutino, Gianluca Violante, as well as participants of the seminars at the Dallas Fed, St. Louis Fed, UC Irvine, as well as SED, NASMES, MWM meetings for their insightful comments. This work was supported by computational resources provided by the BigTex High Performance Computing Group at the Federal Reserve Bank of Dallas. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Reserve Banks of Dallas or St. Louis or the Federal Reserve System.

1 Introduction

According to existing models of monopsony power firms pay wages below marginal product (Robinson (1933)) because workers face search frictions or because jobs are differentiated (Card, 2022). Yet canonical models typically treat these two sources of monopsony power in isolation, leaving the literature ill-equipped to explain conflicting empirical evidence on worker behavior and wage setting. In this paper, we overcome this limitation by developing an information-based theory that unifies both perspectives. Rather than imposing search behavior mechanically, our framework endogenizes the degree of randomness in job search and the elasticity of labor supply to the firm. This yields a closed-form wage equation that parses the structural drivers of monopsony power and allows us to resolve three prominent empirical puzzles on sorting and inequality.

The first puzzle is the *sorting puzzle*: matched employer-employee data exhibit strong positive assortative matching, yet wage profiles are surprisingly flat with respect to firm productivity (Bonhomme, Lamadon, and Manresa, 2019, henceforth BLM). If wage signals are too weak to incentivize workers to find better jobs, what drives the strong sorting patterns observed in the data? The second puzzle is the *AKM puzzle*: the literature following Abowd, Kramarz, and Margolis (1999) finds that firm fixed effects explain a small share of wage variance, while a parallel literature documents large wage markdowns, implying substantial firm wage-setting power. If firms wield such significant power to dictate pay, why does the identity of the employer explain so little of the variation in wages? And the third puzzle is the *elasticity puzzle*: platform data and quasi-experimental evidence show that job applications respond strongly to posted wages, implying a highly elastic labor supply. Under standard theory, this high responsiveness should severely limit market power. If workers readily redirect their search toward better-paying jobs, how can firms sustain such substantial wage markdowns without losing their applicant pools? Standard models struggle to reconcile these facts: directed search models generate sorting via steep wage gradients, while random search models generate flat wages but struggle to produce sorting.

To study the drivers of monopsony power when there are both search frictions and job differentiation, and resolve these three empirical puzzles, we build on the targeted search framework of Cheremukhin, Restrepo-Echavarria, and Tutino (2020) by intro-

ducing sequential wage posting and strategic firm behavior. The matching process operates as a sequential game: firms first post wage menus and commit to screening policies, after which workers observe these offers and choose their application targets. Both sides of the market utilize probabilistic strategies subject to endogenous information frictions, modeled as a relative entropy cost. Workers face a cognitive cost of deviating from applying randomly, while firms face a parallel cost to selectively screen the resulting applicant pool. This rational inattention structure provides a microfoundation for multinomial logit discrete choice models and allows the framework to nest canonical directed and random search as limiting cases.

Our first contribution is a closed-form expression for the equilibrium wage that isolates the structural determinants of monopsony power. The equation decomposes pay into a base productivity share, a screening cost adjustment, and a monopsony markdown. The anatomy of this markdown algebraically unites several distinct branches of the modern monopsony literature. We demonstrate that information frictions endogenously restrict the set of viable alternatives for workers, creating a localized captivity factor within specific submarkets. This captivity mechanism generates the pricing power typically attributed to granular oligopsony, comparative advantage, or finite-market directed search, allowing the framework to produce substantial markdowns without requiring assumptions about firm concentration or specific productivity distributions. Furthermore, the model provides an informational microfoundation for the upward-sloping labor supply curves traditionally driven by unobserved worker preferences or job differentiation. While unobserved amenity shocks and information costs produce mathematically equivalent multinomial logit choice probabilities, our framework clarifies how they diverge sharply in their welfare implications.

Second, we characterize the socially optimal allocation and demonstrate that the competitive market is generically inefficient. The inefficiency stems from two persistent deviations: firms suppress wages to extract monopsony rents, and they price and screen workers by their prevalence in the applicant pool rather than by their marginal contribution to match creation. Because private pricing responds to rent extraction and demographic pool composition rather than to true technological margins, the market systematically misallocates worker search effort. By microfounding these distortions, our framework provides a practical toolkit for applied researchers to gauge the local allocative welfare losses from monopsony pricing. We develop a bridge from reduced-

form empirical estimates of labor supply to normative welfare metrics, showing how observed behavioral responses must be re-evaluated at the efficient benchmark, where congestion feedback damps the application response. This translation allows policymakers to evaluate targeted interventions, such as graduated minimum wage schedules, using observable data rather than full structural estimation.

Third, we estimate the model using moments from the Swedish labor market to demonstrate how this framework resolves the three empirical puzzles mentioned above. The estimation reveals a pronounced asymmetry in information frictions: worker search costs are high, while firm screening is inexpensive. This asymmetry grants firms substantial monopsony power, enabling them to compress wages down to the workers' history-dependent outside options. This dynamic resolves the AKM puzzle. Because firm productivity is absorbed into profits rather than passed through to wages, the employer's identity leaves a minimal footprint on compensation. Furthermore, these flat wage profiles provide weak directional signals, leaving worker applications diffuse. The market instead operates in a firm-driven sorting regime. Assortative matching is sustained by firms, who leverage their low-cost screening technology to select the most productive pairings from their applicant pools, resolving the sorting puzzle. Finally, this regime resolves the elasticity puzzle. Because worker search costs are high, the baseline labor supply is highly inelastic across most of the market. However, firms direct intense screening toward high-value targets, which locally elevates meeting rates and amplifies worker responsiveness in those specific submarkets. Because empirical estimates rely on realized matches, they systematically over-sample these targeted areas. The high aggregate elasticity therefore reflects the behavior of workers in the narrow submarkets where firms concentrate their hiring activity, rather than indicating widespread worker mobility or bargaining power.

Finally, we quantify the macroeconomic consequences of this firm-driven sorting regime. We find that the competitive market is inefficiently over-selective. Because firms possess substantial monopsony power and face low screening costs, they hoard their screening capacity for the most lucrative pairings. They suppress wages and refuse to pay the premiums necessary to clear the rest of the market, artificially restricting the overall volume of employment. The social planner corrects this misallocation by prescribing higher wages to induce broader worker search. This intervention shifts the economy from a restricted state to a high-volume state, increasing the total number

of matches by 82 percent. By bringing a large swath of previously unmatched workers into the market, the planner accepts a reduction in average net surplus per match of 0.16 log points. The primary cost of monopsony in this environment is therefore an extensive margin failure, where severe rent extraction prevents the formation of mutually beneficial employment relationships.

The paper proceeds as follows. Section 2 presents the model of sequential targeted search with endogenous information costs and characterizes both the competitive equilibrium and the social planner's allocation. Section 3 dissects the closed-form wage equation to isolate the structural sources of monopsony power, establishes the framework for quantifying welfare losses, and defines the two regimes of sorting. Section 4 presents the empirical estimation on the Swedish labor market, resolves the three monopsony puzzles, quantifies the macroeconomic misallocation, and examines robustness. Section 5 concludes. Proofs and supplemental derivations are provided in the Appendix.

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 costless targeting, leading to submarket-based equilibria that are often efficient and feature assortative matching (Eeckhout and Kircher, 2010; Guerrieri, Julien, Kircher, and Wright, 2021). Our framework generalizes this class of models by introducing information costs that determine how accurately workers and firms can target one another. A companion paper explores in depth how our framework maps into the canonical models of this literature, showing how finite information costs resolve coordination multiplicities and reveal that captive-audience monopsony branches persist alongside efficient equilibria at the frictionless limit. In the current paper, we focus on the empirical implications of these frictions. Specifically, our firm-driven sorting regime provides a structural mechanism for the hypothesis of Eeckhout and Kircher (2011), demonstrating how monopsony power can render wage data largely uninformative about underlying sorting patterns. A broader perspective on these modeling choices and their implications for sorting and matching

technologies is provided in Chade, Eeckhout, and Smith (2017).

Second, we contribute to the monopsony literature. Recent empirical studies document substantial wage markdowns across a wide range of labor markets (e.g., Staiger, Spetz, and Phibbs, 2010; Azar, Marinescu, and Steinbaum, 2019; Dube, Naidu, and Reich, 2022; see Azar and Marinescu, 2024, for a survey). Our model unites diverse sources of monopsony power through an endogenous captivity factor. Because information frictions make searching costly, workers narrow their applications to a limited set of targets, leaving them captive to those specific submarkets. This captivity mechanism provides a common algebraic foundation for the pricing power typically attributed to granular oligopsony (Berger, Herkenhoff, and Mongey, 2022; Berger, Herkenhoff, Kostol, and Mongey, 2023) or comparative advantage (Bils, Kaymak, and Wu, 2025). We generate this concentration endogenously without needing to assume finite firm counts or specific productivity distributions. Furthermore, the framework provides an informational microfoundation for the upward-sloping labor supply curves traditionally driven by job differentiation or unobserved worker preferences (Manning, 2003; Card, Cardoso, Heining, and Kline, 2018; Lamadon, Mogstad, and Setzler, 2022). While unobserved compensating differentials (Sorkin, 2018; Hall and Mueller, 2018) and information costs can produce mathematically equivalent multinomial logit choice probabilities, our framework highlights how they diverge in their normative implications. Treating dispersed choices as consumed amenity surplus masks the allocative welfare losses and dissipated resource costs that arise when search is genuinely frictional.

Third, our framework offers an alternative resolution to the empirical puzzles surrounding sorting and wage variance, contrasting with dynamic or ex-post mechanisms in the literature. To explain why firm fixed effects account for a small share of wage variance, models relying on sequential auctions (Postel-Vinay and Robin, 2002; Cahuc, Postel-Vinay, and Robin, 2006) use dynamic offer-arrival timing to pin wages to a worker's employment history. We replicate this cross-sectional outcome, where wages track history-dependent outside options rather than current firm identity, using static information frictions. Furthermore, while Taber and Vejlin (2020) incorporate outside-offer renegotiation to show that search frictions primarily affect the level rather than the variance of wages, our model generates this configuration endogenously through the monopsony markdown. To reconcile strong sorting with these compressed wage profiles, existing approaches often depend on ex-post mechanisms. For instance, Borovičková

and Shimer (2025) rationalize the empirical sorting patterns by relying on ex-post selection through unobserved match-specific shocks in a random search environment, while Hagedorn, Law, and Manovskii (2017) and Lopes de Melo (2018) demonstrate that bargaining can decouple wages from firm rankings. Our resolution operates instead through ex-ante strategic firm screening and wage posting. By allowing firms to target specific worker types, we show that sorting can be entirely firm-driven. This allows the framework to jointly address the elasticity puzzle and uncover the associated welfare misallocation without relying on dynamic contracting or ex-post match shocks.

Fourth, we contribute to the literature that nests directed and random search to generate intermediate degrees of randomness in matching. Early approaches typically rely on mechanical hybrid structures, such as cheap-talk partitions (Menzio, 2007) or exogenous mixtures of informed and uninformed searchers (Lester, 2011; Rabinovich and Wolthoff, 2022). More recent work incorporates choice frictions, applying a worker-side entropy penalty within a dynamic Nash bargaining environment (Lentz, Maibom, and Moen, 2022). We advance this literature by building on our foundational work in Cheremukhin, Restrepo-Echavarria, and Tutino (2020), which introduced endogenous information costs to govern the targetability of search on both sides of a matching market. Extending that theoretical machinery to sequential wage posting generates a continuous spectrum between uniform random matching and directed search. Operating within this specific information-frictional paradigm, Wu (2024) studies a labor market environment that constitutes a special case of our framework, restricting the analysis to a single worker type and assuming zero screening costs for firms. By maintaining two-sided heterogeneity and active firm screening, our generalized model characterizes the full structure of monopsony power across heterogeneous submarkets and identifies the firm-driven sorting regime that explains the matched employer-employee data.

Finally, our approach rests on the decision-theoretic foundations of rational inattention (Sims, 2003; Matějka and McKay, 2015), which provide the microfoundation for the entropic information costs. Methodologically, introducing these costs connects to recent advances utilizing entropic regularization in optimal transport (Cuturi, 2013) and convex approximations in spatial assignment models (Oberfield, Rossi-Hansberg, Sarte, and Trachter, 2024). This mathematical structure smooths the payoff discontinuities inherent in frictional assignment, rendering previously intractable coordination games smoothly estimable and allowing the data to determine the equilibrium configuration.

2 Model

We develop a model of a frictional labor market in which heterogeneous workers and firms strategically search for each other. Our model 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 social planner's solution, and the welfare cost of the competitive equilibrium. Throughout this section we keep the derivations to the minimum necessary; complete step-by-step derivations are collected in Appendices A–E.

2.1 The Economic Environment

The economy contains a large, finite number of workers indexed by type $x \in \{1, ..., W\}$ and firms indexed by type $y \in \{1, ..., F\}$, and there is a measure μ_x of workers of type x and a measure μ_y of firms of type y.[1]

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 ω_{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: $f_{xy} \geq \omega_{xy} \geq 0$.

Sequential Choices. The matching process unfolds in three stages. First, firms post type contingent wage menus to attract workers (they post a different wage for each worker of type x, $\{\omega_{xy}\}_{x=1}^{W}$) and they commit to a screening strategy specifying which worker types they will later pursue among their applicants. This gives them a first mover advantage because they can anticipate how workers will react in response to their choices. Second, workers observe the wages and screening strategies of firms and choose probabilistic application strategies to try to target their preferred job. Third, firms observe their applicant pools (which are equal to those they expected) and apply the

[1]We model each type as a continuum of identical agents, with masses μ_x and μ_y. Individual application, screening, and matching outcomes are binary draws; by the law of large numbers, the corresponding population masses (c_{xy}, the pool composition a_{xy}, examinations d_{xy}, and matches M_{xy}) equal their expected values exactly, so all aggregate quantities are deterministic and "ex ante" and "ex post" coincide. Masses are real-valued measures, not headcounts.

probabilistic screening strategies that they committed to in the first stage to generate meetings with the workers. After meetings are realized matches are formed according to a matching technology and the game ends. Below we describe these application and screening search strategies, the information frictions that govern them, and the technology that transforms choices into matches.

