Monopsonistic Search*

Pietro Garibaldi† Espen R. Moen‡ Enrico D. Turri§

November 17, 2026

Abstract

We develop a search-and-matching model in which firms can influence the flow of recruits by posting higher wages. Workers draw a privately observed match-specific amenity cost on contact, so the firm's recruiting probability is increasing in its own wage, giving rise to monopsony power. We show that the decentralized equilibrium is inefficient: a marginal wage increase raises aggregate welfare, and may increase employment. If amenities are drawn from a constant-elasticity distribution, the model can be calibrated using aggregate data only. Calibrated to United States data for 2022–2024, the model implies an inverse acceptance elasticity of 2.9, a wage markdown of 14 percent, and a welfare-maximizing wage 10 percent above the decentralized level. We extend the model to account for fiscal externalities and show that the efficient wage is still 9.5 percent higher than the monopsonistic wage.

Keywords: Monopsony; search and matching; wage posting; job amenities; private information; efficiency; free entry; recruiting costs; fiscal externalities.
JEL Codes: E24, J31, J38, J42, J63, J64.

*We thank Tito Boeri, Thomas Le Barbanchon, and seminar participants at Bocconi University and the BSE Summer Forum in Barcelona for comments and suggestions.

†Collegio Carlo Alberto, University of Torino, CEPR, IZA, pietro.garibaldi@unito.it
‡Norwegian Business School, CEPR, espen.r.moen@bi.no
§London School of Economics, e.turri@lse.ac.uk

1 Introduction

Firms pay workers less than they produce. In United States manufacturing, the average plant pays its workers about 65 cents for each dollar of marginal revenue product they generate (Yeh et al., 2022), and estimates of the elasticity of labor supply to the individual firm, the textbook measure of wage-setting power, fall far short of the infinite value that perfect competition would require (Manning, 2021; Kline, 2025). Monopsony has thus become a central concept in modern labor economics, and recent work shows that its consequences reach beyond the individual firm to aggregate employment, wages and welfare (Berger et al., 2022, 2026). The standard model of the aggregate labor market, however, has no room for it. In the Diamond–Mortensen–Pissarides (DMP) model (Diamond, 1982a; Mortensen, 1982; Pissarides, 1985; Mortensen and Pissarides, 1994), the firm takes the probability of filling its vacancy as given: a firm cannot hire faster by paying more. This raises the question: what does monopsony imply for the efficiency of a frictional labor market with free entry, and how large is the distortion?

This paper develops a search-and-matching model of monopsony and uses it to study how the decentralized and efficient equilibria differ. We embed an upward-sloping labor supply curve into a standard search-and-matching model with a matching function and free entry, and derive the implied Beveridge curve. When a worker meets a firm, she draws a match-specific amenity cost, which she observes and the firm does not, and accepts the offer only if the wage compensates it. A firm that pays more is therefore more likely to have its offers accepted. The firm's recruiting probability is thus the product of an acceptance probability, which rises with its own wage, and a matching probability, which the firm takes as given. This is the search-and-matching counterpart of the recruiting function of dynamic monopsony models, in which employment at a firm is the ratio of its recruits to its quits (Manning, 2003, 2021). Monopsony power is then summarized by a single object, the elasticity of the acceptance probability to the wage.

Monopsonistic search sits between two traditions in search theory. Wage-posting models in the tradition of Burdett and Mortensen (1998) give monopsony a search-theoretic foundation and generate wage dispersion among identical workers, but they take contact rates as exogenous, abstract from free entry and have no Beveridge curve, which limits their use for macroeconomic and policy analysis. Directed and competitive search, following Moen (1997), let posted wages affect hiring while keeping free entry, but their equilibria are typically efficient: firms compete for workers through the wages they post, and monopsony power in the sense of Manning (2021) disappears. In our model wages affect hiring and each firm faces its own upward-sloping supply of recruits.

We conduct our analysis of monopsonistic search in three steps. First, we characterize the decentralized equilibrium and show that it is inefficient: at the market wage, a marginal increase in the mean wage raises aggregate welfare. A firm that raises its wage trades a higher probability that a contacted worker accepts against a smaller surplus from the filled job, and ignores the worker's surplus from the match. Since free entry dissipates profits, welfare is the value of a searching worker, and a marginal wage increase, leaving the firm's value unchanged to first order, raises it. Comparing the firm's first-order condition with that of a constrained

planner, who chooses a common wage subject to the same matching technology and the same free-entry condition, we show that the wedge between the monopsonistic and optimal wages is governed by three elasticities: of the acceptance probability, of firm surplus and of worker surplus to the wage. We also show that monopsonistic search changes the relation between wages and employment. A higher wage raises employment through acceptance, as in the classic monopsony model of Robinson (1969), but lowers it through firm entry, as in the standard search-and-matching model. With constant elasticities, employment is inverse-U shaped in the wage: at the market wage a wage increase may raise or lower employment, while at the optimal wage it always lowers it.

Second, we show that monopsony power can be measured from aggregate data alone. When amenities follow a constant-elasticity distribution, a single parameter γ governs the acceptance elasticity and the equilibrium is unique. The monopsonistic wage then expresses γ as the ratio of the worker's rent to the firm's, and the value of unemployment has a closed form. Combining the two gives a quadratic in γ with a unique positive root, which depends only on the average wage, the flow value of non-employment, net output per worker, the job-finding rate and the separation rate. No micro data on wages or amenities are thus needed to calibrate the model.

Third, we take our model of monopsonistic search to United States data, over the period 2022–2024. The calibration finds that the inverse acceptance elasticity is $1/\gamma = 2.9$, and that the firm's labor cost is 14 percent below the marginal product in wage units,[1] at the conservative end of empirical estimates (Manning, 2021; Yeh et al., 2022). The welfare-maximizing mean wage is 10.1 percent above the market wage. It raises welfare, measured as the value of searching for a job, by 5.6 percent, but it also raises unemployment from 3.8 to 6.4 percent.

We then generalize the model to account for fiscal externalities. A monopsonist does not internalize the effect of its wage on the public budget: a lower wage narrows the payroll tax base directly, and affects employment, and with it both the tax base and the number of unemployed workers who claim benefits. Revenue is hump-shaped in the wage, because the direct effect on the tax base dominates near the market wage, and the effect through employment dominates at higher wages: receipts peak at a wage 4.6 percent above the market wage. With a marginal value of public funds of 1.25, internalizing the externality on the public budget lowers the efficiency premium from 10.1 to 9.5 percent, and the correction stays below one percentage point for every value of the marginal value of public funds and every definition of the tax base we consider. In an economy with higher payroll taxes instead, fiscal considerations would strengthen the case for higher wages: the sign of the fiscal externality depends on institutional design, not on monopsony inherently.

Related literature. This paper contributes to three strands of literature. The first is the large literature on monopsony in labor markets, surveyed by Manning (2021), Card (2022), Azar and Marinescu (2024) and Kline (2025). Empirical work estimates the elasticity of labor supply to the firm from wage variation and reads it as a measure of monopsony power

[1] The markdown equals one minus the ratio of the labor share to net output, and is therefore independent of the wage measure.

(Staiger et al., 2010; Ransom and Sims, 2010; Falch, 2010; Azar et al., 2022; Amodio and De Roux, 2024; Kroft et al., 2025), while production-side approaches estimate markdowns directly (Brooks et al., 2021; Yeh et al., 2022). Quantitative models derive wage-setting power from workers' heterogeneous preferences over employers, with a labor supply to the firm that responds to wages instantaneously (Berger et al., 2022; Lamadon et al., 2022; Chan et al., 2020). A related macroeconomic literature studies the consequences of labor market power for wages, the labor share and inequality (Deb et al., 2022, 2024; Berger et al., 2026), and for aggregate productivity (Mottironi, 2025; Armangué-Jubert et al., 2026; Bachmann et al., 2026). Our amenity mechanism shares with these models, and with the evidence that workers trade wages against the non-wage characteristics of jobs (Sorkin, 2018), the idea that preferences over employers give firms wage-setting power. In our model, however, these preferences are revealed at contact in a frictional market, so that monopsony power, vacancy creation and unemployment are determined jointly.

The second strand studies wage setting in frictional labor markets. Dynamic monopsony models in the spirit of Burdett and Mortensen (1998) capture labor market dynamics and generate a firm-level supply of labor from search frictions (Bontemps et al., 2000; Manning, 2003; Gouin-Bonenfant, 2022; Gottfries and Jarosch, 2023; Jarosch et al., 2024; Bloesch et al., 2024), but typically abstract from the DMP framework and free entry. Our acceptance margin is related to stochastic job matching (Pissarides, 2000, ch. 6), in which not every meeting turns into a match, and whose efficiency properties go back to Pissarides (1984). Here, however, the match-specific component is an amenity that the firm neither observes nor values, as in competitive search under asymmetric information (Moen and Rosén, 2011). Closest to us is Malgieri and Souchier (2026), who likewise generate monopsony power from privately observed job amenities within a DMP environment, but organize the analysis around on-the-job search and the dispersion of markdowns along the job ladder. We abstract from the ladder and ask instead what a single posted wage implies for aggregate efficiency and free entry, and how monopsony power can be measured from aggregate data.

The third strand studies the efficiency of search equilibria and the scope for wage policy under monopsony. In the DMP model with bargaining, the decentralized equilibrium is efficient only when the worker's bargaining share equals the elasticity of the matching function (Diamond, 1982b; Hosios, 1990), and competitive search delivers this condition through wage posting (Moen, 1997; Wright et al., 2021). With full information and full wage-setting power, instead, firms push workers down to their reservation wage, as in Diamond (1971). In monopsonistic search firms post wages without observing the worker's amenity, which leaves workers an information rent. The efficient wage premium we compute therefore speaks to the classic case for wage floors under monopsony (Robinson, 1969; Card and Krueger, 1994; Manning, 2004). It complements Berger et al. (2025), who quantify the efficiency and welfare effects of minimum wages in an estimated oligopsony model with heterogeneous employers and find limited scope for welfare gains. In our framework the wedge operates through the acceptance margin under free entry, and its size is pinned down by aggregate moments.