Search Strategies and Information Costs. Agents face an information friction: although workers know the wage that they would get if matched with a specific firm of type y, they don't know who is who, so they cannot perfectly and costlessly submit an application to a specific type of firm. Instead, they exert search effort at a given cost to try to identify who is who. This results in them choosing a discrete probability distribution over types, aiming to target their preferred type(s) of firm(s). This probability distribution represents the probability with which a worker of type x will submit an application to a firm of type y, and we denote it by $\bar{p}_x(y)$.

Likewise, firms know the composition of their pool of applicants but don't know who is who, so they exert search effort to try to locate their preferred worker type. They choose a discrete probability distribution over worker types that represents the probability with which they will consider the application of a particular worker of type x that has submitted an application to their firm. We denote this probability $\bar{q}_y(x)$.

We measure search effort as the relative entropy (Kullback-Leibler divergence) between the chosen strategy (discrete probability distribution) and a uniform prior. For the workers, the prior is given by $\tilde{p}_x(y) = 1/\delta_x$, where $\delta_x = \sum_y \mu_y$ is the total number of firms.[2]

We define the distance between the chosen strategy (or posterior distribution) and the uniform prior of the worker, as their search intensity $p_x(y) = \bar{p}_x(y)/\tilde{p}_x(y)$, such that the workers' search effort is given by

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

[2]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 Cover and Thomas (1991), Chapter 2). See also Cheremukhin, Popova, and Tutino (2015) and Matejka and McKay (2015) for applications to stochastic discrete choice with information costs.

Firms do not screen the general population; they screen only those who apply to them. Therefore, the firm's prior is the empirical composition of its applicant pool, not the population distribution. The pool composition of a firm y is defined by $a_{xy} = \frac{\mu_x p_x(y)}{\delta_y}$, where $\delta_y = \sum_x \mu_x p_x(y)$ is the effective (expected) applicant pool size. Hence, the firm's screening effort is given by

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

where $q_y(x)$ is the screening intensity relative to the prior, $q_y(x) = \bar{q}_y(x)/\tilde{q}_y(x)$.[3]

This means that if a worker or a firm doesn't want to exert search effort because it is too costly, they can acquire no information about how to locate a specific counterpart, and choose a uniform probability distribution over firm types as their strategy. This would result in a search cost of zero because there is no distance between the posterior chosen distribution and the prior distribution.

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

The Matching Technology and Congestion. Given the worker's targeting intensities, $p_x(y)$, the mass of workers of type x applying to firms of type y is given by $c_{xy} = \mu_x p_x(y) \frac{\mu_y}{\delta_x}$. The term μ_y/δ_x converts individual targeting intensities into mass of applications, reflecting that workers choose among firm types rather than individual firms. Similarly, given $q_y(x)$, the mass of firms of type y that looks at the application of workers of type x out of their pool of applicants is given by $d_{xy} = \mu_y q_y(x) a_{xy}$. Screening is the firm's counterpart to application: just as a worker directs applications toward a firm type, a firm directs its limited attention toward worker types within its pool. Neither side identifies individuals; a worker-firm type combination is simply the submarket in which a meeting occurs, so the firm pays the committed wage ω_{xy} without inspecting applicants.

In sub-market (x, y), the masses c_{xy} and d_{xy} of type-x workers and type-y firms come together and attempt to pair off. Because both sides are homogeneous, who meets

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

whom is irrelevant and coordination is imperfect: in the spirit of the urn-ball, or "agents meeting in a bar," technology, some agents are approached by several counterparts at once while others meet no one. Following the literature, the resulting number of matches is given by a matching function $M(c_{xy}, d_{xy})$ of the two masses present in the sub-market.

To facilitate the individual optimization problems, we rewrite the matching function via a *congestion function* ϕ_{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 the matching technology M to sort out the imperfect coordination that arises within the sub-markets.[4] With this setup not all meetings turn into matches, leaving some vacancies unfilled and some workers unemployed even though no one is turned away. The relationship between the congestion function and the total number of matches is given by:

$$\phi\left(\mu_x, \mu_y, p_x\left(y\right), q_y\left(x\right)\right) = M(c_{xy}, d_{xy})/\mu_x\mu_y p_x(y) q_y(x).$$

Assumption 3. The matching function M (and congestion function ϕ) are twice continuously differentiable in strategies p and q.

As in standard models of frictional markets, agents are atomistic and take matching rates as given: individual workers and firms do not internalize how their own search and screening strategies marginally affect the sub-market congestion rates ϕ_{xy}.

Assumption 4. Agents do not internalize congestion: workers and firms take the congestion function, ϕ, as given when choosing their strategies.[5]

While these congestion effects do not enter private decisions, they are central to the efficiency analysis, because the social planner does internalize them. It is useful to record the elasticities of the congestion function with respect to the chosen strategies, $\varepsilon_{\phi,p} \equiv \frac{\partial \phi_{xy}}{\partial p_x(y)} \frac{p_x(y)}{\phi_{xy}}$ and $\varepsilon_{\phi,q} \equiv \frac{\partial \phi_{xy}}{\partial q_y(x)} \frac{q_y(x)}{\phi_{xy}}$. From the properties of the matching function, the corresponding elasticities of the total match probability with respect to applications and screening are $1 + \varepsilon_{\phi,p} = \varepsilon_{m,c} + (1 - a_{xy})\varepsilon_{m,d}$ and $1 + \varepsilon_{\phi,q} = \varepsilon_{m,d}$, where $\varepsilon_{m,c}$ and $\varepsilon_{m,d}$ are the elasticities of the matching function with respect to the effective masses

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

[5]A companion paper develops the general version of the model in which workers and firms partially internalize congestion and firms internalize the compositional spillovers across their applicant pools. The generalization adds considerable complexity without changing the empirical conclusions, so we adopt the transparent non-internalizing benchmark throughout.

of searchers on each side. These objects reappear in the social planner's problem in Section 2.5.

2.2 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 $\{\omega_{xy}\}$, the firm's screening intensities $q_y(x)$, and the submarket congestion rates ϕ_{xy} as given. They choose $p_x(y) \in S_x$ to maximize expected wage income net of search costs:

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

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

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

Proof. Because the worker takes the congestion rates ϕ_{xy} as given, the expected-wage term is linear in $p_x(y)$, while the entropic cost $p \ln p$ is strictly convex; for $\theta_x > 0$ the objective is therefore strictly concave on the convex compact set S_x, and the marginal cost of search diverges as probability mass vanishes, guaranteeing a unique interior maximum. Setting up the Lagrangian with multiplier $\theta_x \lambda_x$ on the constraint, the first-

order condition is

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

Exponentiating and using the constraint to solve for the multiplier yields (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 expected return of an application, $q_y(x)\phi_{xy}\omega_{xy}$: the product of the expected meeting rate and the posted wage.

Wage Elasticities of Labor Supply. The parameter θ_x (the marginal cost of exerting search effort) governs the baseline sensitivity of the application response to wages. A single elasticity object, formally derived in Appendix A, summarizes the mechanics of the worker's choice and plays a central role in everything that follows: the wage semi-elasticity of applications,

$$z_{xy} \equiv \frac{q_y(x)\phi_{xy}}{\theta_x/\delta_x}\left(1 - \frac{\mu_y}{\delta_x}p_x(y)\right), \tag{2.2}$$

which captures how workers respond to wage variations. It is the familiar MNL own-price semi-elasticity: proportional to the expected meeting rate, inversely proportional to the information cost, and damped by the own-application share, $\frac{\mu_y}{\delta_x}p_x(y)$. Because agents take congestion as given (Assumption 4), z_{xy} is also the application response the firm perceives when it evaluates its posted wage, $z_{xy} = \frac{d\ln p_x(y)}{d\omega_{xy}}$, holding ϕ_{xy} fixed.

This elasticity formally defines the effective labor supply curve facing the firm in the first stage. Examining its limits reveals how information costs bound the firm's pricing power across the two friction regimes of the model. When $\theta_x \to \infty$, information costs dominate, the elasticity z_{xy} approaches zero, and workers apply randomly across all firms. When $\theta_x \to 0$, information frictions vanish and z_{xy} diverges, eroding the firm's pricing power—unless equilibrium sorting concentrates the worker's applications on a single firm type, in which case the captivity factor $\left(1 - \frac{\mu_y}{\delta_x}p_x(y)\right)$ keeps the elasticity bounded. We return to this captivity mechanism, which is central to our results, in Section 3.

2.3 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 that dictates which applications they will look at $\{q_y(x)\}$. By committing to both instruments simultaneously, firms overcome the time-inconsistency problem that would arise if they were to choose their strategies after their pool of applicants has formed. Firms possess a first-mover advantage: they fully internalize how their posted contracts affect the workers' application strategies (Eq. 2.1) and, through them, the firm's own standing in its applicant pool (a_{xy}), while taking the physical congestion of the submarket (ϕ_{xy}) as given, in line with Assumption 4.[6]

A parallel form of non-internalization concerns *informational* congestion. The contract a firm posts for worker type x also alters the pool shares of every other applicant type: attracting additional type-x applications mechanically dilutes $a_{x'y}$ for all $x' \neq x$ (the cross-derivatives $\frac{\partial a_{x'y}}{\partial \omega_{xy}}$ and $\frac{\partial a_{x'y}}{\partial q_y(x)}$). In the same spirit as Assumption 4, we assume that firms do not take these compositional spillovers into account: the firm exhibits *submarket myopia* (narrow bracketing), evaluating each worker type in isolation while internalizing the effect of its contract on that type's own pool share. This is also the more plausible reading of how decentralized human-resources decisions are made.

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:

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

It is important to reiterate the interpretation of the firm's choice set and cost function. While the firm knows the aggregate population μ_x, it can only form matches with workers in its applicant pool. The term a_{xy} in the admissibility constraint and in the

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

cost function ensures the firm chooses among workers in their applicant pool and pays to pursue 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 characterize the optimal wages and screening intensities. The equilibrium wage has a closed-form solution with bounds:*

$$\omega_{xy}^* = \left[f_{xy}\, a_{xy} + \frac{\theta_y/\delta_y}{\phi_{xy}} \left(1 - a_{xy}\right) - \frac{1}{z_{xy}} \right]_0^{f_{xy}}. \tag{2.3}$$

Moreover, whenever the firm retains a positive margin on the match ($\omega_{xy}^* < f_{xy}$), the optimal screening strategy collapses to a simple multinomial logit driven entirely by the physical surplus:

$$q_y^*(x) = \frac{\exp\left(\frac{\delta_y}{\theta_y}\phi_{xy} f_{xy}\right)}{\sum_{x'=1}^{W} a_{x'y}^* \exp\left(\frac{\delta_y}{\theta_y}\phi_{x'y} f_{x'y}\right)}. \tag{2.4}$$

Proof. See Appendix B. □

Equation (2.3) provides the exact closed-form solution for the equilibrium wage. Because of its complexity and central importance to our results, we defer a detailed term-by-term economic deconstruction of this equation to Section 3.

The screening strategy (Equation 2.4) contains a striking simplification that makes the equilibrium tractable: at an interior wage, the wage-dependent terms cancel out of the combined first-order conditions, and screening depends only on the physical surplus f_{xy}. The same simple rule continues to apply at the lower corner, where the monopsony markdown censors the wage: there the worker's application response to screening vanishes along with the worker's gain, and the wage-dependent terms drop out on their own. This robustness is central to the results that follow: firms keep sorting applicants by surplus even in submarkets where wages are pinned to the floor. Only at the upper corner ($\omega_{xy}^* = f_{xy}$), where the firm surrenders the entire surplus, does screening deviate from (2.4); that case is characterized by the general first-order condition (B.2) in Appendix B. With the agents' optimal strategies fully characterized, we can now define the aggregate market equilibrium.

2.4 Equilibrium

An equilibrium requires that agents optimize taking the congestion rates of the market as given, 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 $\{q_y^*(x), \omega_{xy}^*\}$ and the submarket congestion rates ϕ_{xy}.

2. **Firm Optimization:** For every type y, $\{q_y^*, \omega_{xy}^*\}$ solves the firm's problem. Firms internalize the worker's full best response function $p_x^*(\omega, q)$ and its effect on their own pool shares a_{xy}, taking the congestion rates ϕ_{xy} as given.

3. **Consistency:** The congestion rates ϕ_{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 ϕ_{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. There also exists a threshold $\underline{\theta}$ such that for sufficiently high search costs $(\theta_x, \theta_y) > \underline{\theta}$, the matching equilibrium is unique. In the frictionless limit ($\theta \to 0$), there are generically multiple coordination equilibria.*

Proof. See Appendix C. □

The multiplicity in the frictionless limit is not a defect of the model; it is the imprint of the coordination multiplicities known from directed search (e.g., Burdett, Shi, and Wright, 2001). Information costs act as an equilibrium selection device: for θ above the threshold, the equilibrium is unique and varies smoothly with fundamentals, which is what makes the structural estimation in Section 4 feasible.

2.5 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:

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

The planner's first-order conditions, derived in Appendix D, take a transparent form. The screening condition equates the marginal entropy cost of screening to the social return of a match—the total surplus scaled by the firm's technological contribution:

$$\frac{\theta_y / \delta_y}{\phi_{xy}} \left(\ln q_y(x) + 1 + \lambda_y \right) = f_{xy} \varepsilon_{m,d}. \tag{2.5}$$

The application condition is more subtle, because an extra application to firm y changes the applicant pool that all other applicants face. It generates 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.

Remarkably, at the planner's optimum these two externalities exactly offset: substituting the planner's screening condition (2.5) into the marginal social cost, the pool outside option cancels out of both sides (Appendix D), and the planner's application condition reduces to a purely submarket-local expression:

$$\frac{\theta_x/\delta_x}{q_y(x)\phi_{xy}}\left(\ln p_x(y) + 1 + \lambda_x\right) = f_{xy}\varepsilon_{m,c} + \left(1 - \frac{1}{q_y(x)}\right)\frac{\theta_y/\delta_y}{\phi_{xy}}. \tag{2.6}$$

This cancellation (the Envelope Theorem at work, since the planner sets screening optimally) is the key to the efficiency analysis below: the planner wants workers to respond only to the surplus they personally help create, not to the composition of the pool they join.

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 (2.6).

Proposition 4. *Under Assumptions 1-4, the efficient shadow wage required to induce the socially optimal application strategies is:*

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

The firm's residual shadow profit is simply the remaining surplus: $\pi_{xy}^{PO,F} = f_{xy} - \omega_{xy}^{PO}$. The planner's screening strategies, however, cannot be decentralized by wages alone: the firm's private return to screening is the full surplus f_{xy}, whereas the social return is the technological share $f_{xy}\varepsilon_{m,d}$.

Proof. First, compare the socially optimal application condition to the worker's private

FOC:

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

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

Because the worker takes congestion as given, the posted wage transmits the marginal social return to an application one-for-one, without any baseline scalar distortion. Equating the right-hand sides isolates the shadow wage ω_{xy}^{PO} that aligns private and social application incentives.

Second, compare the socially optimal screening condition to the firm's private screening FOC for an interior wage:

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

$$\text{Firm: } \frac{\theta_y/\delta_y}{\phi_{xy}}\left(\ln q_y(x) + 1 + \lambda_y\right) = f_{xy}$$

Because the firm takes the congestion rate as given, it credits its screening with the full marginal surplus f_{xy}, whereas the marginal social product of screening is only the technological share $f_{xy}\varepsilon_{m,d}$. The two conditions coincide only in the knife-edge case $\varepsilon_{m,d} = 1$, in which matches are linear in screening. □

The standard efficiency requirement dictates that an agent receives a share of the physical surplus exactly equal to the elasticity of the matching function with respect to their input ($\varepsilon_{m,c}$), mirroring the classic Hosios (1990) condition. The second term in the efficient wage 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 pool, a heavily targeted worker ($q_y(x) > 1$) is systematically easier for the firm to identify relative to the baseline. The planner recognizes this structural efficiency and passes the information-cost savings onto the worker via a higher shadow wage.

Proposition 4 also reveals an asymmetry between the two margins. The worker's application margin is guided purely by the wage signal: because the worker takes congestion as given, a suitably chosen wage schedule can implement the planner's application strategies exactly. The firm's screening margin has no such corrective price. The firm

is the residual claimant on the matching technology, and because it does not internalize congestion, it perceives the full surplus rather than the technological share $\varepsilon_{m,d}$ as its return to screening. Since screening intensities must average to one within the pool, this inflated perceived return makes private screening differentiate across worker types more sharply than the planner's: the firm is systematically over-selective, a force that resurfaces quantitatively in Section 4.

Proposition 5. *Under Assumptions 1-4, for finite positive costs of search ($\theta_x, \theta_y > 0$), the competitive equilibrium is generically constrained inefficient.*

Proof. For the competitive market to achieve the socially efficient allocation, the competitive wage must equal the shadow wage ($\omega_{xy}^{CE} = \omega_{xy}^{PO}$) in every submarket. Subtracting the shadow wage of Proposition 4 from the competitive interior wage (2.3) isolates the residual distortions that would have to perfectly offset each other:

$$f_{xy}\left(a_{xy} - \varepsilon_{m,c}\right) + \frac{\theta_y/\delta_y}{\phi_{xy}}\left(\frac{1}{q_y^*(x)} - a_{xy}\right) = \frac{1}{z_{xy}}.$$

The left side collects the firm's demographic (rather than technological) valuation of the worker's baseline surplus share and its biased subjective valuation of the screening premium, both of which depend on the endogenous pool composition. The right side is the monopsony markdown, which pushes the wage downward based on the exogenous worker information cost θ_x. Because the left side is endogenously determined by cross-submarket heterogeneity and the right side is exogenous, this equality generically fails to hold for any finite $\theta_x > 0$. Moreover, even at wages satisfying this condition, the screening margin remains distorted (Proposition 4), so the competitive allocation generically differs from the planner's. □

The comparison in the proof isolates three wedges between competitive and efficient pricing. First, the myopic firm anchors the base wage to the worker's demographic share of the applicant pool, a_{xy}, rather than to the technological contribution $\varepsilon_{m,c}$ that the Hosios logic prescribes: it pays workers for their prevalence in the pool, not for their marginal contribution to match creation. Second, the firm evaluates the screening cost premium against the worker's raw demographic share $(1 - a_{xy})$, whereas the planner benchmarks it against the worker's relative targeting intensity $(1 - 1/q_y(x))$: the firm's subjective reference point misprices the information-cost savings that heavily targeted

workers generate. Third, and most importantly, the monopsony markdown $1/z_{xy}$ has no counterpart in the planner's shadow wage. It is pure rent extraction, made possible by the firm's first-mover advantage over an imperfectly elastic application response.

Note that the applicant pool composition enters neither wage through cross-applicant terms: because the myopic firm evaluates each worker type in isolation, its wage contains no compensation for the rest of the pool, and the Envelope-Theorem cancellation of the physical cannibalization and information spillover externalities (2.6) removes pool composition from the planner's application condition as well. Both the competitive and the efficient wage are submarket-local objects, which makes the three wedges above a complete description of the pricing distortion.

Ultimately, the competitive market is inefficient because firms set wages to extract rents rather than to price the true social product of an application, distorting relative price signals away from pure physical surplus, while their screening is too selective relative to the planner's. We systematically dissect these distortions using the closed-form wage equation, quantify the resulting aggregate welfare loss, and explore how these forces shape equilibrium sorting in Section 3.

3 Unifying Search and Monopsony: Intuition and Sorting Regimes

In this section, we deconstruct the closed-form wage equation to build intuition for the economic mechanics of the model. We then demonstrate how this framework algebraically unifies the modern monopsony and search literatures, provides a basis for normative welfare quantification, and defines the conditions under which sorting decouples from wage signals, setting the stage to resolve the empirical puzzles in Section 4. We focus on the mechanisms that organize the literature and resolve the empirical puzzles.

3.1 The Anatomy of Wage-Setting Power

Rewriting the closed-form solution (2.3) with its components labeled, the firm's optimal posted wage decomposes into three additively separable structural parts:

$$\omega^*_{xy} = \left[\underbrace{f_{xy}\, a_{xy}}_{\text{Base Productivity Share}} + \underbrace{\frac{\theta_y/\delta_y}{\phi_{xy}}\,(1 - a_{xy})}_{\text{Screening Cost Premium}} - \underbrace{\frac{1}{z_{xy}}}_{\text{Monopsony Markdown}} \right]_0^{f_{xy}} \tag{3.1}$$

The outer brackets denote censoring to the feasible range $[0, f_{xy}]$: when the interior expression falls outside this range, the wage binds at the corner and the wage first-order condition holds as an inequality. As established in Proposition 2, the simple screening rule (2.4) continues to apply at the lower corner; only at the full-appropriation upper corner is screening governed by the general condition in Appendix B.

Base Productivity Share. The first term represents the baseline value of the match attributable to the focal worker: the maximum wage a firm would pay if it evaluated the worker in a vacuum. Because the firm takes matching rates as given, it evaluates the worker by their demographic frequency in the applicant pool: if a worker constitutes a given fraction of the applicant pool, their baseline claim is that fraction of the surplus. This demographic valuation is the private counterpart of the planner's technological valuation, which credits the worker with the marginal contribution of applications to match formation, the generalized Hosios share $f_{xy}\varepsilon_{m,c}$ (Proposition 4). The gap between the two is the first wedge of Proposition 5: the firm pays workers for their prevalence in the pool, not for their contribution to match creation.

Screening Cost Premium. The second term is a wage adjustment reflecting the firm's relative informational burden, proportional to the firm's screening cost θ_y and inversely proportional to the applicant pool size δ_y and congestion ϕ_{xy}. Because $q_y(x)$ is an intensity that averages to 1 across the pool, a heavily targeted worker ($q_y(x) > 1$) is systematically easier for the firm to identify relative to the baseline applicant, and the firm passes a portion of these information-cost savings to the worker as a premium; an avoided worker ($q_y(x) < 1$) is penalized. The myopic firm evaluates this premium against the worker's raw demographic share, $1 - a_{xy}$, whereas the planner benchmarks

it against the relative targeting intensity, $1 - 1/q_y(x)$ (Proposition 4); the difference between the two reference points is the second wedge of Proposition 5. In assortative equilibria where the pool is homogeneous, the firm does not need to selectively screen ($q_y(x) \to 1$) and the premium vanishes.

Monopsony Markdown. The final term, $-1/z_{xy}$, is the rent-extraction discount stemming from the firm's first-mover advantage over an upward-sloping labor supply curve. Our wage equation recovers the structure of the canonical inverse-elasticity rule of monopsonistic competition (e.g., Manning, 2003), $\omega = MRP - (\partial \ln L_y(\omega)/\partial\omega)^{-1}$: the first two terms constitute the effective opportunity cost facing the firm, and $1/z_{xy}$ is the inverse semi-elasticity of the firm's labor supply. But where standard models assume this elasticity arises from exogenous preference heterogeneity (e.g., Type I Extreme Value taste shocks), our framework microfounds it. The elasticity is endogenously determined by the information friction θ_x, the meeting rate $q_y(x)\phi_{xy}$, and the strategic pool composition, as in equation (2.2). As search costs rise, worker targeting becomes less precise, labor supply becomes less elastic, and the markdown increases.

The markdown can also persist even as search costs vanish ($\theta_x \to 0$) in assortative equilibria. When workers target a single firm type, the own-application share $\bar{p}_{xy}$ approaches one, and the captivity factor $(1 - \bar{p}_{xy})$ in equation (2.2) drives the elasticity down exponentially fast. Economically, in an assortative equilibrium firms face a captive audience. With no effective wage competition because workers are cornered into a specific match, the firm can extract nearly the entire surplus via the markdown. This underscores that monopsony power is not a function of aggregate labor supply elasticity alone, but is fundamentally shaped by the equilibrium sorting pattern.

3.2 Unifying the Monopsony Literature and Welfare

Collecting these results, the model isolates four distinct structural sources of wage-setting power, each tied to an observable primitive:

1. **Worker search frictions (θ_x).** Information costs limit the resolution of worker targeting and hence the labor supply elasticity, feeding the monopsony markdown $1/z_{xy}$. The markdown can remain large even as $\theta_x \to 0$ when sorting makes workers captive.

2. **Firm screening costs (θ_y).** Entropy costs of screening generate the screening cost premium: firms pass information-cost savings to easily identified workers and penalize those they must search harder to find, tying posted wages to the composition of the applicant pool rather than to surplus alone.

3. **Equilibrium sorting (a_{xy}).** The composition of the pool determines both the base productivity share and the size of the markdown. Full sorting ($a_{xy} \to 1$) eliminates the screening premium while maximizing the markdown, making assortative matching a force for wage compression rather than wage competition.

4. **Sequential wage setting.** By pre-committing to wage menus before workers apply, firms extract rents from the workers' reactions; the first-mover advantage converts the elasticity objects into pricing power.

Market concentration is not absent from the model; it operates through the market's relative thickness rather than a fixed number of competitors. The elasticity zxy scales with the number of firms a worker can reach (δ_x) and with pool size and congestion ($\delta_y, \phi xy$), all governed by the relative masses of workers and firms. A thinner firm side lowers z_{xy} and raises the markdown, exactly as employer concentration does in oligopsony - here endogenously, through search tightness.

The captivity factor $(1 - \bar{p}_{xy})$ within the application elasticity z_{xy} functions as the algebraic engine uniting several disparate branches of the modern monopsony literature. It mirrors the mechanisms driving market power in granular oligopsony models (Berger, Herkenhoff, and Mongey, 2022), Roy-model comparative advantage frameworks (Bils, Kaymak, and Wu, 2025), and finite-market power in directed search (Burdett, Shi, and Wright, 2001). In oligopsony models, labor supply elasticity depends on the firm's payroll share relative to the market. In comparative advantage models, the markdown depends on the population distribution of second-best productivities, which shrinks the worker's viable outside options. Our model generates this concentration endogenously through the application-portfolio share $\bar{p}_{xy}$. We do not need to assume finite firms or copula-distributed comparative advantage; information frictions and choice-set limits naturally generate identical pricing power.

This formulation also makes precise which distortions survive the frictionless limit. If information costs vanish ($\theta_x \to 0$) while applications remain interior, the elasticity

explodes ($z_{xy} \to \infty$), the monopsony markdown approaches zero, and the wage converges to the demographic share $f_{xy}a_{xy}$ plus the vanishing screening premium. Even then, competitive pricing does not coincide with the planner's: the demographic base share replaces the Hosios share $\varepsilon_{m,c}$, so the first wedge of Proposition 5 survives frictionlessness. And along assortative branches, where workers become captive ($\bar{p}_{xy} \to 1$), the markdown itself survives the limit and wages collapse toward the outside option rather than converge to a competitive split.

For any positive information costs, however, the $1/z_{xy}$ 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. 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, for a given allocation of match volume, the resulting allocative inefficiency arises from distortions to the *relative* payoffs across different firms. While the absolute magnitude of the monopsony markdown determines the aggregate distribution of surplus, the *within-submarket* component of welfare loss is driven by the net wage distortions

$$\Delta\omega_{xy} = f_{xy}\left(\varepsilon_{m,c} - a_{xy}\right) + \frac{\theta_y/\delta_y}{\phi_{xy}}\left(a_{xy} - \frac{1}{q_y^*(x)}\right) + \frac{1}{z_{xy}}$$

across potential matches, holding the allocation of match volume fixed. It is this relative variance that distorts the price signals guiding worker sorting. A second, quantitatively larger channel, the reallocation of search across submarkets and the resulting change in total match volume, operates on the extensive margin and is quantified in Section 4.