Outline. The paper proceeds as follows. Section 2 develops the model of monopsonistic search, derives the monopsonistic wage and the efficiency wedge, and specializes it in Sec-

tion 2.8 to the case with constant elasticities, in which monopsony power is identified from aggregate data. Section 3 takes the model to United States data: Section 3.1 presents the moments and the accounting conventions, Section 3.2 reports the calibration and the efficient wage, and Section 3.3 discusses what the calibrated recruiting cost measures. Section 4 introduces the government budget and computes the fiscal-adjusted optimal wage. Section 5 concludes.

2 A Model of Monopsonistic Search

This section presents the model of monopsonistic search. We first describe the labor market and the firm's recruiting function, and then derive the value functions, the worker's reservation cost and the firm's wage-setting problem, which give an explicit expression for the monopsonistic wage. We then define the equilibrium and compare it with the allocation chosen by a constrained planner in Section 2.7. Finally, Section 2.8 specializes the model to a canonical case with constant elasticities, in which the equilibrium is unique and monopsony power can be measured from aggregate data.

2.1 The Labor Market and Firm Recruiting Function

The labor market is populated by a unit measure of infinitely lived, risk-neutral and ex ante identical workers, and by firms that post vacancies. Unemployed workers and vacancies meet through a matching function $m(u, v)$ with constant returns to scale, where u is unemployment and v the number of vacancies. A vacancy meets a worker at rate $q(\theta) = m(u, v)/v$, where $\theta = v/u$ is market tightness and $q'(\theta) < 0$, and an unemployed worker meets a vacancy at rate $\theta q(\theta)$. As in the standard model, the matching function governs contacts. Unlike the standard model, a contact turns into a match only if the worker accepts the firm's offer.

When a worker and a firm meet, the worker draws a match-specific cost x of working at that job, which we also call an amenity cost. The cost is drawn independently across matches from a continuous distribution F with support $\Omega = [0, \bar{x}]$, and it is private information to the worker: the job is an inspection good, whose cost the worker learns at the meeting and the firm never observes. We think of x as any feature of a job that workers care about and firms do not, and over which workers' preferences differ, such as commuting distance, the tasks to be performed and how hard the worker finds them, or the colleagues she would work with. The assumption is close to the match-specific productivity of the stochastic matching model (Pissarides, 2000, ch. 6), with one difference: x affects the worker's utility but not the firm's output.

The firm sets the wage when it meets an unemployed worker and has all the bargaining power: it makes a take-it-or-leave-it offer, which the worker accepts or rejects. The firm is committed to its offer, and there is no renegotiation, since the worker has no credible way to reveal her cost x. This is in line with work on competitive search under asymmetric information (Moen and Rosén, 2011). With full information, the firm would simply pay each worker her reservation wage.[2]

[2]Appendix C solves the model under full information and describes an information-congestion paradox.

Suppose that firm i offers the wage w_i to a worker who has drawn the cost x. Let $W(x, w_i)$ be the worker's present discounted value of the job, strictly decreasing in x and increasing in w_i, and let U, the value of unemployment, be her outside option. The worker accepts the offer if $W(x, w_i) \geq U$. Since W is continuous and strictly monotonic, there is a unique reservation cost $x^*(w_i)$, defined by $W(x^*(w_i), w_i) = U$, such that the worker accepts if and only if $x \leq x^*(w_i)$. A contact therefore turns into a match with probability $F(x^*(w_i))$, which increases with the wage, and the firm's recruiting function is

$$R(w_i, \theta) = F\big(x^*(w_i)\big)\, q(\theta) \ .$$

The firm takes market tightness θ and the other aggregate variables as given, but it can raise its recruiting rate by paying more. The recruiting function $R(w_i, \theta)$ is the search-and-matching counterpart of the recruiting function of dynamic monopsony models (Manning, 2003, 2021).

The model relates to three familiar benchmarks. With full information, it collapses to the paradox of Diamond (1971): firms that hold all the wage-setting power pay workers their reservation wage. The firm's full bargaining power also makes the model comparable with Nash bargaining when workers have no bargaining power (Mortensen, 1982), and with competitive search (Moen, 1997). We use these comparisons when we derive the optimal wage.

Each firm has one job. A filled job produces output y, and a vacancy costs c per unit of time. Firms enter freely, and opening a job requires a fixed cost κ, so that in equilibrium the value of a vacancy equals the entry cost,

$$V^* = V(w^*, \theta) = \kappa \ ,$$

where w^* is the monopsonistic wage, the wage that maximizes the value of a vacancy, and V^* is that value. Unemployed workers receive a flow utility z, for example from home production, which does not depend on the equilibrium of the labor market.

2.2 Value Functions

Jobs are destroyed exogenously at rate λ. Let w_i denote the wage offered by job i. In principle, firms could post heterogeneous wages in equilibrium, in which case the value of unemployment would depend on the distribution of wages offered in the market. However, without on-the-job search, workers sample wages only when they leave unemployment, and the equilibrium wage distribution is degenerate: all firms post the same wage.

Let J be the value to a firm of a filled job, V the value of a vacancy, W the value to a worker of being employed, and U the value of unemployment. These values solve the

With full information, firms capture the whole surplus, which free entry dissipates in vacancy costs, so welfare is zero. With private information, workers keep a rent and welfare is positive. By contrast, a perfectly discriminating monopolist in a static setting achieves efficiency, and so would firms with full information in a setting without free entry.

following system of asset equations:

$$rJ = y - w_i + \lambda(V - J), \tag{1}$$

$$rV = \max_{w_i} \{-c + q(\theta)F(w_i - rU)(J - V)\}, \tag{2}$$

$$rW = w_i - x + \lambda(U - W), \tag{3}$$

$$rU = z + \theta q(\theta) \left[\int_0^{\bar{x}} \max\{W(x), U\} dF(x) - U \right], \tag{4}$$

where r is the discount rate and y, c, z and x are output, the flow cost of a vacancy, the flow utility of unemployment and the worker's amenity cost, as defined above.

Throughout the analysis we impose three regularity conditions on the distribution of match-specific amenities. These assumptions are standard in search models with endogenous acceptance decisions, and they ensure that firms face a well-defined and smooth labor supply when setting wages.

Assumption 1 (Amenities distribution)**.** *The match-specific amenity x is distributed on a compact support $[0, \bar{x}]$ with cumulative distribution function F that is continuous and strictly increasing on its support. The associated density f exists and is continuous on $(0, \bar{x})$.*

Assumption 2 (Interior acceptance)**.** *In equilibrium, the reservation cost $x^*(w) = w - rU$ lies strictly inside the support of the distribution, so that*

$$0 < F(x^*(w)) < 1.$$

This guarantees that job acceptance decisions are interior and that firms' wage-setting problems admit interior solutions.

Assumption 3 (Elasticity regularity)**.** *The acceptance elasticity*

$$\varepsilon_F(x) \equiv \frac{f(x)\, x}{F(x)}$$

is finite and continuous on $(0, \bar{x})$.

These assumptions ensure that small changes in wages generate smooth changes in workers' acceptance decisions and that firms face a well-defined monopsonistic labor supply at the equilibrium wage. The results below rely only on local properties of the acceptance margin and do not require a particular functional form for F. Stronger assumptions are introduced in Section 2.8 to establish existence and uniqueness of equilibrium in the canonical constant-elasticity case.

Equation (1) describes the value of a filled job. A firm employing a worker earns flow profits $y - w_i$ and loses the worker at rate λ. Upon separation, the firm is left with a vacancy of value V. The term $\lambda(V - J)$ therefore captures the expected capital loss from job destruction. Equation (2) characterizes the value of a vacancy. Maintaining a vacancy entails a flow cost c. At rate $q(\theta)$, the firm contacts a worker, who accepts the job if and

only if the offered wage yields non-negative surplus relative to unemployment. Since the firm does not observe x, its offer is accepted with probability $F(w_i - rU)$. Conditional on acceptance, the firm obtains the surplus $J - V$. Equation (3) gives the value of employment for a worker with cost x. While employed, the worker receives wage w_i but incurs flow cost x. Jobs are destroyed at rate λ, in which case the worker becomes unemployed and obtains value U. Finally, equation (4) defines the value of unemployment. An unemployed worker receives flow utility z and meets firms at rate $\theta q(\theta)$. Upon meeting a firm, the worker observes the realization of x and accepts the job whenever $W(x) \geq U$. The integral term therefore captures the expected value of job offers net of the value of remaining unemployed.

In equilibrium, the value of unemployment U depends on the wage posted by firms in the market. However, when choosing its wage w_i, an individual firm takes U as given. Since each firm is infinitesimal, a change in its wage has no effect on workers' job-finding rate and hence on the equilibrium value of unemployment. As a result, wage setting affects worker behavior only through the acceptance margin $w_i - rU$.

2.3 Worker Asset Equations and Reservation Cost

Each match is characterized by the cost x, which the worker observes and cannot credibly reveal to the firm. For a worker employed at wage w, the value of employment $W(w, x)$ satisfies

$$(r + \lambda)W(w, x) = w - x + \lambda U,$$

where λ is the exogenous job destruction rate and U is the value of unemployment. The worker's rent from employment relative to unemployment is therefore

$$Q^W(w, x) \equiv W(w, x) - U = \frac{w - x - rU}{r + \lambda}. \tag{5}$$

To simplify the analysis, we assume that the support Ω of x is wide enough that no firm can offer a wage high enough to guarantee acceptance with probability one while making positive profits. The maximum wage a firm would ever post satisfies $J(w_{\max}) = V = \kappa$, which from (9) below yields $w_{\max} = y - r\kappa$. The maximum cost consistent with acceptance at any profitable wage, evaluated at the lowest possible outside option $rU = z$, is therefore $\bar{x} = w_{\max} - z = y - r\kappa - z$. We therefore set

$$\Omega = [0,\ y - r\kappa - z]\ .$$

For a given wage w and value of unemployment U, a worker accepts the job whenever $Q^W(w, x) \geq 0$. This defines a reservation cost $x^*(w)$ such that the worker is indifferent between employment and unemployment. Setting $Q^W(w, x^*) = 0$ in (5) gives the reservation rule

$$x^*(w) = w - rU. \tag{6}$$

Workers with $x \leq x^*(w)$ accept the job, while workers with a higher cost reject it.