Quantifying the Welfare Loss. We can directly quantify the aggregate welfare destroyed by the competitive market's strategic pricing. Because the social planner's optimal allocations can be decentralized by forcing the market to adopt the shadow wage ω^{PO}, we map the aggregate welfare loss directly to the submarket-specific wage distortions, defined as the gap $\Delta\omega_{xy} = \omega_{xy}^{PO} - \omega_{xy}^{CE}$.

Proposition 6. *Let $\mathcal{W}(\omega)$ denote the aggregate indirect welfare evaluated at a mandated wage vector ω. Using a second-order Taylor expansion around the socially optimal wage*

vector ω^{PO}, the aggregate welfare loss in the competitive equilibrium is proportional to the sum of the squared wage distortions, scaled by the physical match volume and the planner's normative labor supply elasticity:

$$\mathcal{W}(\omega^{PO}) - \mathcal{W}(\omega^{CE}) \approx \frac{1}{2} \sum_{x=1}^{W} \sum_{y=1}^{F} M_{xy} \hat{z}_{xy}^{PO} (\Delta\omega_{xy})^2,$$

where $M_{xy} = \mu_x \mu_y p_x(y) q_y(x) \phi_{xy}$ is the realized mass of matches, and $\hat{z}_{xy}^{PO} = \frac{z_{xy}}{1 - \omega_{xy}^{PO} z_{xy} \varepsilon_{\phi,p}}$ is the true physical application semi-elasticity, which incorporates the congestion feedback that the planner internalizes but atomistic agents ignore.

Proof. See Appendix E. The shadow wage ω^{PO} decentralizes the planner's optimum, so indirect welfare $\mathcal{W}$ is maximized there and its gradient vanishes by the Envelope Theorem. The loss is then a second-order curvature term; with many firms the MNL cross-substitution terms vanish and the Hessian is diagonal. Expressing the curvature through the inverse labor supply curve implied by the worker's optimality condition turns each diagonal element into $-M_{xy}\hat{z}_{xy}^{PO}$, which yields the Harberger form. □

This derivation links our strategic pricing distortions directly to classical public finance and empirical welfare metrics. It represents the information-frictional equivalent of a Harberger (1954) triangle. The welfare loss in any given submarket is half the square of the wage distortion ($\Delta\omega_{xy}$), weighted by the physical size of the submarket (M_{xy}) and the normative sensitivity of workers to that distortion ($\hat{z}_{xy}^{PO}$). Notice that the relevant elasticity is purely physical: it incorporates the true congestion feedback ($\varepsilon_{\phi,p}$) that the atomistic agents of the private market ignore. This formulation highlights why the firm's strategic pricing is so pernicious: the wage distortions are largest in the submarkets where monopsony power runs deepest, causing large relative misallocations of aggregate worker search effort.

Mapping Empirical Elasticities to Welfare Loss. Applying the welfare loss formula requires carefully mapping standard empirical estimates of labor supply to the normative semi-elasticity $\hat{z}_{xy}^{PO}$. Empirical studies using modern platform data (e.g., Marinescu and Wolthoff, 2020) typically estimate the log-log elasticity of applications with respect to the posted wage, $e_p \equiv \frac{\partial \ln p_x(y)}{\partial \ln \omega_{xy}}$, which captures the realized semi-elasticity at the competitive wage: $\hat{z}_{xy}^{data} = e_p / \omega_{xy}^{CE}$. This object cannot be substituted directly into

the welfare formula for two reasons: realized application flows embed the physical congestion feedback that atomistic agents do not internalize, and they are measured at the depressed competitive wage rather than at the shadow wage. Inverting the congestion feedback at the competitive wage to recover the baseline elasticity z_{xy}, and re-applying the feedback at the shadow wage (both steps are spelled out in Appendix E), collapses into a single closed-form bridge:

$$\hat{z}_{xy}^{PO} = \frac{\hat{z}_{xy}^{data}}{1 - \hat{z}_{xy}^{data} \varepsilon_{\phi,p} \left(\omega_{xy}^{PO} - \omega_{xy}^{CE} \right)}. \tag{3.2}$$

This mapping highlights a crucial adjustment for empirical welfare analysis in frictional markets. Because congestion elasticities are negative and the shadow wage exceeds the competitive wage in the submarkets where monopsony bites, the normative elasticity lies below the raw empirical estimate: as mandated wages rise toward the efficient benchmark, the additional applications they attract physically crowd the submarket and damp the marginal response. A naive Harberger calculation that carries the estimated elasticity unadjusted to the counterfactual therefore overstates the local allocative loss, and with it the gain from marginal wage corrections. The mapping supplies the discipline to evaluate interventions at the correct counterfactual point using only reduced-form inputs.

The Amenity Duality. The framework also clarifies the observational equivalence between information frictions and unobserved amenities. By the Matějka-McKay (2015) duality, entropic information costs (θ_x) and amenity taste shocks (Type-I extreme value) produce mathematically identical multinomial logit choice structures. When observing application flows alone, a model with compensating differentials and a model with information frictions are indistinguishable. However, their implications diverge on welfare and pass-through. Amenity dispersion represents consumed utility, meaning flat wages might reflect efficient compensating differentials. Conversely, information costs represent a dissipated resource loss, implying severe misallocation. Analyzing these mechanisms within a single structural framework provides the tools to parse this duality.

3.3 Nesting Search Models and the Two Regimes of Sorting

The cost parameters (θ_x, θ_y) determine the fundamental nature of the equilibrium, spanning the friction environments of the search literature as limiting cases.

In the high-cost limit ($\theta_x \to \infty$), the entropic penalty dominates and application strategies converge to the uniform prior. This reproduces the random matching environments of Diamond (1971) and Burdett and Mortensen (1998), with wage determination governed by the logic of static monopsonistic competition (Manning, 2003; Berger, Herkenhoff, and Mongey, 2022). Workers are captive not because of the time it takes to receive another offer, but because the cognitive cost of identifying a better one is too high. Unlike Burdett and Mortensen (1998), where wage dispersion among identical firms arises from mixed strategies in a dynamic game, our model predicts that agents of the same type behave identically, generating wage dispersion through the interaction of heterogeneity and screening costs. The model also captures the fundamental tension of the random-search tradition between the Diamond paradox and Bertrand competition, mediated by labor market tightness: when workers are abundant relative to firms, rising θ collapses wages to the outside option; when firms are abundant, competition for scarce labor pushes wages toward marginal product.

In the frictionless limit ($\theta \to 0$), the model approaches the directed search benchmarks. With identical agents and an increasing-returns matching technology, it replicates Burdett, Shi, and Wright (2001), including their equilibrium multiplicity. With heterogeneous firms and constant returns, the mixing equilibrium reproduces the defining features of competitive search (Moen, 1997): high-productivity firms attract larger applicant pools and submarkets clear through queue lengths. Because our firms price myopically rather than internalizing the congestion their postings create, the wage split retains the demographic-valuation wedge and does not converge to the Hosios rule.[7] Moreover, assortative captive-audience branches persist alongside the competitive-search-like equilibria; these branches are characterized by extreme monopsony power, an overlooked feature of the directed search paradigm that foreshadows our empirical findings.

[7]These surviving wedges are distortions to the wage, not necessarily to the allocation. In the limit, wages are transfers and information costs vanish, so limit welfare depends only on which matches form: monopsony becomes purely distributional. The competitive allocation attains the planner's whenever the market's log-supermodular ranking of matches coincides with the planner's level-supermodular one, which holds for standard surplus structures such as comparative advantage and vertical complementarity.

At moderate information costs, this multiplicity is resolved. Large enough θ guarantees a unique, smooth equilibrium, acting as a purification device for the equilibrium that otherwise spans the continuum between frictionless directed search and uniform random search. This uniqueness makes estimation feasible.

Within this continuous space, the interaction of the wage-equation forces reveals that sorting can be generated by two distinct mechanisms with opposite implications for wage inequality. The relative magnitude of information frictions (θ_x versus θ_y) and their interaction with the outside option floor (see Section 4.1) dictate the sorting regime. This distinction is central to resolving the empirical puzzle of flat wage profiles in sorted markets.

Competitive (Worker-Driven) Sorting. The worker-driven regime arises in environments with low information costs (θ_x) or weak monopsony power. Because workers can target efficiently, the monopsony markdown remains small. 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, strong sorting is always accompanied by high wage dispersion and high pass-through. This aligns with the intuition of standard directed search models, where wages play the primary allocative role. However, this regime struggles to rationalize modern employer-employee matched data, which feature flat wage profiles and small firm fixed effects.

Monopsony (Firm-Driven) Sorting. A different regime emerges under asymmetric frictions, where worker search costs are high but firm screening costs are low ($\theta_x \gg \theta_y$). High θ_x suppresses the application elasticity (z_{xy}), granting firms massive monopsony power. The markdown aggressively crushes wages down against the censoring floor defined by the worker's outside option .

Because wages are pinned to the outside option of the worker across the market, the wage gradient across heterogeneous firms becomes flat, weakening the price signal. Workers, observing minimal wage differences, have little incentive to exert search effort, and their applications (p_x) become diffuse and undirected. However, sorting does not disappear. It is rescued and enforced by the firms. If productive complementarities are strong, firms find it profitable to exert effort to cherry-pick the best matches from their

relatively undirected applicant pools using their low-cost screening technology (q_y).

This firm-driven regime decouples sorting from wage signals. It provides the mathematical configuration within the framework that allows strong positive assortative matching to coexist with flat wage profiles and small firm fixed effects. Section 4 will demonstrate that the Swedish labor market seems to operate in this regime.

4 Resolving the Monopsony Puzzles

Matched employer-employee data presents a constellation of stylized facts that are difficult to reconcile within a single theoretical framework. A successful structural model of the labor market must jointly rationalize three robust empirical regularities that appear contradictory.

The first is the sorting puzzle. 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, requiring high-productivity firms to post steep wage gradients to attract high-skill applicants.

However, the same data reveals that wage profiles are surprisingly flat with respect to firm productivity. While wages increase as workers move to better firms, the gradient is shallow. The pass-through of firm productivity into wages is low, meaning workers capture a small fraction of the additional surplus generated by moving up the job ladder. This creates a tension: if the wage premium for moving to a top firm is small, why do high-skill workers sort there so effectively? The coexistence of strong sorting and weak price signals is difficult to explain.

The second is the AKM puzzle, which highlights a conflict regarding the source of wage variation. The literature following Abowd, Kramarz, and Margolis (1999) consistently finds that firm fixed effects explain a small share of the variance in wages, with the majority explained by worker fixed effects. This suggests 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 firms possess signif-

icant market power. The puzzle is how firms can simultaneously exercise substantial monopsony power, setting wages below marginal product, while generating wage data dominated by worker characteristics rather than firm identity.

The third is the elasticity puzzle. Quasi-experimental and platform evidence finds that job applications respond strongly to posted wages, with firm-level labor supply elasticities estimated around three (e.g., Manning, 2003; Azar, Marinescu, and Steinbaum, 2019; Marinescu and Wolthoff, 2020). Taken through the lens of the textbook inverse-elasticity rule, a single aggregate elasticity rigidly ties the markdown to one number everywhere. Yet the evidence above suggests wages sit near workers' outside options, with firms capturing nearly all of the match surplus. This degree of wage-setting power goes far beyond what a mechanical one-to-one mapping from an aggregate elasticity of three would suggest. How can highly wage-elastic search coexist with large markdowns?

Standard frameworks struggle to resolve these tensions simultaneously. Directed search models 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 generate monopsony power and 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 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.

4.1 Data, Estimation Strategy, and Model Fit

To investigate these puzzles quantitatively, we fit our model to moments from the Swedish labor market, using the dynamic moment estimates of Bonhomme, Lamadon, and Manresa (2019).

BLM characterize the labor market with 6 worker types and 10 firm classes. Because their model focuses on job-to-job transitions, 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 class $y_{prev} \in \{1, \ldots, 10\}$. This yields 60 distinct worker types searching for jobs at 10 distinct firm classes. The masses of searching workers μ_s and firms μ_y are calibrated directly from the steady-state distribution of workers across firms reported by BLM.

An important extension relative to the baseline model is that workers differ in their outside option based on their current employment. We denote the outside option of searcher type s by V_s. A match between permanent type x and firm class y produces output of value $V_s + f_{xy}$, where the productivity gain f_{xy} does not depend on the worker's previous employer. The surplus of a match over the worker's state-dependent reservation value is therefore the gain f_{xy}. All wage-related quantities (V_s, f_{xy}, wages, markdowns) are expressed in log-wage units.

Unrestricted Parameters. We estimate the deep parameters of the model. We impose no structural assumptions on the productivities or the outside options: the productivity gains f_{xy} form an unrestricted 6×10 matrix, and the outside options V_s form an unrestricted 60-vector. We assume a flexible, type-pair-specific matching technology, specified as Cobb-Douglas with constant returns, $M(c, d) = \psi_{xy}\, c^{\alpha_{xy}} d^{1-\alpha_{xy}}$, and estimate the matrices of matching efficiencies ψ_{xy} and elasticities α_{xy} (6×10 each). Finally, we estimate the vectors of information costs: worker search costs θ_x by permanent type (6 parameters) and firm screening costs θ_y by firm class (10 parameters). Two scale normalizations on the weighted means of θ_x and of ψ_{xy} pin down units that are not separately identified from the level of congestion. This leaves 254 estimated parameters.