Given the reservation rule (6), the value of unemployment satisfies

$$rU = z + \theta q(\theta) \int_0^{x^*(w)} [W(w, x) - U]\ dF(x),$$

or, equivalently,

$$rU = z + \theta q(\theta) \int_0^{x^*(w)} Q^W(w,x)\, dF(x), \tag{7}$$

which highlights that unemployed workers anticipate receiving wage offers posted by monopsonistic firms.

Substituting (5) into (7) yields

$$rU = z + \frac{\theta q(\theta)}{r+\lambda} \left[(w - rU) F(x^*(w)) - \int_0^{x^*(w)} x\, dF(x) \right].$$

This expression can be rewritten as

$$rU = \frac{r+\lambda}{r+\lambda+\theta q(\theta) F(x^*(w))} z + \frac{\theta q(\theta) F(x^*(w))}{r+\lambda+\theta q(\theta) F(x^*(w))} \left(w - \mathbb{E}[x \mid x \le x^*(w)] \right), \tag{8}$$

which defines U implicitly as a nonlinear function of the equilibrium wage.

Although we write U as a function of w, here w is the equilibrium wage posted by all firms in the market, which a single firm takes as given.

2.4 Firm Monopsonistic Problem

Consider a monopsonistic firm that posts a wage w and employs at most one worker. Jobs are destroyed exogenously at rate λ. For a given wage w, the value of a filled job satisfies

$$(r+\lambda) J(w) = y - w + \lambda V, \tag{9}$$

where y denotes output and V is the value of a vacancy. The firm's rent from a filled job, net of the continuation value of a vacancy, is therefore given by

$$Q^F(w) \equiv J(w) - V = \frac{y - w - rV}{r+\lambda}.$$

The value of a vacancy to the monopsonist solves

$$rV = \max_w \left\{ -c + q(\theta) F(x^*(w)) \left[J(w) - V \right] \right\} = \max_w \left\{ -c + q(\theta) F(x^*(w)) Q^F(w) \right\},$$

where c is the flow vacancy cost, $q(\theta)$ is the job-filling rate, and $F(x^*(w))$ is the probability that a contacted worker accepts the job. Acceptance depends on the reservation cost $x^*(w)$, which is determined on the worker side and satisfies

$$x^*(w) = w - rU.$$

When choosing its wage, the firm takes the value of unemployment U as given, as discussed above, so its wage affects worker behavior only through the acceptance margin $x^*(w)$.

The first-order condition for optimal wage posting equates the marginal benefit of a higher wage, through a higher probability of acceptance, to its marginal cost in terms of lower firm rents:

$$\frac{\partial F(w-rU)}{\partial w}Q^F(w) = F(w-rU)\frac{\partial Q^F(w)}{\partial w}. \tag{11}$$

This condition can be rewritten in elasticity form as

$$\varepsilon_F(x^*(w)) = \epsilon_{Q^F,w}(w)\frac{w-rU}{w},$$

where ε_F is the acceptance elasticity of Assumption 3 and elasticities are expressed in absolute value.

As in the static monopsony problem, optimal wage setting balances the elasticity of the acceptance probability with respect to wages against the elasticity of firm rents. The additional term $(w-rU)/w$ reflects the dynamic nature of the problem and captures the fact that wages affect acceptance only through the worker's net surplus relative to unemployment.

2.5 An Explicit Expression for the Wage

To obtain an explicit expression for the monopsonistic wage, we start from the firm's first-order condition for wage setting in (11). Using the definitions of $Q^F(w)$ and $x^*(w)$, this condition can be written as

$$f(w-rU)\frac{y-w-rV}{r+\lambda} = \frac{F(w-rU)}{r+\lambda}, \tag{12}$$

where $f(\cdot)$ and $F(\cdot)$ are the density and distribution function of the cost x.

It is convenient to express the firm's problem in terms of the elasticity of the acceptance probability with respect to the reservation cost. Define

$$\epsilon_{F,x}(x) \equiv \frac{f(x)\,x}{F(x)}, \tag{13}$$

which measures the elasticity of the acceptance probability with respect to the reservation cost.

Using (13), the first-order condition (12) can be rewritten as

$$\epsilon_{F,x}(w-rU) = \frac{w-rU}{y-w-rV}.$$

Solving this condition for the wage yields the optimal wage posted by a monopsonistic firm,

$$w^* = \frac{\epsilon_{F,x}(w^*-rU)}{\epsilon_{F,x}(w^*-rU)+1}(y-rV) + \frac{1}{\epsilon_{F,x}(w^*-rU)+1}rU. \tag{14}$$

Equation (14) expresses the equilibrium wage as a weighted average of output net of the flow value of a vacancy, $y-rV$, and the flow value of unemployment, rU. The weights depend on the elasticity of the acceptance probability, which measures the firm's monopsony power.[3]

[3]When F belongs to the constant-elasticity family with parameter γ, the elasticity $\epsilon_{F,x}(\cdot)$ is constant and equal to γ.

Finally, free entry implies that the value of a vacancy equals the entry cost,

$$V = \kappa. \tag{15}$$

Substituting (15) and the optimal wage (14) into the vacancy value equation (10) yields an equilibrium condition that, together with the worker-side equations, pins down market tightness,

$$r\kappa = -c + q(\theta)F(w^* - rU)\frac{y - w^* - r\kappa}{r + \lambda}. \tag{16}$$

2.6 Unemployment and Equilibrium Definition

Aggregate unemployment evolves according to a standard Beveridge curve. Let u denote the unemployment rate. Employment relationships are destroyed at exogenous rate λ, so that a measure $\lambda(1-u)$ of workers flows into unemployment per unit of time. Unemployed workers meet firms at rate $\theta q(\theta)$ and accept job offers with probability $F(w^* - rU)$, where w^* denotes the equilibrium wage. The law of motion for unemployment is therefore given by

$$\dot{u} = \lambda(1-u) - \theta q(\theta)F(w^* - rU)u.$$

In steady state, unemployment is constant and satisfies

$$u = \frac{\lambda}{\lambda + \theta q(\theta)F(w^* - rU)}. \tag{17}$$

We can now define the equilibrium of the model.

Definition 1. *An equilibrium is a set of value functions (J, V, W, U), a wage w^*, a reservation cost x^*, a market tightness θ, and an unemployment rate u such that:*

1. *Given U, the wage w^* solves the firm's monopsonistic wage-setting problem and satisfies* (14)*.*
2. *Free entry holds and market tightness θ satisfies* (16)*.*
3. *Given w^*, workers optimally accept jobs according to the reservation rule* (6)*, and the value of unemployment U solves* (8)*.*
4. *The unemployment rate u is determined by the steady-state Beveridge curve* (17)*.*

2.7 The Efficiency Wedge: A Tale of Three Elasticities

This subsection presents the main theoretical results of the monopsonistic search model. We focus on the welfare and unemployment effects of wage increases around the decentralized equilibrium. Throughout, it is useful to interpret these results in light of the firm's wage-setting problem derived in Section 2.4, where wages affect outcomes exclusively through workers' acceptance decisions and firms' vacancy creation incentives.

The main result is established by the following theorem.

Theorem 1. *At the decentralized monopsonistic equilibrium, a marginal increase in the mean wage raises aggregate welfare.*

Proof. See Appendix B.2. □

The welfare result follows from a standard envelope argument. At the decentralized equilibrium, firms choose wages to maximize their own surplus, taking the value of unemployment and market tightness as given. As a consequence, the private first-order condition eliminates any first-order effect of wages on firm profits.

From a social perspective, however, wages play an additional role. A higher wage raises the reservation cost, and with it the probability that a meeting results in a match. Since firms do not internalize this acceptance externality when setting wages, the decentralized monopsonistic wage is inefficiently low. A marginal increase in the mean wage therefore improves allocative efficiency and raises welfare.

We next explore the implications of a wage increase for employment.

Theorem 2. *At the decentralized monopsonistic equilibrium, the effect of a marginal increase in the mean wage on unemployment is ambiguous.*

Proof. See Appendix B.3. □

The ambiguity arises because wage increases operate through three channels that work in opposite directions. First, a higher wage directly raises workers' acceptance probability, increasing the flow from unemployment to employment. Second, higher wages raise the value of unemployment by improving expected job surplus. This indirect effect reduces acceptance incentives, but is locally dominated by the direct effect at the decentralized equilibrium. Third, higher wages reduce firms' expected surplus from a filled job, discouraging vacancy creation. This reduces market tightness and lowers the probability that unemployed workers meet firms.

Whether unemployment rises or falls therefore depends on the relative strength of the acceptance margin and the vacancy creation margin, which in turn is governed by parameters such as the elasticity of job acceptance, vacancy posting costs, and the job destruction rate.

To clarify the source of inefficiency in the decentralized equilibrium, we compare monopsonistic wage setting with the allocation chosen by a constrained planner. The planner selects a common wage w to maximize steady-state welfare, measured by the value of unemployment,

$$\max_{w} \; rU = z + \theta q(\theta) F(w - rU)\, Q^W(w),$$

subject to the free-entry condition

$$r\kappa + c = q(\theta) F(w - rU)\, Q^F(w).$$

The planner internalizes how wages affect both worker acceptance decisions and firms' vacancy creation. Worker and firm rents are given by

$$Q^W(w) = \mathbb{E}\left[\frac{w - x - rU}{r + \lambda} \,\middle|\, x \le w - rU \right], \qquad Q^F(w) = \frac{y - w - r\kappa}{r + \lambda}.$$

Decentralized monopsony. In the decentralized equilibrium, a firm takes market tightness θ as given when choosing its wage. Locally, $\partial\theta/\partial w = 0$. The firm balances two opposing forces: a higher wage increases the probability that a contacted worker accepts the job, but reduces the surplus from a filled job. The first-order condition can be written as

$$\epsilon_{F,w-rU}\frac{w}{w-rU} = \epsilon_{Q^F,w}, \tag{18}$$

which equates the elasticity of the acceptance probability to the elasticity of firm surplus. Monopsonistic firms therefore choose wages to maximize their own expected rents, ignoring the broader equilibrium effects on workers and vacancy creation.