Target Moments and Implementation. We estimate the model via the Generalized Method of Moments (GMM), targeting 735 moments derived from the BLM estimates and complementary aggregate evidence:

- **600 transition flows:** the number of movers from each of the 60 searcher states to each of the 10 destination firm classes, capturing detailed sorting patterns in worker flows (transformed by asinh to moderate the influence of the largest cells);
- **60 inflow composition shares:** the worker-type composition of each firm class's inflows (also asinh-transformed);

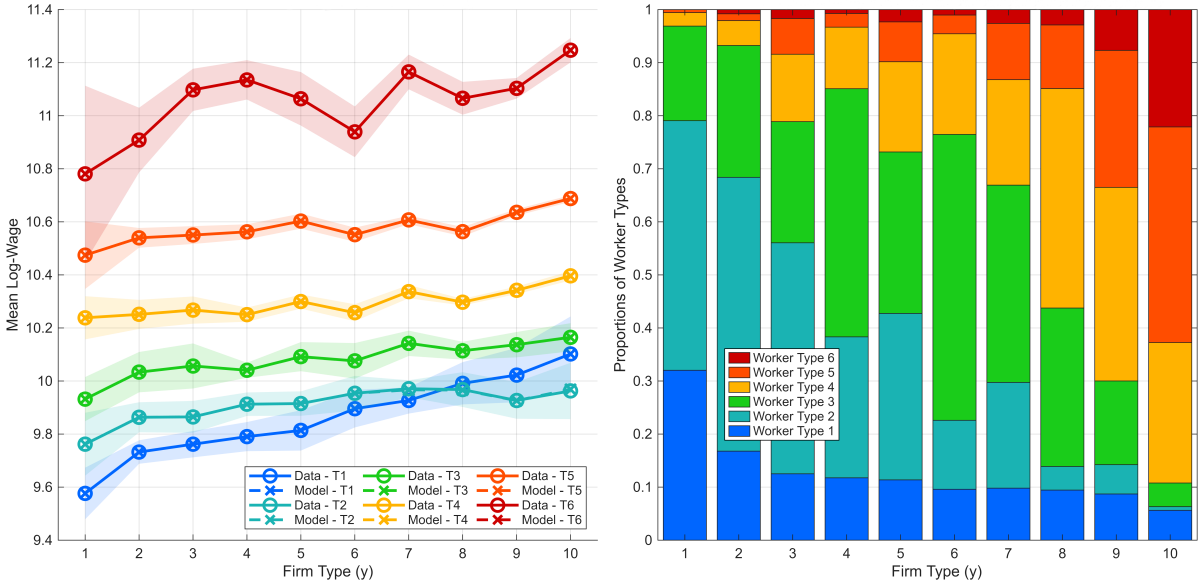
Table 1: Model Fit by Moment Block (baseline estimation)

Moment block	# moments	RMS deviation	Notes
Transition flows	600	2.46	asinh counts
Mean wages	60	0.0007	normalized units
Inflow composition	60	0.0065	asinh shares
AKM decomposition	3	0.23	shares and correlation
Labor supply elasticity	1	0.28	model 3.3 vs. target 3.0
Variance of productivity	1	0.0005	model vs. target 0.344
Markdowns by firm class	10	0.29	log points

Notes: Root-mean-squared deviation between model and data moments within each block, in the natural units of each block. Wages, composition, and the productivity variance are matched almost exactly; the elasticity and the markdown structure are matched in magnitude; the AKM deviation is concentrated in the lightly weighted worker-effect share (the firm-effect share and the sorting correlation are matched exactly).

- **60 mean wages:** the mean log-wage for each cell (worker type × destination class), capturing the gradient of wages with respect to firm productivity;
- **3 AKM moments:** the share of wage variance attributed to firm fixed effects (4.2%), the share attributed to worker fixed effects (60.2%), and the correlation between worker and firm fixed effects (0.41), derived from the AKM decomposition of the same Swedish data;
- **1 aggregate labor supply elasticity:** the match-weighted firm-level labor supply elasticity, targeted at 3.0, representing the consensus magnitude in the empirical monopsony literature (Manning, 2003; Azar, Marinescu, and Steinbaum, 2019);
- **1 variance of productivity:** the cross-match variance of log productivity (0.344), constructed from Swedish value-added data adjusted for tax wedges and aggregation corrections;
- **10 wage markdowns:** class-specific log markdowns, defined as the gap between adjusted value added per worker and adjusted gross earnings, averaging approximately 0.2 log points across firm classes.

Figure 4.1: Matched BLM Moments and Model Fit.



These moments discipline the objects at the heart of the monopsony puzzles. The model must be simultaneously consistent with elastic application behavior, large average markdowns, and the AKM variance decomposition. The GMM objective uses a diagonal weighting matrix that places high weight on the wage, composition, productivity-variance, and the first two AKM blocks relative to the fine-grained flow cells. Because the objective is high-dimensional and multimodal, we minimize it with a large-scale multi-start global optimization routine to ensure robust convergence.[8]

Model Fit. Despite being heavily over-identified (254 parameters against 735 moments), the estimated model achieves a strong overall fit to the data. Table 1 summarizes the fit by moment block. The model reproduces the detailed wage profiles, the inflow compositions, and the productivity variance very closely. It captures the broad shape of the 600-cell transition matrix and delivers an aggregate labor supply elasticity and a markdown structure of the right magnitude.

Figure 4.1 illustrates the fit to the core BLM facts. The left panel plots the mean

[8]We use a multi-start TikTak global optimizer (several hundred Sobol-sequenced restarts feeding a local quasi-Newton polish), retaining the best-fitting economically coherent optimum. The diagonal GMM weighting matrix places high weight on the 60 mean-wage, 60 inflow-composition, productivity-variance, and the first two AKM blocks, and low weight on the 600 fine-grained flow cells; an economic-coherence screen discards optima with degenerate corner-solution wage or matching configurations.

log-wages for each of the 6 worker types across the 10 firm classes. The model (dashed lines) tracks the data (solid lines) with high fidelity. While wages generally rise with firm class, the wage profiles are flat relative to the estimated productivities, indicating substantial wage compression. The right panels compare the worker-type composition of each firm class's hires in the data and in the model. High-wage workers are disproportionately employed by, and move to, high-wage firm classes, demonstrating strong positive assortative matching. The model replicates these flow compositions closely.

This combination of flat wage profiles alongside strong sorting is the central empirical puzzle. The model's ability to fit these moments simultaneously, while also matching the AKM firm-effect share (4.2%) and sorting correlation (0.41), 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.

4.2 Extreme Monopsony and the AKM Resolution

The estimated structural parameters, displayed in Figure 4.2, reveal the mechanisms driving these aggregate outcomes. While there is significant heterogeneity in productivity, matching efficiency, and information costs across the market, the cost estimates exhibit a pronounced asymmetry. The estimated firm screening costs (θ_y) are, on average, an order of magnitude smaller than the worker search costs (θ_x). This disparity places virtually all the informational rent-extraction capacity on the firm side.

The implication of this asymmetry becomes visible when analyzing how the match surplus is split. The bottom-right panel of Figure 4.2 shows that the estimated outside options V_s, although unrestricted in the estimation, align closely with the current wages of the workers. Figure 4.3 details this surplus decomposition across all submarkets.[9] For the vast majority of matches, the equilibrium wage series lies directly on top of the outside option V_s, far below the estimated match productivity. Firms exercise severe monopsony power, capturing almost the entire surplus as profit.

This pricing behavior resolves the AKM puzzle. Wages are determined primarily

[9]We compute standard errors from the GMM sandwich covariance, with the moment Jacobian obtained by two-sided numerical differentiation and a Moore-Penrose pseudo-inverse to accommodate the rank-deficiency induced by cells at the wage or screening bounds. The error bands in the figures are a parametric bootstrap: we draw 1,000 parameter vectors from the estimated asymptotic distribution $\mathcal{N}(\hat{\theta}, \hat{\Sigma})$, re-solve the equilibrium for each draw, and report percentile intervals of the recomputed object.

Figure 4.2: Estimated Structural Parameters.

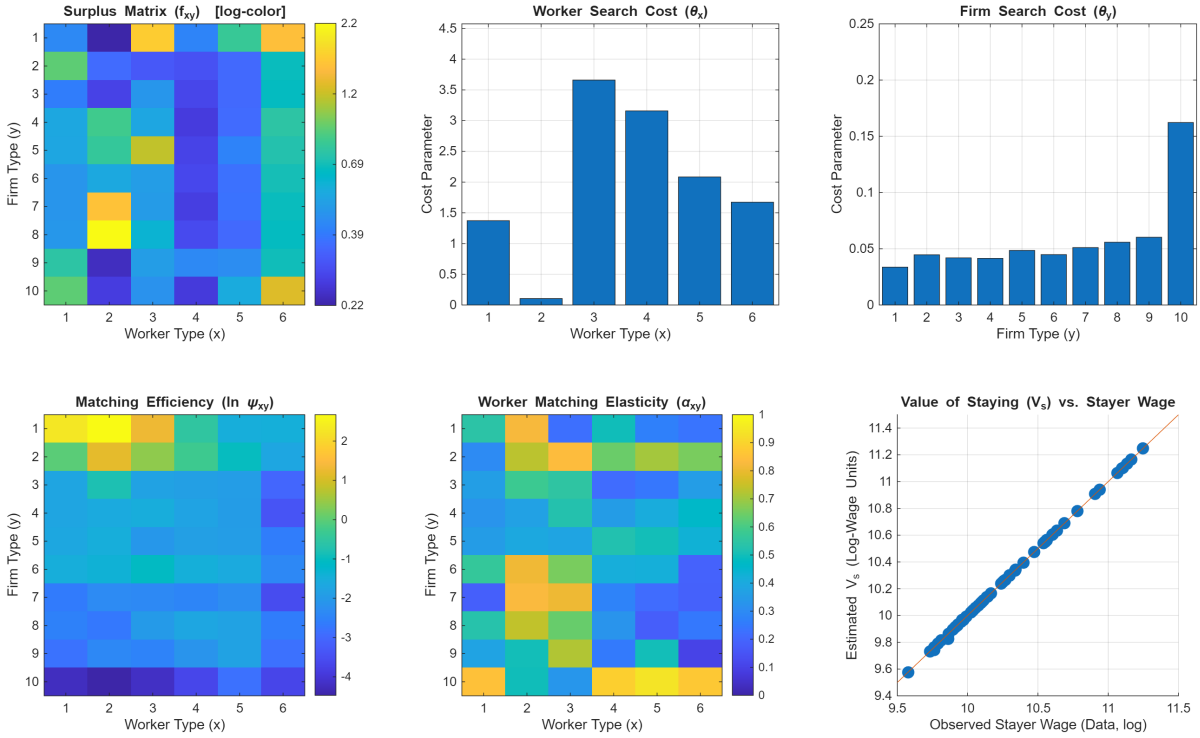
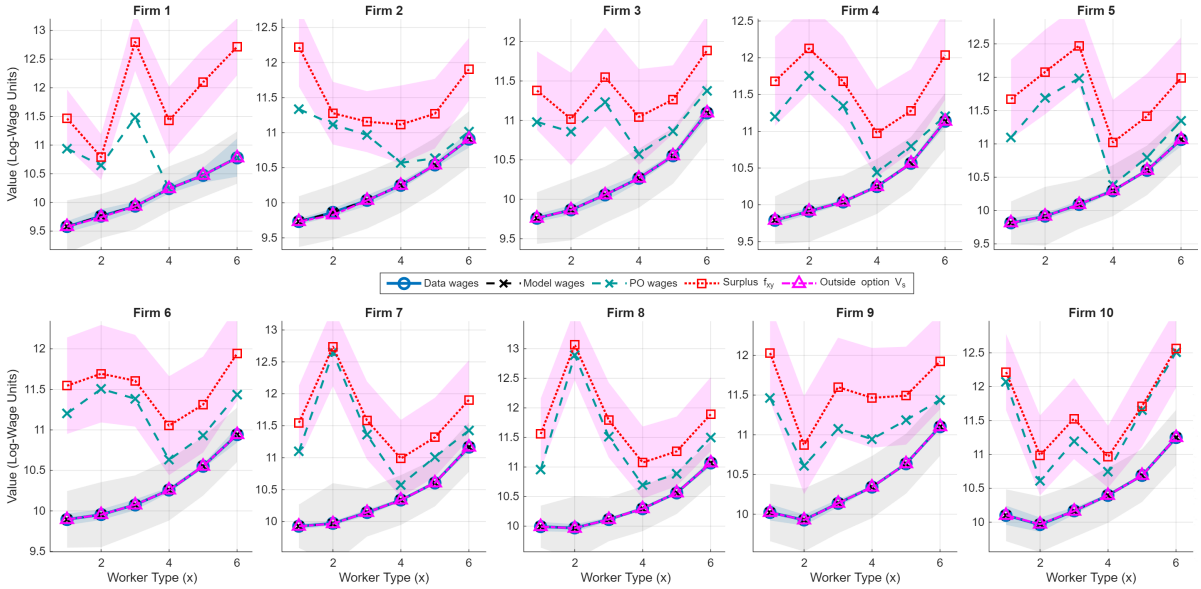


Figure 4.3: Firm Payoffs and Surplus Decomposition.



by the worker's history, embedded in their state-dependent outside option V_s, rather than the productivity of the new match. Worker fixed effects dominate the variance decomposition because wages track the worker's history. Conversely, firm fixed effects are small because a high-productivity firm does not pay significantly more than a low-productivity firm; it captures the additional productivity as profit. The model reproduces the core of the data's AKM configuration (a 4.2% firm-effect share combined with a 0.41 sorting correlation) as an equilibrium outcome of firm wage-setting under information frictions. While the model attributes a larger portion of wage variation to history-dependent outside options than a pure worker fixed effect captures, it successfully suppresses the firm identity as a driver of wage levels.

A direct consequence of this monopsony power is a strong disconnect between wages and productivity. Figure 4.4 illustrates this relationship: the correlation between estimated match productivity and equilibrium wages is weakly positive (0.16). Because high-productivity firms effectively capture their applicant pools, they exert stronger monopsony power, flattening the wage gradient. Wages cease to function as a reliable signal of productivity. This finding provides support for the hypothesis of Eeckhout and Kircher (2011) that wage data alone may be insufficient, or even misleading, for identifying sorting patterns in the presence of frictions.