The planner's problem. By contrast, the planner recognizes that wages affect market tightness. Higher wages reduce firms' expected surplus, lower vacancy creation, and thus decrease θ. Let $\eta_{q,\theta}$ denote the elasticity of the matching function with respect to market tightness. Log-differentiating the free-entry condition yields

$$\eta_{q,\theta}\,\epsilon_{\theta,w} = \epsilon_{F,w-rU}\frac{w}{w-rU} + \epsilon_{Q^F,w}, \tag{19}$$

which captures how wages propagate to market tightness through (i) acceptance decisions and (ii) firm surplus.

The planner's first-order condition can be written as

$$(1-\eta_{q,\theta})\epsilon_{\theta,w} + \epsilon_{F,w-rU}\frac{w}{w-rU} + \epsilon_{Q^W,w} = 0. \tag{20}$$

Combining (19) and (20) yields

$$\eta_{q,\theta}\,\epsilon_{Q^W,w} + \epsilon_{F,w-rU}\frac{w}{w-rU} = (1-\eta_{q,\theta})\epsilon_{Q^F,w}. \tag{21}$$

A tale of three elasticities. Equation (21) highlights three key elasticities governing efficiency:

1. **The recruiting elasticity** $\epsilon_{F,w-rU}$, which measures how strongly worker acceptance responds to wages. A low recruiting elasticity implies strong monopsony power: wages have little effect on hiring, allowing firms to compress wages below the efficient level.

2. **The firm-surplus elasticity** $\epsilon_{Q^F,w}$, which captures how sensitive firm rents are to wage changes. This term governs the vacancy-creation channel.

3. **The worker-surplus elasticity** $\epsilon_{Q^W,w}$, which measures how much worker rents respond to wages. This term is entirely absent from the decentralized condition (18).

The decentralized equilibrium (18) balances only the first two elasticities. The planner condition (21) introduces the third elasticity and weights firm and worker surplus elasticities by the matching elasticity $\eta_{q,\theta}$. Intuitively, $\eta_{q,\theta}$ governs how strongly vacancy creation responds to profitability. When $\eta_{q,\theta}$ is high, wages have large effects on market tightness, and the vacancy-creation externality becomes quantitatively important.

Relation to competitive search and Hosios. The decentralized and planner allocations coincide only in the limiting case $\eta_{q,\theta} = 1$. In that case, the planner condition collapses to

$$\epsilon_{F,w-rU} \frac{w}{w - rU} = \epsilon_{Q^F,w},$$

so that the monopsonistic outcome is efficient. This limiting case corresponds to competitive search (Moen, 1997) and satisfies the Hosios condition (Hosios, 1990), under which private wage setting internalizes matching externalities.

Outside this knife-edge case, monopsonistic firms underweight worker surplus and ignore the impact of wages on vacancy creation. As a result, equilibrium wages are inefficiently low. Efficiency ultimately depends on the interaction of the three elasticities $\epsilon_{F,w-rU}$, $\epsilon_{Q^F,w}$ and $\epsilon_{Q^W,w}$, weighted by the matching elasticity $\eta_{q,\theta}$, which together determine the magnitude of monopsony power and the welfare wedge.

Wages and employment at the optimal wage Theorem 2 states that raising the mean wage near the monopsonistic equilibrium may increase or decrease employment. One may then ask how wages and employment are related at the optimal wage. The next proposition provides the answer.

Proposition 1. *At the optimal equilibrium, a marginal increase in the mean wage increases unemployment.*

Proof. See Appendix B.4. □

Intuitively, since free entry exhausts firm profits, welfare coincides with workers' utility. If wages could be raised without raising unemployment, workers' utility would increase, so the wage could not be optimal. This implies that the optimal wage lives in the tradeoff region when a marginal increase increases unemployment.

2.8 The Canonical Model

To obtain sharper analytical results, we now specialize the model to a canonical environment in which both workers' acceptance behavior and the matching technology exhibit constant elasticity. This parametrization allows us to derive closed-form expressions for wages and to characterize the existence and uniqueness of the equilibrium and its comparative statics.

2.8.1 Constant-Elasticity Acceptance and Matching

We assume that the cost x follows a constant-elasticity distribution. A result from the statistical literature (see Veres-Ferrer and Pavia, 2022) establishes that a random variable supported on $[0, \bar{x}]$ has constant elasticity if and only if its cumulative distribution function can be written as

$$F(x) = \begin{cases} 0, & x < 0, \\ \left(\frac{x}{\bar{x}}\right)^{\gamma}, & 0 \le x \le \bar{x}, \\ 1, & x > \bar{x}, \end{cases}$$

where $\gamma > 0$ is the elasticity parameter. We apply this result with elasticity γ and the support of Section 2, whose upper bound $\bar{x} = y - r\kappa - z$ is the maximum cost consistent with acceptance at any profitable wage. This ensures that there always exists a wage that induces full acceptance, while preserving interior acceptance decisions in equilibrium. In addition, we assume a constant-elasticity matching function,

$$q(\theta) = A\theta^{-\alpha},$$

with $\alpha \in (0, 1)$. Together, these assumptions imply that both the probability of acceptance and the job-filling rate respond to wages and market tightness with constant elasticities.

2.8.2 Existence and Uniqueness of Equilibrium

Under constant elasticity, the equilibrium of the model can be summarized by three equations. First, the monopsonistic wage takes the closed-form expression

$$w = \frac{\gamma}{\gamma + 1}(y - r\kappa) + \frac{1}{\gamma + 1} rU. \tag{22}$$

Second, free entry implies

$$(r\kappa + c)(r + \lambda) = q(\theta)F(w - rU)(y - w - r\kappa).$$

Finally, the value of unemployment satisfies

$$rU = \frac{(r + \lambda)z + \theta q(\theta)F(w - rU)\left(w - \mathbb{E}[x \mid x \leq w - rU]\right)}{r + \lambda + \theta q(\theta)F(w - rU)}. \tag{23}$$

Substituting the constant-elasticity expressions into these conditions, the equilibrium system can be rewritten as

$$\begin{aligned}
(r\kappa + c)(r + \lambda) &= A\theta^{-\alpha}\left(\frac{\gamma}{\gamma + 1}\frac{y - r\kappa - rU}{\bar{x}}\right)^{\gamma}\frac{y - r\kappa - rU}{\gamma + 1} \\
(r + \lambda)(rU - z) &= A\theta^{1-\alpha}\left(\frac{\gamma}{\gamma + 1}\frac{y - r\kappa - rU}{\bar{x}}\right)^{\gamma}\frac{\gamma}{\gamma + 1}\frac{y - r\kappa - rU}{\gamma + 1}
\end{aligned}$$

which can be combined to give

$$\theta = \frac{\gamma + 1}{\gamma}\frac{rU - z}{r\kappa + c}$$

and

$$(r\kappa + c)^{1-\alpha}(\lambda + r)\frac{(rU - z)^{\alpha}}{\bar{x}} = A\left(\frac{\gamma}{\gamma + 1}\right)^{\gamma+1+\alpha}\frac{1}{\gamma}\left(\frac{y - r\kappa - rU}{\bar{x}}\right)^{\gamma+1},$$

Since the left-hand side increases and the right-hand side decreases in rU, exactly one $rU \in (z, y - r\kappa)$ satisfies this condition.

Theorem 3. *In the canonical monopsonistic search model there exists a unique equilibrium.*

2.8.3 Identifying Monopsony Power from Aggregate Data

A practical attraction of the constant-elasticity model is that its single index of monopsony power, γ, can be recovered from standard aggregate moments — the wage, the value of non-employment, output, and the worker-flow rates — without any micro data on the amenity distribution. Two equilibrium relations pin it down. First, the monopsonistic wage (22) inverts to express the elasticity as the ratio of the worker's rent to the firm's,

$$\gamma = \frac{w - rU}{(y - r\kappa) - w}. \tag{24}$$

Second, under constant elasticity the value of unemployment (23) collapses to the closed form

$$rU = \frac{(r + \lambda)\, z + \dfrac{p}{\gamma + 1}\, w}{r + \lambda + \dfrac{p}{\gamma + 1}}, \tag{25}$$

where $p \equiv \theta q(\theta) F(x^*)$ is the job-finding rate, obtained from the steady-state Beveridge relation $p = \lambda(1 - u)/u$. Substituting (25) into (24) and writing $S \equiv y - r\kappa$ for the net flow surplus yields a quadratic in γ,

$$(S - w)\,\gamma^2 + \left[\frac{r + \lambda + p}{r + \lambda}\,(S - w) - (w - z)\right]\gamma - (w - z) = 0. \tag{26}$$

Its left-hand side equals $-(w - z) < 0$ at $\gamma = 0$ and tends to $+\infty$ as $\gamma \to \infty$, so it has a *unique positive root*. Monopsony power is thus point-identified from the observables $\{w, z, y, r\kappa, r, \lambda, p\}$: this is the precise sense in which monopsonistic search can be taken to aggregate data. Section 3 implements the identification on United States data, where the net surplus S is output net of the user cost of capital and of the employer wedge on the wage bill, $S = y_{net}/(1 + \tau^*)$, and the fixed entry cost κ is set to zero so that the whole recruiting cost is carried by the flow c.

2.8.4 Comparative Statics in Monopsony Power

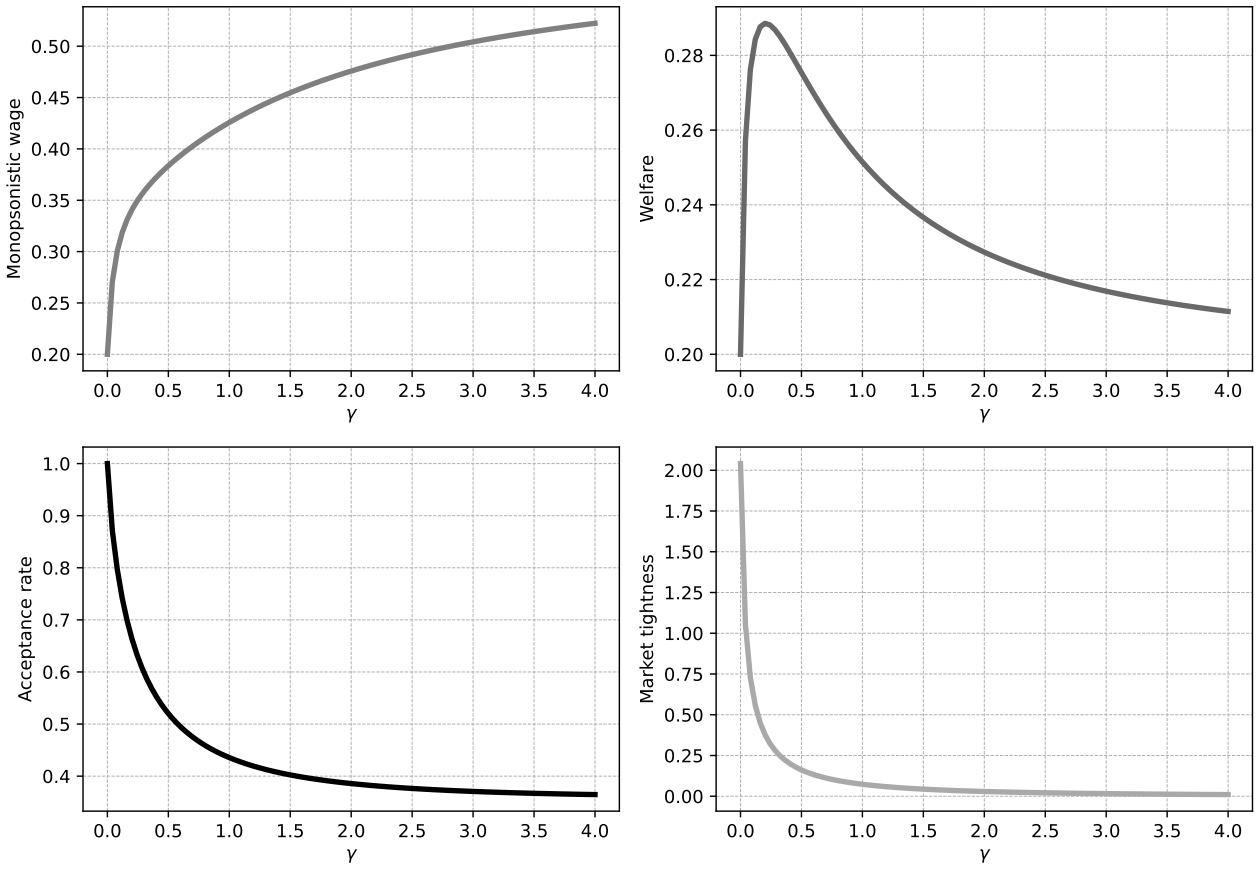
The elasticity parameter γ governs the responsiveness of workers' acceptance decisions to wages. A lower value of γ implies that acceptance probabilities are less sensitive to wage increases, thereby strengthening monopsony power. This interpretation is transparent from the wage formula (22): as γ decreases, wages place more weight on the worker's outside option and less on productivity.

Since γ governs monopsony power, we derive the comparative statics of the monopsonistic equilibrium with respect to it. The effect of γ on welfare is non-monotone.

The limiting cases can be derived analytically.

Proposition 2. *As $\gamma \to 0$ and monopsony power strengthens, (i) $rU \to z$, (ii) $w^* \to z$, (iii) the acceptance probability converges to one, and (iv) market tightness θ converges to a strictly positive constant.*

Figure 1: Comparative statics in monopsony power



Monopsonistic wage w^*, welfare rU, acceptance probability $F(x^*)$ and market tightness θ in the canonical model, against the acceptance elasticity γ; monopsony power is stronger at lower γ. As $\gamma \to 0$ the wage and welfare fall to z while almost every contact is accepted; as γ rises the wage approaches $y - r\kappa$, acceptance falls, tightness collapses and welfare falls back toward z. Welfare peaks near $\gamma = 0.2$. Parameter values: $c = 1$, $r = \lambda = 0.1$, $\alpha = 0.5$, $y = 1$, $z = 0.2$, $\kappa = 4$, $A = 1$, and $\gamma \in (0, 4]$.

Proposition 3. *As $\gamma \to \infty$ and monopsony power weakens, (i) $rU \to z$, (ii) $w^* \to y - r\kappa$, (iii) the acceptance probability converges to zero, and (iv) market tightness θ converges to zero.*

The proofs follow from observing that

$$\left(\frac{\gamma}{\gamma+1}\right)^{\gamma+1+\alpha} \frac{1}{\gamma} \sim \begin{cases} \gamma^{\alpha} & \gamma \to 0 \\ \frac{1}{e\gamma} & \gamma \to \infty \end{cases}$$

and that

$$\frac{y - r\kappa - rU}{\bar{x}} \in (0, 1) \ .$$

Figure 1 plots the monopsonistic wage, welfare, measured by the value of unemployment, the acceptance probability and market tightness against γ. It shows the two limits of Propositions 2 and 3 and the path between them. When γ is close to zero, almost all workers have

an amenity cost close to zero, so almost every contact is accepted even though the wage is close to z: workers find jobs quickly but gain little from them, and welfare is close to z. As γ rises, the wage rises toward the break-even wage $y - r\kappa$, but acceptance falls and firm entry collapses, so that market tightness is close to zero already at $\gamma = 1$ and welfare falls back toward z. Welfare is therefore hump-shaped in γ, peaking at an intermediate degree of monopsony power, around $\gamma = 0.2$ with these parameters, where workers earn a rent and enough contacts are accepted to sustain entry.

2.8.5 The Beveridge Curve Representation

The model admits a Beveridge curve which, with constant elasticities, reads

$$u = \frac{\lambda}{\lambda + A\left(\frac{v}{u}\right)^{1-\alpha}\mathcal{AR}}$$

where $\mathcal{AR}$ is the acceptance rate.

The acceptance probability acts much like the efficiency parameter of the matching function: any change in γ (monopsony power) affects the probability of acceptance and through it shifts the Beveridge curve. In particular, as depicted in panel (a) of Figure 2, an increase in monopsony power (a decrease in γ) shifts the Beveridge curve inward. This happens because, although wages fall when monopsony power is stronger, the equilibrium acceptance rate rises, so stronger monopsony power acts like higher matching efficiency.

Using free entry, we find the linear relation

$$v = u\left[\frac{A(y - w^* - r\kappa)\mathcal{AR}}{(r\kappa + c)(r + \lambda)}\right]^{\frac{1}{\alpha}} .$$

An increase in monopsony power affects this relationship in two ways: it increases net profits per worker and increases the acceptance rate. As a result, the job-creation ray rotates upward. Panel (b) of Figure 2 superimposes this job-creation ray on the Beveridge curve of panel (a); their intersection is the equilibrium.

2.8.6 Unemployment with Monopsonistic Search: Robinson meets DMP

In a standard search-and-matching model with homogeneous productivity, the relationship between unemployment and wages is monotonically increasing: higher wages reduce profits, but since the value of firms is anchored by free entry, market tightness must fall, increasing unemployment.

The classical intuition for monopsony, due to Robinson (1969), is that higher wages instead raise employment. However, the Robinson intuition is derived in partial equilibrium and abstracts from entry on the firm side.

In monopsonistic search, the result is nuanced. As Theorem 2 shows, raising the wage around the monopsonistic equilibrium may increase employment or reduce it, depending on parameter values. This reflects a combination of the Robinson and the standard search-and-matching intuitions, which push in opposite directions. Indeed, in the canonical model we

can plot wages and employment and derive the global relationship between the two (panel (c) of Figure 2).

In particular, we see that the relationship between wages and employment is inverse-U shaped: for low wages, the Robinson effect dominates and raising wages increases employment through rising acceptance, but only up to a certain wage level above which firm entry, and with it market tightness, declines so fast that employment decreases. Indeed, in monopsonistic search the relationship between wages and firm entry need not be monotone decreasing: for low wages, acceptance rates can be very low and discourage firm entry, as depicted in panel (d) of Figure 2.

The monopsonistic equilibrium may lie on either side of the peak. The optimal wage, instead, always lies to its right: by Proposition 1, a further wage increase at the optimum raises unemployment.

3 Calibration to US Data

This section takes the canonical model of Section 2.8 to United States data for 2022–2024. The model has a single separation rate: the rate at which a worker loses her job is the rate at which the firm loses the worker, λ, so the same hazard discounts the job value and enters the Beveridge curve (17). The United States is a natural first application, since turnover is high and employment protection is light, so the observed $E \to U$ rate is a good measure of how fast jobs end.

Section 3.1 presents the data and the two accounting choices on which the results depend, the price of capital and the treatment of the labor share. Section 3.2 reports the calibration and the efficient wage, and Section 3.3 explains what the calibrated recruiting cost measures and why it does not depend on the wage protocol.