4.3 Firm-Driven Sorting and the Elasticity Puzzle

The severe monopsony power documented above results in large markdowns. Because markdowns drive wages to the reservation floor across the market, the flat wage profiles destroy the signal value of pay, meaning wages cease to function as a reliable indicator of underlying firm productivity. If wages offer little signal of productivity, how does the market achieve strong assortative matching?

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, and the pie charts within them indicate the surplus split, highlighting that the vast majority is captured by firms. The lines connecting agents to submarkets illustrate the search intensities. Worker strategies (bottom) are diffuse, reflecting the weak wage incentives. Firm strategies (top), however, are highly targeted. The thickest screening lines emanate from firms targeting

Figure 4.4: The Productivity-Wage Disconnect.

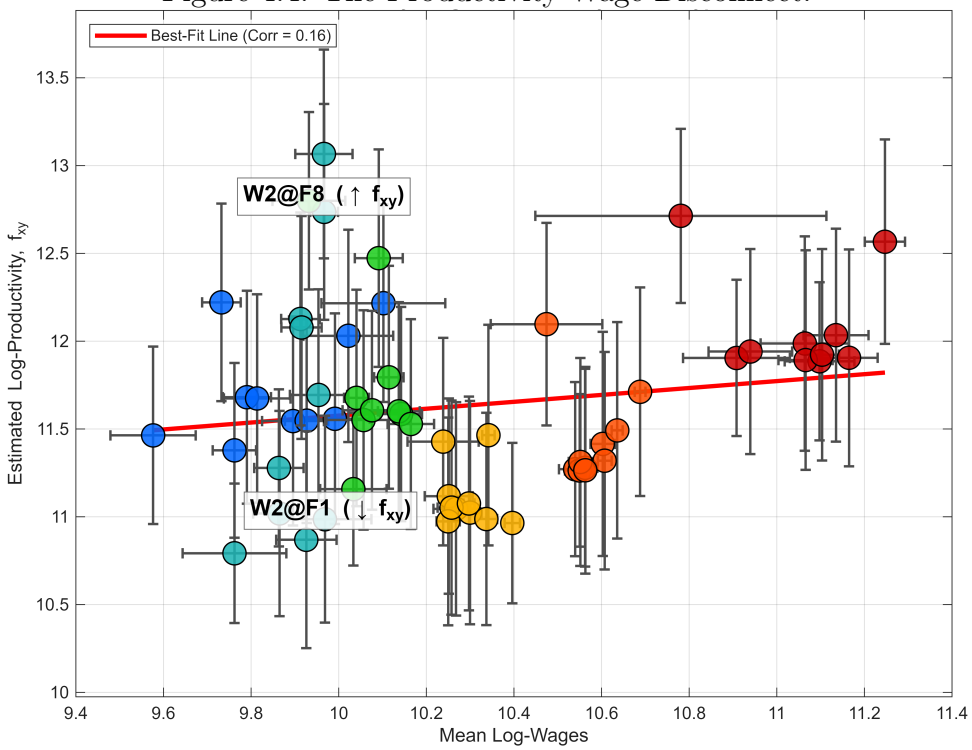
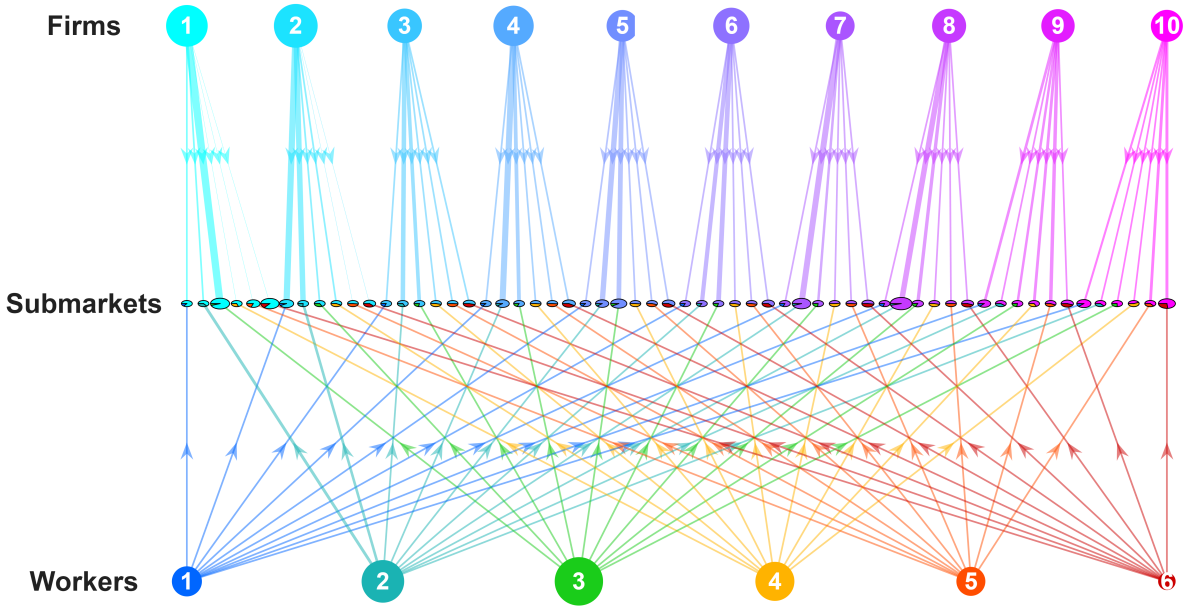


Figure 4.5: Equilibrium Search and Screening Strategies.



high-surplus matches. Because firms capture nearly the entire surplus, they possess a powerful incentive to direct their screening capacity toward these pairings. The market operates in a firm-driven sorting regime: the strong sorting observed in the data is driven by firms' screening efforts q_y, which direct the flow of matches despite the lack of targeting from workers.

This firm-driven sorting mechanism provides the key to resolving the elasticity puzzle. The estimated model yields an aggregate labor supply elasticity of 3.3, aligning closely with the consensus target of 3.0 in the empirical literature. In the textbook inverse-elasticity rule, an aggregate elasticity of three bounds the markdown to a much lower level, which seemingly contradicts our finding of pervasive monopsony where wages sit at workers' outside options. Indeed, the estimated model generates an average markdown of approximately 0.46 log points across matches, similar to the data, even as the aggregate elasticity sits at 3.3. Reconciling highly wage-elastic search behavior with substantial firm wage-setting power requires examining how these matches are distributed.

The tension is resolved by recognizing that aggregate empirical elasticities inherently reflect the submarkets where economic activity is concentrated. Consider the structural form of the application semi-elasticity, which is proportional to the expected meeting

rate over the information cost ($e_{xy} \propto q_y(x)\phi_{xy}/\theta_x$). Because the estimated worker search costs are large on average ($\theta_x > 1$), the unweighted baseline elasticity across the market is low. For the vast majority of potential submarkets, workers are unresponsive, granting firms the pervasive monopsony power required to push base wages down to the outside-option floor.

However, because firm screening costs are negligible ($\theta_y \ll \theta_x$), firms actively elevate the screening intensity for high-value targets. While non-targeted workers face a lower baseline screening rate, targeted workers become upward outliers ($q_y(x) \gg 1$). In these targeted submarkets, the high screening intensity acts as a multiplier, overwhelming the high worker search cost and causing the local application elasticity to spike.

When researchers estimate labor supply elasticities using cross-sectional regressions, instrumental variables, or difference-in-differences designs, the identifying variation relies on realized matches or active recruitment channels. These methods systematically capture the elevated elasticities of the targeted submarkets where matches actually form. Even application-level data will exhibit high responsiveness because researchers observe robust application flows in the very submarkets where firms are actively screening and matching. This represents a profound structural feature of the labor market: high measured elasticity reflects the responsiveness of workers in the specific submarkets where firms have elevated the meeting rate. The broader, highly inelastic baseline of the market grants firms their monopsony power, while the targeted, high-elasticity peaks drive the aggregate empirical estimates.

4.4 Misallocation and Aggregate Welfare

The decoupling of wages from productivity has substantial welfare implications. The inefficiency is not merely distributional; the monopsony distortion actively prevents the formation of matches, artificially starving the economy of employment.

In the competitive equilibrium, firms expend significant screening effort hunting for high-surplus matches. Because the monopsony markdown suppresses the wage signal, workers are not incentivized to direct their search, leaving their applications diffuse. Armed with a cheap screening technology ($\theta_y \ll \theta_x$), firms respond by becoming drastically over-selective. They use their screening intensity to cherry-pick only the most lucrative pairings, leaving a massive swath of potential matches unformed.

Table 2: Welfare gain from the planner's reallocation, decomposed.

Panel A. Levels

	Matches N	Net surplus/match g	Welfare $W = N \cdot g$
Equilibrium (CE)	45,373	0.606	27,496
Optimum (PO)	82,518	0.446	36,803
Change	+81.9%	−26.4%	+33.8%

Panel B. Decomposition of the welfare gain ΔW

Margin	Contribution	Share of CE welfare
Extensive $(\Delta N \cdot \bar{g})$: more matches	+19,538	+71.1%
Intensive $(\Delta g \cdot \bar{N})$: surplus per match	−10,231	−37.2%
Net gain (ΔW)	+9,307	+33.8%

Notes: Welfare $W = N \cdot g$, with N the number of matches and g the match-weighted net surplus per match $(f_{xy} - V_s)$, in log-wage units. The decomposition $\Delta W = \Delta N\,\bar{g} + \Delta g\,\bar{N}$ uses CE–PO midpoints $\bar{N}, \bar{g}$ and is exact (no interaction residual). The planner's gain is the net of a large extensive-margin expansion and a partly offsetting intensive-margin decline.

This misallocation becomes apparent when contrasting the equilibrium outcomes with the socially optimal allocation. To achieve the first-best allocation, the planner prescribes a radically different wage structure. The green line in Figure 4.3 plots the planner-prescribed shadow wage ω^{PO} required to induce efficient worker application strategies. In almost all submarkets, the equilibrium wage sits substantially below this efficient level. The planner utilizes higher wages to pull workers into the market and incentivize broader search.

Shifting from the competitive equilibrium to the planner's allocation increases total surplus by 33.8 percent (Table 2). Decomposing this aggregate gain reveals the mechanics of the inefficiency: the welfare gain is driven entirely by the extensive margin. The planner increases the total number of matches by 82 percent (from 45,373 to 82,518). Because the planner forms so many additional matches, the average net surplus per match actually falls by 26 percent, dropping from 0.606 to 0.446 log-wage units.

The competitive equilibrium leaves workers unmatched because firms hoard their screening capacity for the highest-value targets and refuse to pay the wage premiums necessary to clear the rest of the market. The planner corrects this by raising wages, trading a lower average net surplus per match for a massive expansion in total match

volume and aggregate surplus.

A Practical Toolkit for Policy Evaluation. Quantifying the aggregate, extensive-margin cost above requires solving the model in full. A distinct and more immediately answerable question can be addressed with reduced-form inputs alone: how large is the *local* allocative distortion that monopsony pricing creates in a given market, and how would a marginal intervention move it? The Harberger-style formula and elasticity mapping of Section 3 are exactly this toolkit; they let an analyst evaluate marginal, within-market interventions without re-estimating the structural model. The required inputs are modest.

To approximate this local distortion in a given labor market, an analyst requires an empirical estimate of the local labor supply elasticity (readily available from the quasi-experimental literature), the observed competitive wage, and a standard macroeconomic calibration, or estimate, of the matching function elasticity, which then implies the congestion elasticity. These inputs map the observed behavioral elasticity to the normative physical elasticity $\hat{z}_{xy}^{PO}$ using equation (3.2).

To approximate the efficient shadow wage (ω^{PO}), policymakers can proxy the physical match surplus (f_{xy}) using observed firm-level value-added or revenue per worker, net of the worker's outside option. Multiplying this surplus by a standard matching elasticity (e.g., 0.5) yields the Hosios base share. Comparing this theoretical benchmark to observed competitive wages provides an empirical estimate of the wage distortion ($\Delta\omega_{xy}$).

Feeding these components into the normative Harberger formula allows policymakers to evaluate specific interventions, such as minimum wages. The formula reveals that this local distortion is driven by the variance of relative wage distortions across submarkets, not by the absolute level of wages. A uniform minimum wage might inadvertently compress the incentive structure further. Because the competitive market is over-selective, an optimal policy must target relative prices to induce broader match formation.

A graduated minimum-wage schedule that varies by firm productivity or firm class could artificially restore the missing wage gradient. This intervention would compel high-productivity firms to pay the premiums necessary to direct worker search and expand match volume, effectively using policy to substitute for the absent competitive

price signals.

4.5 Robustness

The quantitative results above rely on the behavioral benchmark of Assumption 4 and on a specific construction of the generalized method of moments objective. To validate these structural choices, we re-estimated the full model under alternative behavioral assumptions and aggregation methods.

Internalization. The baseline model assumes that agents do not internalize congestion or the composition of applicant pools. Re-estimating a generalized version of the model in which firms optimally internalize the physical congestion their contracts create delivers comparable fit and identical qualitative economics, preserving the 0.41 AKM sorting correlation: internalization merely reweights base compensation between the worker's demographic share of the pool and their technological matching contribution, and the flexible, match-specific matching technology parameters absorb this reweighting. The data are largely uninformative about this margin, and our conclusions do not depend on it.

Allowing firms to internalize the composition of their applicant pools, by contrast, makes matters distinctly worse. Such firms aggressively raise the posted wages of applicants they rarely screen, because a larger and more diverse pool lowers their average entropic cost of screening; this anchors compensation to pool-average productivity and compresses wage dispersion across worker types within any given firm class. The Swedish data feature pronounced within-class wage dispersion, which conflicts with this behavior: the re-estimated fit deteriorates by a factor of two to six, the required monopsony markdown balloons, driving equilibrium wages into corner solutions, and the aggregate labor supply elasticity collapses to a range between 0.15 and 0.74, falling far below the empirical consensus of 3.0. The data thus firmly support the benchmark in which firms evaluate each hiring market in isolation.

Alternative Moment Aggregation and Identification. Finally, we confirm the robustness of the markdown structure to accounting conventions. The baseline objective weights wage moments using the model's endogenous match flows, which can entangle

the wage fit with the flow fit. Re-estimating the model using empirical flows to weight the wage-related moments decouples these blocks.

Under this alternative aggregation, the core narrative remains unchanged. The severe asymmetry in information costs, the firm-driven sorting regime, and the presence of substantial markdowns all persist, with the average markdown aligning even closer to the empirical targets. While individual cell-level parameters for matching efficiency and matching elasticity exhibit weak local identification, the aggregate objects driving our macroeconomic conclusions are robust across specifications.

5 Conclusion

This paper develops an information-based theory of the labor market that bridges the search and monopsony literatures. By modeling search and screening as probabilistic choices subject to entropic costs, the framework yields a closed-form wage equation that isolates the structural sources of wage-setting power. The resulting endogenous captivity factor provides a common algebraic foundation for modern monopsony, reproducing the localized pricing power of granular oligopsony or comparative advantage models without requiring assumptions about finite firm concentration. The framework adopts a transparent behavioral benchmark in which agents treat market congestion as given and firms evaluate each hiring submarket in isolation. Re-estimations that relax this benchmark confirm it: allowing firms to internalize congestion leaves the results unchanged, while allowing them to internalize the composition of their applicant pools sharply degrades the fit. This bounded behavior maintains realistic wage dispersion, offering a behavioral anchor for future models of wage setting.

Estimating this framework on Swedish administrative data reconciles the apparent tensions between strong assortative matching, flat wage profiles, and highly elastic application flows. The estimation identifies a firm-driven sorting regime characterized by asymmetric information costs. High worker search costs grant firms substantial monopsony power, compressing wages toward history-dependent outside options and leaving worker applications diffuse. Assortative matching is instead sustained by firms utilizing low-cost screening to select productive pairings. This dynamic highlights that sorting can decouple entirely from wage signals. For the broader literature on inequality and misallocation, this suggests that observed wage dispersion may be a poor proxy

for underlying productivity differences, and that aggregate labor supply elasticities can reflect compositional selection by screening firms rather than widespread worker mobility.

The macroeconomic consequence of this regime is a substantial extensive-margin welfare loss. Armed with high monopsony power and inexpensive screening, the competitive market becomes inefficiently over-selective. Firms hoard their screening capacity for the highest-surplus matches and suppress wages, leaving a large volume of potential employment unrealized. The social planner corrects this by prescribing higher wages to pull workers into the market, accepting a lower average net surplus per match in exchange for a broad expansion in total matches. To assist in evaluating such interventions, we provide a normative framework that maps reduced-form behavioral elasticities directly to allocative welfare losses. This translation demonstrates that targeted policies, such as graduated minimum wage schedules, could counteract the over-selective monopsony trap by forcing firms to pay the premiums necessary to expand aggregate match volume.

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] **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

[8] **Borovičková, K. and R. Shimer (2025)** "Assortative Matching and Wages: The Role of Selection." Working paper.

[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] **Cahuc, P., F. Postel-Vinay, and J.-M. Robin (2006)** "Wage Bargaining with On-the-Job Search: Theory and Evidence." *Econometrica*, 74(2), pp. 323-364. https://doi.org/10.1111/j.1468-0262.2006.00665.x

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

[13] **Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018)** "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics*, 36(S1), pp. S13-S70. https://doi.org/10.1086/694153

[14] **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

[15] **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

[16] **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

[17] **Cover, T. M. and J. A. Thomas (1991)**. "Elements of Information Theory." John Wiley & Sons, Inc.

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

[19] **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

[20] **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.

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

[22] **Eeckhout, J. and P. Kircher (2011).** "Identifying Sorting—In Theory." *Review of Economic Studies*, 78(3), pp. 872-906. https://doi.org/10.1093/restud/rdq034

[23] **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

[24] **Hagedorn, M., T. H. Law, and I. Manovskii (2017)** "Identifying Equilibrium Models of Labor Market Sorting." *Econometrica*, 85(1), pp. 29-65. https://doi.org/10.3982/ECTA11301

[25] **Hall, R. E., and A. I. Mueller (2018)** "Wage Dispersion and Search Behavior: The Importance of Nonwage Job Values." *Journal of Political Economy*, 126(4), pp. 1594-1637. https://doi.org/10.1086/697739

[26] **Harberger, A. C. (1954)** "Monopoly and Resource Allocation." *American Economic Review*, 44(2), pp. 77-87.

[27] **Hosios, A. J. (1990)** "On the Efficiency of Matching and Related Models of Search and Unemployment." *Review of Economic Studies*, 57(2), pp. 279-298. https://doi.org/10.2307/2297382

[28] **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

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

[30] **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

[31] **Lopes de Melo, R. (2018)** "Firm Wage Differentials and Labor Market Sorting: Reconciling Theory and Evidence." *Journal of Political Economy*, 126(1), pp. 313-346. https://doi.org/10.1086/695505

[32] **Marinescu, I. and R. Wolthoff (2020)**. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics*, 38 (2). https://doi.org/10.1086/705903

[33] **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

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

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

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

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

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

[39] **Postel-Vinay, F. and J.-M. Robin (2002)** "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity." *Econometrica*, 70(6), pp. 2295-2350. https://doi.org/10.1111/j.1468-0262.2002.00441.x

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

[41] **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.

[42] **Rosen, J. B. (1965)** "Existence and Uniqueness of Equilibrium Points for Concave N-Person Games." *Econometrica*, 33(3), pp. 520-534. https://doi.org/10.2307/1911749

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

[44] **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

[45] **Sims, C. A. (2003)** "Implications of Rational Inattention." *Journal of Monetary Economics*, 50(3), pp. 665-690. https://doi.org/10.1016/S0304-3932(03)00029-1

[46] **Sorkin, I. (2018)** "Ranking Firms Using Revealed Preference." *Quarterly Journal of Economics*, 133(3), pp. 1331-1393. https://doi.org/10.1093/qje/qjy001

[47] **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.

[48] **Taber, C. and R. Vejlin (2020)** "Estimation of a Roy/Search/Compensating Differential Model of the Labor Market." *Econometrica*, 88(3), pp. 1031-1069. https://doi.org/10.3982/ECTA14441

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

A Wage and Screening Elasticities of Labor Supply

This appendix derives the wage semi-elasticity z_{xy} used throughout the paper, and the screening elasticity of applications v_{xy} used in the proof of Proposition 2 (Appendix B).

The worker's optimality condition from Proposition 1 is

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

From the resource constraint captured by the Lagrange multiplier it follows that

$$d\left[\frac{\theta_x}{\delta_x}(1 + \ln p_x(y) + \lambda_x)\right] = \frac{\theta_x/\delta_x}{(1-\bar{p}_{xy})}\, d\ln p_x(y), \qquad \bar{p}_{xy} = \frac{\mu_y}{\delta_x}p_x(y).$$

This motivates the definition of the wage semi-elasticity of applications,

$$z_{xy} \equiv \frac{q_y(x)\phi_{xy}}{\theta_x/\delta_x}\,(1-\bar{p}_{xy})\,,$$

which is the response of applications to wages holding the matching rate ϕ_{xy} fixed.

The wage semi-elasticity. Totally differentiating the optimality condition with respect to ω_{xy}, holding $q_y(x)$ and, by Assumption 4, the congestion rate ϕ_{xy} fixed:

$$q_y(x)\phi_{xy}\, d\omega_{xy} = \frac{\theta_x/\delta_x}{(1-\bar{p}_{xy})}\, d\ln p_x,$$

so that

$$\frac{d\ln p_x(y)}{d\omega_{xy}} = z_{xy}.$$

Because agents do not internalize the dependence of ϕ_{xy} on applications, z_{xy} is both the worker's behavioral response and the application response the firm perceives when setting its contract in the first stage.

The screening elasticity of applications. Differentiating the worker's optimality condition with respect to $q_y(x)$, again holding ϕ_{xy} fixed:

$$\omega_{xy}\phi_{xy}\, dq_y(x) = \frac{\theta_x/\delta_x}{(1-\bar{p}_{xy})}\, d\ln p_x.$$

Recognizing that $\frac{dq_y(x)}{q_y(x)} = d\ln q_y(x)$ and substituting z_{xy} isolates the screening elasticity of applications used in the text:

$$v_{xy} \equiv \frac{d\ln p_x}{d\ln q_y} = \omega_{xy} z_{xy}.$$

Limits. When $\theta_x \to \infty$, $z_{xy} \to 0$ and workers apply randomly. When $\theta_x \to 0$, z_{xy} diverges and the firm's pricing power erodes, unless the own-application share approaches one ($\bar{p}_{xy} \to 1$): the captivity factor $(1-\bar{p}_{xy})$ then keeps the elasticity bounded, sustaining monopsony power even without information costs (Section 3).

B Proof of Proposition 2

With the elasticities of Appendix A in hand, we can compute the firm's first-order conditions. Recall that $\delta_y = \sum_x \mu_x p_x^*(y)$ and pool weights $a_{xy}^* = \mu_x p_x^*(y)/\delta_y$. By the quotient rule, the own-type derivatives are $\frac{\partial a_{xy}^*}{\partial \omega_{xy}} = a_{xy}(1-a_{xy})z_{xy}$ and $\frac{\partial a_{xy}^*}{\partial q_y(x)} = a_{xy}(1-a_{xy})\frac{v_{xy}}{q_y(x)}$. Consistent with the firm's submarket myopia (Section 2), the compositional spillovers onto other applicant types are not taken into account: the firm evaluates its contract for type x setting $\frac{\partial a_{x'y}^*}{\partial \omega_{xy}} = \frac{\partial a_{x'y}^*}{\partial q_y(x)} = 0$ for $x' \neq x$. The full objective function is:

$$Y_y = \sum_{x=1}^{W} \mu_x p_x^*(y)\, q_y(x)\, \phi_{xy}\, (f_{xy} - \omega_{xy}) - \theta_y \sum_{x=1}^{W} a_{xy}^* q_y(x)(\ln q_y(x) + \lambda_y) + \theta_y \lambda_y$$

Step 1: the screening FOC. Taking the derivative w.r.t. $q_y(x)$, applying the derivatives involving v_{xy}, and dividing by $\delta_y a_{xy}^* \phi_{xy}$ yields the screening FOC:

$$(f_{xy} - \omega_{xy})(1 + v_{xy}) = \frac{\theta_y/\delta_y}{\phi_{xy}} \left[(1 + v_{xy}(1-a_{xy}))(\ln q_y(x) + \lambda_y) + 1\right], \tag{B.1}$$

where $v_{xy} = \omega_{xy} z_{xy}$ is the screening elasticity of applications derived in Appendix A. Collecting the $\ln q_y(x) + \lambda_y$ terms on one side and solving:

$$\ln q_y(x) + \lambda_y = \frac{\frac{\delta_y}{\theta_y}\phi_{xy}(f_{xy} - \omega_{xy})(1 + v_{xy}) - 1}{1 + v_{xy}(1 - a_{xy})}, \tag{B.2}$$

which, combined with the constraint that intensities average to one ($\sum_x a^*_{xy} q_y(x) = 1$, pinning down λ_y), characterizes the screening strategy. Note that this derivation did not involve any assumptions about wage determination, and therefore holds regardless of the wage determination process, including wages set exogenously or constrained by some upper or lower bound.

Step 2: the wage FOC and the invariant shadow cost. Taking the derivative of the objective function w.r.t. ω_{xy} and dividing by $\delta_y a^*_{xy} \phi_{xy} q_y(x) z_{xy}$ yields the wage FOC:

$$\frac{\theta_y/\delta_y}{\phi_{xy}}(1 - a_{xy}) \left(\ln q_y(x) + \lambda_y\right) = (f_{xy} - \omega_{xy}) - \frac{1}{z_{xy}}. \tag{B.3}$$

To combine the two FOCs, rewrite the screening FOC (B.1) as

$$(f_{xy} - \omega_{xy}) + v_{xy} \left[(f_{xy} - \omega_{xy}) - \frac{\theta_y/\delta_y}{\phi_{xy}}(1 - a_{xy}) \left(\ln q_y(x) + \lambda_y\right)\right] = \frac{\theta_y/\delta_y}{\phi_{xy}} \left(\ln q_y(x) + 1 + \lambda_y\right).$$

By the wage FOC (B.3), the bracketed term equals $1/z_{xy}$, and $v_{xy}/z_{xy} = \omega_{xy}$, so the wage-dependent terms cancel:

$$(f_{xy} - \omega_{xy}) + \omega_{xy} = f_{xy} = \frac{\theta_y/\delta_y}{\phi_{xy}} \left(\ln q_y(x) + 1 + \lambda_y\right).$$

The shadow screening cost is therefore invariant to the wage schedule (as long as the wage FOC holds as an equality):

$$\ln q_y(x) + \lambda_y = \frac{\delta_y}{\theta_y} \phi_{xy} f_{xy} - 1, \tag{B.4}$$

and combining with the adding-up constraint yields the simple screening rule, equation (2.4).

Step 3: the closed-form target wage. Substituting the invariant shadow cost (B.4) into the wage FOC (B.3):

$$(f_{xy} - \omega_{xy}) - \frac{1}{z_{xy}} = \frac{\theta_y/\delta_y}{\phi_{xy}}(1 - a_{xy}) \left(\frac{\delta_y}{\theta_y} \phi_{xy} f_{xy} - 1\right) = (1 - a_{xy}) f_{xy} - (1 - a_{xy}) \frac{\theta_y/\delta_y}{\phi_{xy}}.$$

Solving for the wage yields

$$\omega_{xy} = f_{xy}\, a_{xy} + \frac{\theta_y/\delta_y}{\phi_{xy}}\,(1 - a_{xy}) - \frac{1}{z_{xy}},$$

and applying the non-negativity and feasibility constraints delivers the censored closed-form solution in Equation (2.3).