3.1 Data and Accounting Conventions

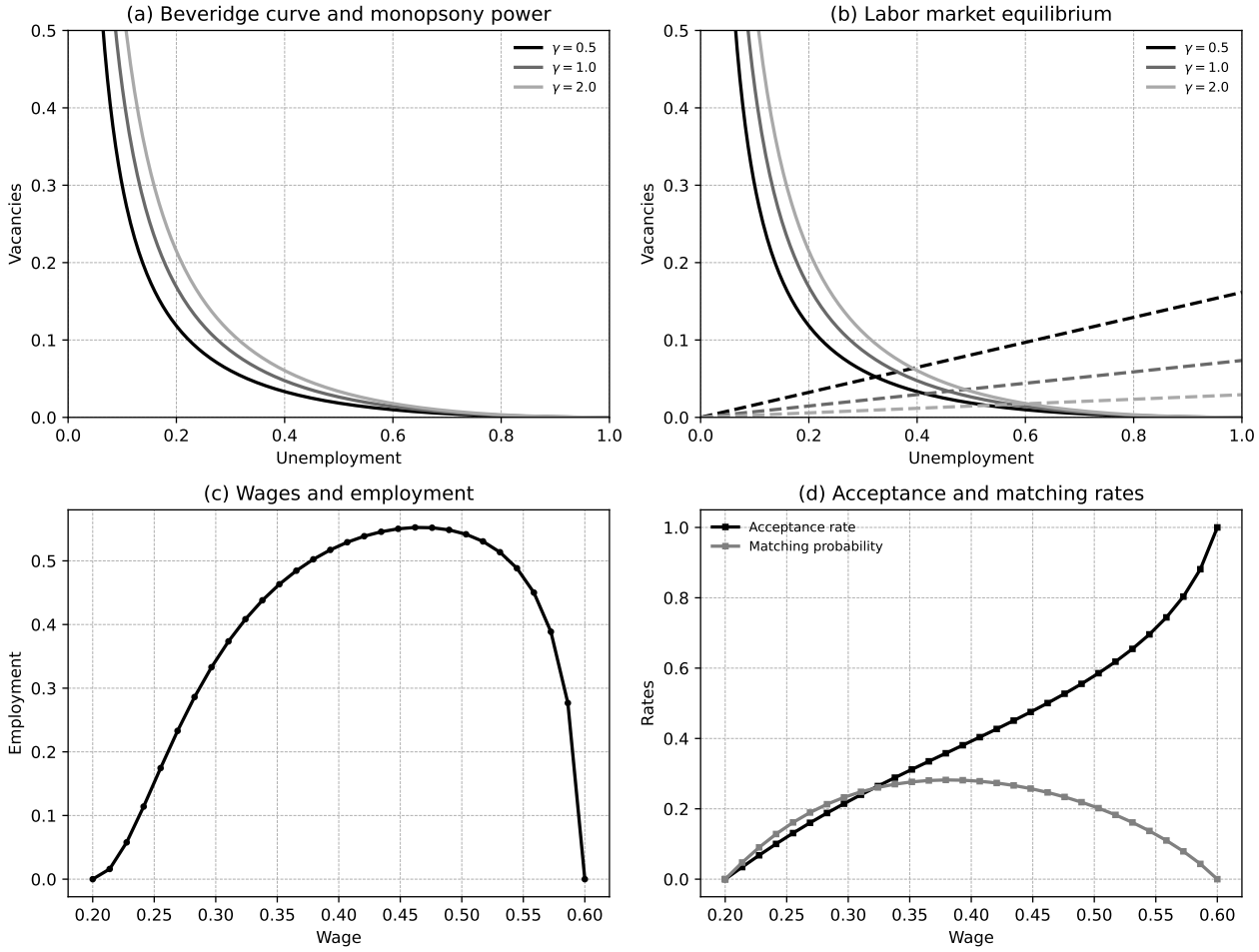
Table 1 collects the moments, averaged over 2022–2024. We normalize gross output per worker to $y = 1$ and express all flows at quarterly rates. Five choices deserve comment.

Output. The normalization uses GDP per worker, \$171,300 per year, computed as GDP divided by household-survey employment. Every level in the model is a ratio to this number.

The wage. We use the QCEW average annual wage of employment covered by unemployment insurance, \$71,667, so that $w = 0.418$. This administrative measure is the cash wage the worker is paid. It is not the firm's cost of a worker, which also includes payroll taxes and non-wage compensation; we account for both below. The BEA measure of wages and salaries per full-time equivalent, \$81,395 or $w = 0.475$, serves as a robustness check.

The outside option. We set z to the average weekly unemployment insurance benefit, annualized, about \$20,167, so that $z = 0.118$. A larger z narrows the worker's rent, lowers the identified γ and so raises measured monopsony power. The low ratio of z to y reflects that unemployment insurance in the United States is much less generous than in continental Europe.

Figure 2: Simulations of the canonical model



(a) Beveridge curves for three values of the acceptance elasticity γ: the curve shifts inward as monopsony power rises (γ falls), because the equilibrium acceptance rate rises and acts like higher matching efficiency. (b) The same curves with the job-creation rays (dashed), which are steeper when monopsony power is stronger, through both higher profits per worker and a higher acceptance rate. (c) Employment against the posted wage for $\gamma = 1$: the relation is inverse-U shaped, as the Robinson effect dominates at low wages and the entry effect at high ones. (d) Acceptance and matching probabilities against the posted wage for $\gamma = 1$: acceptance rises with the wage while the matching probability is hump-shaped, so firm entry need not fall as the wage rises. Parameter values: $c = 1$, $r = \lambda = 0.1$, $\alpha = 0.5$, $y = 1$, $z = 0.2$, $\kappa = 4$, $A = 1$; panels (a) and (b) use $\gamma \in \{0.5, 1, 2\}$.

Table 1: Labor market data for the United States, 2022–2024

Variable	2022	2023	2024	Average	Unit
Gross wage, QCEW (w)	69,400	71,500	74,100	71,667	USD/year
Gross wage, BEA per FTE	78,625	81,359	84,200	81,395	USD/year
Employment (N)	158.300	161.000	161.500	160.267	millions
GDP per capita (y^{pc})	77,200	81,700	86,150	81,683	USD/year
Population (P)	333.300	334.900	340.100	336.100	millions
GDP per worker (y)	162,600	169,900	181,400	171,300	USD/year
Flow value of unempl. (z, UI)	19,500	20,000	21,000	20,167	USD/year
Labor share (s_L)[a]	58.90	58.10	58.30	58.43	% of output
Job-openings rate (JOLTS)	6.80	5.60	4.70	5.70	%
Unemployment rate (u)	3.65	3.64	4.11	3.80	%
$E \to U$ rate (quarterly)	2.90	2.90	2.90	2.90	%

[a]Bureau of Labor Statistics series MPU4910141, *Private Nonfarm Business Sector: Labor Share*, annual. Labor compensation—wages and salaries *plus* employer-paid non-wage compensation and employer payroll contributions—as a fraction of sector output. This is the moment the effective employer wedge τ^* is set to reproduce, and the calibration targets the 2022–24 average $s_L = 0.584$; see Section 3.1.

Flows and tightness. The quarterly $E \to U$ rate is 0.029 and the unemployment rate averages 3.80 percent, so the steady-state Beveridge relation (17) gives a job-finding rate of $p = \lambda(1-u)/u = 0.745$ per quarter. We build market tightness from the JOLTS job-openings rate, 5.70 percent of employment plus openings, and the unemployment rate, which gives $\theta = V/U = 1.530$. The window is unusually tight: tightness falls from 1.93 in 2022 to 1.15 in 2024, against a long-run United States ratio of 0.7 to 1.2. Tightness affects only the recruiting cost, which falls as θ rises, so the tight window makes our recruiting cost low rather than high. The identified γ and the efficiency premium do not depend on θ.

The labor share. We take s_L from the Bureau of Labor Statistics series *Private Nonfarm Business Sector: Labor Share* (MPU4910141), which measures labor compensation, that is, wages and salaries plus employer-paid non-wage compensation and employer payroll contributions, as a share of sector output. It reads 0.589, 0.581 and 0.583 over the three years, and we target the average, $s_L = 0.584$. A nonfarm business share fits a model in which every worker holds a market job producing y; a corporate-sector share would be a few points lower. The share is flat over the window, so the choice of year does not matter, although it has fallen by about three points since the mid-2010s.

Two more accounting choices affect the size of the firm's rent: the price of capital and the total cost of a job to the firm. We discuss them in turn.

Labor is the only factor in the model, so we net capital out of measured output. We price capital at $r = 7\%$ per year, the firm's cost of capital coherently with the long run return on the S&P500 index, and use the same rate to discount. With a depreciation rate $\delta_K = 0.07$ per year and a productive, nonresidential business, capital–output ratio $K/Y = 2.30$, capital's

share of gross output is

$$s_K \equiv (r + \delta_K)k = 0.321, \qquad y_{net} \equiv y - s_K = 0.679,$$

where k is capital per worker in quarterly output units. Net output y_{net}, output per worker net of the user cost of capital, is the productivity term y of the model of Section 2. It does not net out payroll taxes, which enter through the wedge below.

The firm's cost of a worker exceeds the QCEW wage w: the firm also pays the statutory payroll cost and non-wage compensation, such as health insurance, pension contributions and paid leave, which the QCEW wage omits. We capture non-wage compensation with a compensation wedge ϱ, so that the firm's labor cost per worker, which is what the labor share measures, is

$$s_L \equiv (1 + \tau_{\text{pay}} + \varrho)\, w \equiv (1 + \tau^*)\, w, \tag{27}$$

and we set it equal to the labor share, $s_L = 0.584$. With a statutory employer payroll cost $\tau_{\text{pay}} = 0.10$, made of FICA at 7.65 percent plus state unemployment insurance, FUTA and workers' compensation, this implies $\varrho = 0.297$ and an effective wedge $\tau^* = 0.397$. The wage grossed up by the payroll cost alone gives only $(1+\tau_{\text{pay}})w = 0.46$, well below the labor share. The gap is the non-wage compensation that the QCEW wage misses, and a wedge of about 30 percent of the cash wage is in line with the employer cost of health insurance, retirement contributions and paid leave in the Employer Costs for Employee Compensation series. We write the wedge additively, $\tau^* = \tau_{\text{pay}} + \varrho$, because its two parts differ: the first is a tax paid to the government, the second is compensation paid to the worker.

National accounts split gross output, normalized to $y = 1$, into labor compensation s_L and gross operating surplus, $\text{GOS} = 1 - s_L = 0.416$, a residual that lumps together distinct items. Netting out capital's share $s_K = (r + \delta_K)k$, priced from the data rather than read off the surplus, leaves

$$\text{GOS} - s_K = a - s_L = y_{net} - s_L = 0.679 - 0.584 = 0.095 \equiv \pi, \tag{28}$$

where a is labor's output elasticity. With constant returns and no fixed factor, $a = 1 - s_K = y_{net} = 0.679$: the marginal product of labor equals net output per worker, and the whole residual is the flow profit π that free entry capitalizes into the job value $J = \pi/(r + \lambda)$. Labor is paid $s_L = 0.584$ against a marginal product of 0.679, a markdown of $s_L/a = 0.860$: the monopsonist keeps about a seventh of the marginal product, which the accounts record as capital income or profit. The measured labor share is therefore compensation paid, not the marginal product of labor.

The size of the pool depends on both accounting choices. Pricing capital at 3 rather than 7 percent would lower s_K from 0.32 to 0.23 and raise π from 0.095 to 0.19. A larger pool means a larger job value, and hence a larger recruiting cost, but also a larger rent for the planner to reallocate, and hence a larger efficiency premium. The labor share moves the pool one for one, so a corporate-sector share a few points lower would raise the pool, the recruiting cost and the efficiency premium. The wage measure, instead, does not matter: at the same labor share, the BEA wage of 0.475 would imply a smaller wedge and the same π, since by (28) the pool depends only on the total labor share and not on how it is split

between wage and non-wage compensation. The pool is the quantity on which everything that follows depends, and the cost of capital and the labor share fix it.

The two parts of τ^* must be treated differently on the worker's side of the model. The payroll tax $\tau_{\text{pay}} w$ goes to the government and is no part of the worker's compensation, as is standard. Non-wage compensation $\varrho\, w$, instead, is health insurance, retirement contributions and paid leave, which the worker receives. Since we introduce ϱ because the QCEW wage understates what a job pays, the same correction must apply wherever compensation enters the model, including the worker's asset equations. Otherwise $\varrho\, w = 0.124$, about 18 percent of net output, would be a cost to the firm that accrues to nobody: not to profit, not to the government and not to the worker.

We therefore let a fraction $\psi \in [0, 1]$ of the compensation wedge accrue to the worker and define total compensation

$$\tilde{w} \equiv (1 + \psi\varrho)\, w.$$

We set $\psi = 1$ throughout, so that the worker values non-wage compensation at its cost to the firm:

$$\tilde{w} = (1 + \varrho)\, w = 0.542.$$

The worker's asset equations of Section 2 are then written in total compensation rather than in the cash wage,

$$(r + \lambda)\, W(\tilde{w}, x) = \tilde{w} - x + \lambda U, \qquad rU = \tilde{z} + \theta q(\theta) \int_0^{x^*} [W(\tilde{w}, x) - U]\, dF(x), \tag{29}$$

with reservation cost $x^* = \tilde{w} - rU$, and with the outside option measured on the same basis,

$$\tilde{z} \equiv (1 + \psi\varrho)\, z = 0.153.$$

Treating z in the same way as the wage matters. The non-employed also receive value that the cash benefit misses, such as health coverage obtained outside employment and home production, and raising the wage while leaving the outside option at the cash benefit would widen the worker's rent, and so lower measured monopsony power, through a change of units alone.

This has three consequences. First, the compensation wedge drops out of the model. Since $\tau^* = \tau_{\text{pay}} + \varrho$,

$$(1 + \tilde{\tau})\, \tilde{w} = (1 + \tau^*)\, w = s_L, \qquad \tilde{\tau} \equiv \frac{\tau_{\text{pay}}}{1 + \psi\varrho} = 0.077, \tag{30}$$

exactly: on total compensation, the firm's only wedge is the statutory payroll cost, rescaled. The wedge ϱ survives only as the factor that converts the QCEW wage into the model's wage. The firm's problem is then summarized by

$$\pi = y_{net} - (1 + \tilde{\tau})\, \tilde{w} = y_{net} - s_L, \qquad \tilde{S} \equiv \frac{y_{net}}{1 + \tilde{\tau}} = 0.631, \tag{31}$$

so that $\pi = (1 + \tilde{\tau})(\tilde{S} - \tilde{w})$. The break-even wage $\tilde{S}$ is the marginal product expressed in wage units, and it plays the role of $y - r\kappa$ in the wage equation (22) and in the identification

quadratic (26). Second, the identification does not depend on ψ. Every coefficient of the quadratic (26) is homogeneous of degree one in a common scaling of $(\tilde{S}, \tilde{w}, \tilde{z})$, so γ, the markdown and the efficiency premium are the same for any ψ, as long as the outside option is scaled with the wage. What ψ changes are levels: $\tilde{w}$, $\tilde{z}$, rU, x^* and the Nash share of Appendix D. Third, the fiscal analysis of Section 4 does depend on ψ, because the government taxes the cash wage while paying a benefit valued at $\tilde{z}$; we return to this there.

3.2 Calibration Results

Our main quantitative result is that the monopsonistic wage is too low by an economically meaningful amount: the wage that maximizes welfare, measured by the value of search, is 10 percent above the monopsonistic wage and raises welfare by 5.6 percent.

Table 2 reports the calibrated parameters. Calibrating the model, we find that the inverse acceptance elasticity is $1/\gamma = 2.9$, so the firm sets the wage much closer to the worker's outside option than to productivity, and total compensation is 14 percent below the marginal product in wage units. By (30), this markdown equals one minus the ratio of the labor share to net output, so it depends neither on the wage measure nor on ψ. It lies at the conservative end of empirical estimates (Yeh et al., 2022).

Table 3 compares the monopsonistic and the efficient allocations. The welfare gain comes from the acceptance margin: a higher wage turns more meetings into matches. It is limited by the entry margin: a higher wage squeezes profits and discourages vacancy creation, so that at the efficient wage unemployment is 6.4 percent instead of 3.8. At the efficient wage the two margins offset each other, and by Proposition 1 any further increase raises unemployment.

Figure 3 plots welfare and unemployment as the wage rises from the monopsonistic level, through the efficient wage and beyond. It shows that most of the welfare gain is cheap and the rest is expensive. Welfare rises steeply at first and is flat near the optimum (panel a), while unemployment barely moves at first and then rises faster and faster, until entry stops at the break-even wage (panel b). A wage increase of 5 percent already delivers about two thirds of the welfare gain at a small employment cost, while the last third costs more employment than the first two thirds. Beyond the optimum, welfare falls quickly below its initial level.

3.3 Wage Protocols and the Calibrated Cost of Posting a Vacancy

Panel C of Table 2 reports the flow recruiting cost, $c = 0.989$, or 1.46 times quarterly net output per worker. Since a vacancy takes $1/[q(\theta)F(x^*)] = 2.05$ quarters to fill, the expected recruiting outlay per hire is $J = 2.03$, or 2.99 quarters of net output, a large number for any recruiting technology. This subsection explains where the number comes from, and shows that it does not depend on our specific model of monopsony and wage setting.

Two equations determine c, and neither involves wage setting. The first is free entry. A filled job yields the flow profit π of (31) and is destroyed at rate λ, so the firm's Bellman equation $rJ = \pi - \lambda J$ gives

$$\frac{c}{q(\theta)F(x^*)} \;=\; J \;=\; \frac{\pi}{r+\lambda}. \tag{32}$$

Table 2: Calibration of the canonical model, United States, 2022–2024

Object	Description	Value
Panel A. Exogenous parameters and data moments		
y	Gross output per worker (normalization)	1.000
r	Discount rate (annual 7%)	0.0175
δ_K	Capital depreciation (annual)	0.070
K/Y	Productive capital–output ratio (annual)	2.300
$(r+\delta_K)k$	User cost of capital, s_K	0.321
y_{net}	Net flow product $y-(r+\delta_K)k$	0.679
α	Matching function elasticity	0.500
w	Average gross wage (QCEW)	0.418
z	Flow value of non-employment	0.118
τ_{pay}	Statutory employer payroll cost	0.100
ϱ	Employer non-wage compensation wedge	0.297
τ^*	Effective employer wedge $\tau_{\text{pay}}+\varrho$	0.397
$\tilde{w}$	Total compensation $(1+\varrho)w$	0.542
$\tilde{z}$	Non-employment value, same basis $(1+\varrho)z$	0.153
$\tilde{\tau}$	Payroll cost on compensation $\tau_{\text{pay}}/(1+\varrho)$	0.077
$(1+\tau^*)w$	Firm labor cost = labor share (target)	0.584
κ	Fixed entry cost (value of a vacancy)	0.000
λ	Separation rate, worker *and* firm	0.0294
p	Job-finding rate (Beveridge)	0.745
θ	Market tightness V/U	1.530
u	Unemployment rate	3.80%
Panel B. Identification of monopsony power		
$\tilde{S}$	Break-even wage $y_{net}/(1+\tilde{\tau})$	0.631
γ	Acceptance elasticity (positive root)	0.344
$1/\gamma$	Inverse acceptance elasticity	2.91
rU	Flow value of unemployment	0.512
x^*	Reservation cost $\tilde{w}-rU$	0.030
$\bar{x}$	Amenity ceiling $\tilde{S}-\tilde{z}$	0.478
$F(x^*)$	Acceptance probability	0.387
A	Matching efficiency	1.555
$q(\theta)$	Contact rate per vacancy	1.257
$q(\theta)F(x^*)$	Vacancy-filling rate $= p/\theta$	0.487
$1-\tilde{w}/\tilde{S}$	Wage markdown below marginal product	14.0%
Panel C. Free entry and the recruiting cost		
π	Flow profit $y_{net}-(1+\tau^*)w$	0.095
J	Job value $\pi/(r+\lambda)$	2.032
c	Flow recruiting cost $q(\theta)F(x^*)J$	0.989