Censored wages. When the closed-form expression falls outside $[0, f_{xy}]$, the wage binds at a corner and the wage FOC (B.3) holds as an inequality, but the screening FOC (B.1) continues to hold with equality. At the lower corner ($\omega^*_{xy} = 0$) the worker's screening response vanishes along with the worker's gain, $v_{xy} = \omega_{xy} z_{xy} = 0$, and (B.1) reduces directly to the invariant condition (B.4), so the simple screening rule (2.4) continues to apply. At the upper corner ($\omega^*_{xy} = f_{xy}$) screening is characterized by the general condition (B.2) with $v_{xy} = f_{xy} z_{xy}$. ■

C Proof of Proposition 3

Existence. The existence of an equilibrium follows from Brouwer's Fixed Point Theorem. The strategy space for each agent is the product of their bounded search intensities and their wage menus $\omega_{xy} \in [0, f_{xy}]$. The total strategy space is convex and compact. By Assumption 2 (strictly convex costs), the best-response functions $p^*_x(\cdot)$ and $q^*_y(\cdot)$ are continuous. 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 Hessian block. To establish the global uniqueness of the equilibrium, a standard sufficient condition requires the Jacobian matrix of the system of first-order conditions to be globally negative definite (see e.g., Rosen, 1965). Because 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 local Hessian block for a single firm's choice of $\{q_y(x), \omega_{xy}\}$:

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

Because the firm's choice of wage ω_{xy} affects the physical matching environment exclusively through the worker's application response $p_x(y)$, the cross-derivatives of the matching objects follow from the chain rule. Using $z_{xy} = \frac{\partial \ln p_x(y)}{\partial \omega_{xy}}$:

$$\frac{\partial \phi_{xy}}{\partial \omega_{xy}} = z_{xy} \frac{\partial \phi_{xy}}{\partial \ln p_x(y)} = z_{xy}\, \phi_{xy}\, \varepsilon_{\phi,p},$$

where $\varepsilon_{\phi,p} = \frac{\partial \ln \phi_{xy}}{\partial \ln p_x(y)}$ is the physical congestion elasticity of the matching technology.

Substituting this derivative into the cross-term, and expanding the revenue curvature using $\frac{\partial z_{xy}}{\partial \omega_{xy}} = (z_{xy})^2 \Psi_{xy}$, where

$$\Psi_{xy} \equiv 1 - \frac{\bar{p}_{xy}}{1 - \bar{p}_{xy}},$$

the three unique components of the Hessian block, evaluated at the optimum, are:

$$H_{qq} = -\theta_y a^*_{xy} \frac{1 + \omega_{xy} z_{xy}}{q_y(x)} \tag{C.1}$$

$$H_{\omega\omega} = \delta_y a^*_{xy} q_y(x) \phi_{xy} z_{xy} \left[z_{xy} (f_{xy} - \omega_{xy}) \Psi_{xy} - 2 \right] \tag{C.2}$$

$$H_{q\omega} = \delta_y a^*_{xy} \phi_{xy} f_{xy} \varepsilon_{\phi,p}\, z_{xy} \left(1 + \omega_{xy} z_{xy}\right) \tag{C.3}$$

The second derivative with respect to screening, H_{qq}, is unconditionally strictly negative due to the global strict convexity of the firm's entropic screening cost. Because changes in the wage alter the optimal screening intensity only through the worker's application response, the cross-derivative $H_{q\omega}$ is strictly proportional to the application semi-elasticity z_{xy}.

The determinant. To guarantee a unique local maximum, the Hessian must be negative definite, which requires $H_{qq} < 0$, $H_{\omega\omega} < 0$, and a strictly positive determinant $\text{Det}(H) = H_{qq} H_{\omega\omega} - (H_{q\omega})^2 > 0$. Multiplying the diagonal elements, the optimal screening intensity $q_y(x)$ cancels:

$$H_{qq} H_{\omega\omega} = -\theta_y \delta_y (a^*_{xy})^2 \phi_{xy} z_{xy} \left(1 + \omega_{xy} z_{xy}\right) \left[z_{xy} (f_{xy} - \omega_{xy}) \Psi_{xy} - 2 \right],$$

while squaring the cross-term yields

$$(H_{q\omega})^2 = \delta_y^2 (a_{xy}^*)^2 \phi_{xy}^2 f_{xy}^2 \varepsilon_{\phi,p}^2 (z_{xy})^2 \left(1 + \omega_{xy} z_{xy}\right)^2 .$$

Both terms share the strictly positive common multiplier $K_{det} \equiv \delta_y (a_{xy}^*)^2 \phi_{xy} z_{xy} \left(1 + \omega_{xy} z_{xy}\right)$. Factoring it out and rearranging terms isolates the positive anchor from the negative subtractions:

$$\mathrm{Det}(H) = K_{det} \left\{ 2\theta_y - z_{xy} \left[\begin{array}{c} \theta_y (f_{xy} - \omega_{xy}) \Psi_{xy} \\ + \delta_y \phi_{xy} f_{xy}^2 \varepsilon_{\phi,p}^2 \left(1 + \omega_{xy} z_{xy}\right) \end{array} \right] \right\} \tag{C.4}$$

This formulation makes the structural dependencies transparent: the positivity of the determinant (and hence the negative definiteness of H) relies entirely on the exogenous firm screening cost ($2\theta_y$) overcoming the terms multiplied by the application elasticity z_{xy}. We analyze the determinant under symmetric scaling $\theta_x \approx \theta_y = \theta$.

Frictionless limit ($\theta \to 0$). As search costs vanish, $\theta_y = 0$ and the positive anchor disappears, while the elasticity $z_{xy} \sim 1/\theta_x$ diverges at any interior application strategy. Substituting into (C.4):

$$\lim_{\theta \to 0} \mathrm{Det}(H) = K_{det} \Big\{ 0 - z_{xy} \cdot \big[0 + \mathrm{const} \big] \Big\} < 0,$$

with the negative terms exploding quadratically in z_{xy}. Because the determinant is strictly negative, the matrix in (ω, q) space is never strictly negative definite in the frictionless limit. At zero information cost, the two instruments (wage and screening) become collinear, creating a coordinate singularity that prevents a unique local maximum.

Uniqueness threshold ($\underline{\theta}$). Conversely, as search costs become large ($\theta \to \infty$), the elasticity shrinks inversely with the worker's cost, $z_{xy} \sim 1/\theta \to 0$. Substituting $\theta_y = \theta$ and $z_{xy} = 1/\theta$ into the determinant yields:

$$\lim_{\theta \to \infty} \mathrm{Det}(H) \propto 2\theta - \mathrm{const} - 0 > 0.$$

The expression grows linearly with θ, confirming that the exogenous firm screening cost strictly overcomes the aggregate cross-substitution terms. This guarantees uniqueness for high levels of costs. By continuity, there exists a strictly positive threshold $\underline{\theta}$ above which $\text{Det}(H) > 0$. Above this threshold, because H_{qq} is unconditionally negative, ensuring a unique maximum strictly reduces to the requirement that the revenue function is concave ($H_{\omega\omega} < 0$).

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 local Hessian matrix for that submarket drops the wage dimension and collapses to the unconditionally strictly negative scalar H_{qq}. ■

D The Planner's First-Order Conditions

Screening. 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 applicant pool, changing $q_y(x)$ only affects the submarket for type x. Differentiating the welfare function Ω yields:

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

which rearranges to the planner's optimal screening condition (2.5).

Applications. The first-order condition with respect to $p_x(y)$ is more complex because an application increases the aggregate pool size ($\delta_y = \sum_{x'} \mu_{x'} p_{x'}(y)$): by the quotient rule, increasing $p_x(y)$ reduces the effective pool weight $a_{x'y}$ for all other applicants, generating the two cross-submarket externalities discussed in the text.

To formalize the congestion externality, define the planner's expected outside option of the applicant pool using the relative matching weights $w^y_{x'|x} \equiv a_{x'y} \frac{q_y(x')\phi_{x'y}}{q_y(x)\phi_{xy}}$:

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

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

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

Taking the derivative of the aggregate entropy costs with respect to $p_x(y)$ yields the Marginal Social Cost (MSC): the worker's direct search cost plus 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 pool. However, we can substitute the planner's screening FOC (2.5) 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 fully 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 pool outside option O_{xy}^{PO} *cancel out* on both sides. The negative social cost of diluting the applicant pool is 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 (2.6). ■

E Proof of Proposition 6 and the Elasticity Mapping

Proof of Proposition 6. Consider the aggregate welfare function $\Omega(p, q)$. Let $p(\omega)$ and $q(\omega)$ be the competitive equilibrium responses to an exogenous wage vector. We define the indirect welfare function $\mathcal{W}(\omega) \equiv \Omega(p(\omega), q(\omega))$. Because ω^{PO} perfectly

decentralizes the planner's optimum, $\mathcal{W}$ reaches its global maximum at ω^{PO}. By the Envelope Theorem, the first-order derivatives vanish at this point: $\nabla_\omega \mathcal{W}(\omega^{PO}) = \mathbf{0}$.

The welfare loss is therefore captured entirely by the second-order curvature. Assuming a sufficiently large number of firms (F) such that the cross-submarket substitution effects (the off-diagonal Hessian elements of the Multinomial Logit) approach zero and the matrix becomes strictly diagonally dominant, the Taylor expansion simplifies to:

$$\Delta \mathcal{W} \approx -\frac{1}{2} \sum_{x,y} \frac{\partial^2 \mathcal{W}}{\partial \omega_{xy}^2} (\Delta \omega_{xy})^2.$$

Applying the chain rule, and noting that the planner's global first-order conditions ($\nabla_p \Omega = \nabla_q \Omega = 0$) eliminate the cross-derivatives involving $\frac{\partial^2 p}{\partial \omega^2}$, the curvature strictly reduces to:

$$\frac{\partial^2 \mathcal{W}}{\partial \omega_{xy}^2} \approx \left(\frac{\partial p_x(y)}{\partial \omega_{xy}} \right)^2 \frac{\partial^2 \Omega}{\partial p_x(y)^2}.$$

Recall that the social planner's first-order condition requires the marginal social benefit of an application to equal its marginal social cost. Using the worker's private optimality condition as an inverse labor supply curve $\omega_{xy}(p)$, we can express the marginal welfare of an application in wage-equivalent units as $\frac{\partial \Omega}{\partial p_x(y)} = \mu_x \mu_y q_y(x) \phi_{xy} \big(\omega_{xy}^{PO} - \omega_{xy}(p) \big)$. Differentiating this with respect to $p_x(y)$ yields the local physical curvature:

$$\frac{\partial^2 \Omega}{\partial p_x(y)^2} = \mu_x \mu_y q_y(x) \phi_{xy} \left(-\frac{\partial \omega_{xy}(p)}{\partial p_x(y)} \right).$$

Crucially, the curvature must be evaluated along the physical environment, in which the congestion rate ϕ_{xy} adjusts with applications through the equilibrium consistency condition—a feedback the planner internalizes but atomistic agents ignore. Incorporating this feedback yields the true physical semi-elasticity:

$$\hat{z}_{xy}^{PO} = \frac{z_{xy}}{1 - \omega_{xy}^{PO} z_{xy} \varepsilon_{\phi,p}},$$

where $z_{xy} = \frac{q_y(x)\phi_{xy}}{\theta_x/\delta_x} (1 - \bar{p}_{xy})$ is the wage semi-elasticity of Appendix A, evaluated at the planner's allocation. By definition of this elasticity, the slope of the inverse normative supply curve is $\frac{\partial \omega_{xy}(p)}{\partial p_x(y)} = \frac{1}{p_x(y) \hat{z}_{xy}^{PO}}$. Substituting this into the curvature, and multiplying

by the squared behavioral application response $\left(\frac{\partial p_x(y)}{\partial \omega_{xy}}\right)^2 = (p_x(y)\hat{z}_{xy}^{PO})^2$, we obtain:

$$\frac{\partial^2 \mathcal{W}}{\partial \omega_{xy}^2} = (p_x(y)\hat{z}_{xy}^{PO})^2 \left[-\frac{\mu_x \mu_y q_y(x)\phi_{xy}}{p_x(y)\hat{z}_{xy}^{PO}}\right] = -\mu_x \mu_y p_x(y) q_y(x)\phi_{xy}\hat{z}_{xy}^{PO} = -M_{xy}\hat{z}_{xy}^{PO}.$$

Substituting this term back into the Taylor expansion eliminates the negative sign and exactly yields the localized Harberger formulation. ■

Derivation of the elasticity mapping (3.2). Because the observed data is generated by the competitive equilibrium, the empirical log-log elasticity e_p captures the realized semi-elasticity at the competitive wage, $\hat{z}_{xy}^{data} = e_p/\omega_{xy}^{CE}$, with the structural form

$$\hat{z}_{xy}^{data} = \frac{z_{xy}}{1 - \omega_{xy}^{CE} z_{xy} \varepsilon_{\phi,p}}.$$

Note that the realized data reflect the *full physical* congestion feedback $\varepsilon_{\phi,p}$ even though agents do not internalize it: an exogenous wage increase draws additional applications, which physically crowd the submarket and alter the realized response relative to the perceived elasticity z_{xy} of Appendix A. Algebraically inverting this relationship identifies the un-congested baseline elasticity:

$$z_{xy} = \frac{\hat{z}_{xy}^{data}}{1 + \omega_{xy}^{CE} \hat{z}_{xy}^{data} \varepsilon_{\phi,p}}.$$

Finally, we evaluate this baseline at the socially optimal benchmark, re-introducing the physical congestion feedback evaluated at the planner's shadow wage:

$$\hat{z}_{xy}^{PO} = \frac{z_{xy}}{1 - \omega_{xy}^{PO} z_{xy} \varepsilon_{\phi,p}}.$$

Substituting the two steps into one another, the intermediate baseline collapses into the single closed-form mapping (3.2). ■