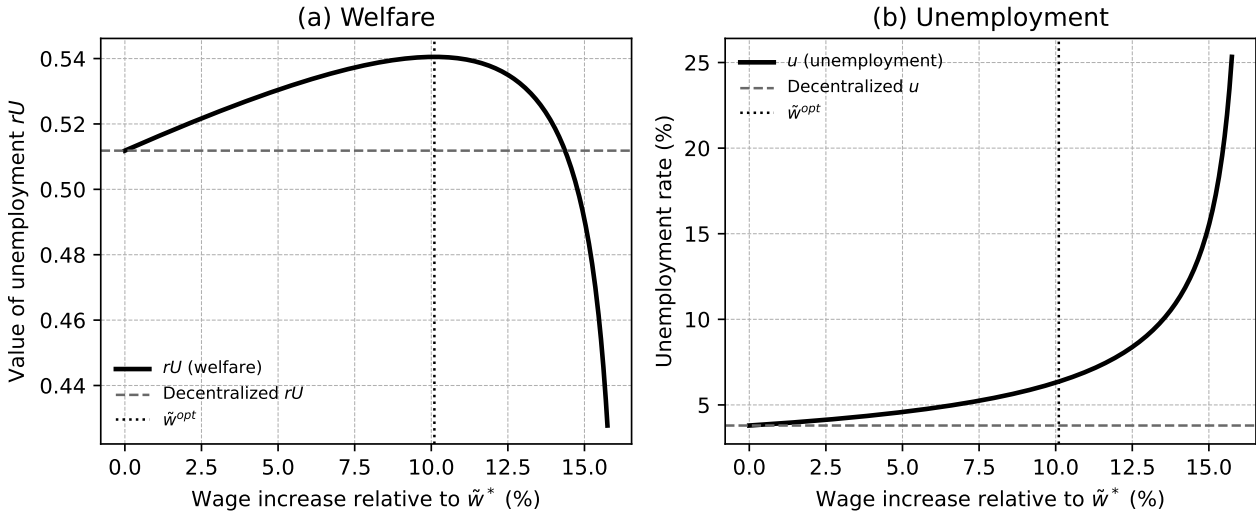
Gross output per worker is normalized to $y=1$ and all flows are per quarter. Panel A collects the exogenous parameters and the data moments of Table 1; τ^* is the only one that is set rather than observed, and it is set so that $(1+\tau^*)w$ equals the aggregate labor share. Panel B implements the identification of Section 2.8.3: γ is the unique positive root of the quadratic (26) with $\tilde{S}=y_{net}/(1+\tilde{\tau})$, and rU, x^* and $F(x^*)$ follow, with $\bar{x}=\tilde{S}-\tilde{z}$. Wage-like quantities in Panel B are in total compensation, $\tilde{w}=(1+\varrho)w$, as in the text. Panel C is the free-entry block: $c=q(\theta)F(x^*)J$ with $q(\theta)F(x^*)=p/\theta$ read directly off the data, and with the fixed entry cost set to $\kappa=0$ so that the whole cost of recruiting is carried by the flow c. The welfare-maximizing wage implied by this calibration is reported separately in Table 3.

Table 3: The efficient wage and its employment cost, United States, 2022–2024

	$\tilde{w}^*$ (decentralized)	$\tilde{w}^{opt}$ (efficient)	change
Total compensation $\tilde{w}$	0.5422	0.5970	+10.10%
Value of search rU	0.5118	0.5405	+5.61%
Unemployment rate u	3.80%	6.37%	+2.57 pp
Market tightness θ	1.530	0.337	−78.0%
Acceptance probability $F(x^*)$	0.387	0.480	+9.2 pp
Job-finding rate p	0.745	0.433	−41.9%
Flow profit π	0.0951	0.0360	−62.1%

The efficient wage maximizes the value of search, $\tilde{w}^{opt} = \arg\max_{\tilde{w}} rU(\tilde{w})$, holding the calibrated primitives $(\gamma, \bar{x}, A, c, \lambda)$ of Table 2 fixed; free entry then determines tightness and the Beveridge curve determines unemployment. A higher wage raises welfare through the acceptance margin—$F(x^*)$ rises by nine percentage points—and lowers it through entry, as the flow profit is squeezed and vacancy creation collapses. At $\tilde{w}^{opt}$ the two exactly offset at the margin, and by Proposition 1 any further wage increase strictly raises unemployment. Section 4 reports the fiscal-adjusted counterpart, $\tilde{w}^{opt}_{\text{fisc}} = 0.5935$ at $\mu = 1.25$.

Figure 3: Welfare and unemployment along the wage path



United States calibration. The horizontal axis is the wage increase relative to the decentralized monopsonistic wage $\tilde{w}^*$. In each panel the dashed line marks the decentralized level and the dotted vertical line the welfare-maximizing wage $\tilde{w}^{opt}$, 10.1 percent above $\tilde{w}^*$. Computed holding the calibrated primitives $(\gamma, \bar{x}, A, c, \lambda)$ fixed, with free entry determining tightness and the Beveridge curve determining unemployment.

The second is the steady-state Beveridge relation (17): the job-finding rate is $p = \lambda(1-u)/u$, so the vacancy-filling rate is

$$q(\theta)F(x^*) \;=\; \frac{p}{\theta} \;=\; \frac{\lambda(1-u)}{u\,\theta}, \tag{33}$$

which depends only on the observed tightness, unemployment and separation rate. Structurally, $p = A\theta^{1-\alpha}F(x^*)$ is a property of the matching technology, and (33) reads it off the data. Combining (32) and (33),

$$c \;=\; q(\theta)F(x^*)\,J \;=\; \frac{p}{\theta}\cdot\frac{\pi}{r+\lambda} \tag{34}$$

The calibrated c is therefore the product of a filling rate read off the data and a job value imputed from the profit flow. It is not a measured recruiting technology, but the value that rationalizes the observed tightness given the rents and turnover.

Appendix D establishes two implications of (34). First, c is invariant to the wage protocol: when the wage is calibrated to the data, Nash bargaining and monopsonistic wage setting deliver the same $c = 0.989$, because each uses the observed wage in π and the same filling rate. The protocols differ only in the second parameter each must impute, none under a fixed wage, the Nash share $\beta = 0.196$ identified linearly, and the acceptance elasticity $\gamma = 0.344$ identified by the quadratic, and in whether that second step is linear or quadratic. The wage protocol therefore cannot make c larger or smaller, and a large calibrated c is evidence neither for nor against monopsony. Third, since the filling rate comes from the data and π from the accounting choices of Section 3.1, the only remaining determinant of c is the effective discount rate $r + \lambda$, that is, firm turnover. At $\lambda = 0.029$ per quarter, a job lasts about twenty-one quarters in discounted terms, and any margin along which a firm loses a productive worker faster than the worker becomes unemployed would lower J and c in proportion. The model of this paper, since it abstract from on-the-job search, has no such margin, and we leave it to future work.

4 Fiscal Externalities and the Optimal Wage

The wage a monopsonistic firm posts affects the public budget, and the firm does not internalize this effect. A lower wage narrows the payroll tax base directly, and through free entry it changes employment, and with it both the tax base and the number of unemployed workers who claim benefits. Neither effect enters the firm's first-order condition (11). This section adds the public budget to the planner's problem. With a standard marginal value of public funds, internalizing the budget lowers the efficient wage premium from 10.1 to 9.5 percent, and across all the values and tax bases we consider the premium stays between 9.2 and 9.8 percent.

4.1 The Government Budget

The government taxes the wage bill of the employed and pays the benefit $\tilde{z}$ to the unemployed, so net public revenue per capita in steady state is

$$T(\tilde{w}) \;=\; \tilde{\tau}\,\tilde{w}\,(1-u) \;-\; \tilde{z}\,u, \tag{35}$$

where u is the equilibrium unemployment rate.

The tax rate in (35) is the statutory employer payroll cost, not the whole wedge on the QCEW wage. Of the firm's labor cost, only $\tau_{\text{pay}} w$ goes to the government. Non-wage compensation $\varrho\, w$, such as health insurance and retirement contributions, goes to the worker and never enters the budget, and counting it as a tax would multiply the fiscal stakes by four. This is why the wedge is written additively in (27). Since the payroll tax is levied on the cash wage $w = \tilde{w}/(1+\psi\varrho)$, the rate on total compensation is $\tilde{\tau} = \tau_{\text{pay}}/(1+\psi\varrho) = 0.077$ of (30), and receipts satisfy $\tilde{\tau}\,\tilde{w} = \tau_{\text{pay}} w$ whatever the wage concept.

The benefit outlay is valued at $\tilde{z}$, the flow value that enters the worker's asset equation (29), rather than at the cash benefit z: the flow value of a non-employed worker should be the same in her Bellman equation and in the budget. This makes the benefit rolls more expensive and the employment effect below larger, so it makes the fiscal correction to the efficient wage larger rather than smaller. It also overstates public outlays, since part of $\tilde{z}$ is publicly provided, such as Medicaid and marketplace subsidies, and part is not, such as coverage through a spouse or home production. With a single separation rate, unemployment is the only non-employed state, so the externality we measure comes from wage setting alone, through the tax base and employment.

At the calibrated equilibrium, payroll receipts are $\tilde{\tau}\,\tilde{w}(1-u) = 0.0402$ per worker and benefit outlays are $\tilde{z}\,u = 0.0058$, so net revenue is $T(\tilde{w}^*) = 0.0344$, 3.44 percent of output. The low statutory payroll rate keeps this number small by European standards, and it shapes the results below.

The marginal social value of public funds is $\mu \geq 1$, with $\mu = 1$ under lump-sum rebates and $\mu > 1$ when taxation is distortionary; $\mu \approx 1.25$ is a standard estimate. The planner maximizes fiscal-adjusted welfare

$$\mathcal{W}(\tilde{w}) \;=\; rU(\tilde{w}) + \mu\, T(\tilde{w}), \tag{36}$$

with first-order condition $\partial rU/\partial\tilde{w} + \mu\,\partial T/\partial\tilde{w} = 0$.

The derivative of revenue with respect to the wage has two terms,

$$\frac{\partial T}{\partial \tilde{w}} = \underbrace{\tilde{\tau}\,(1-u)}_{\text{tax-base expansion}} \;-\; \underbrace{(\tilde{\tau}\,\tilde{w} + \tilde{z})\,\frac{\partial u}{\partial \tilde{w}}}_{\text{employment effect}}. \tag{37}$$

The first is the mechanical expansion of the payroll base: at fixed stocks, a higher wage raises receipts in proportion. The second is the employment effect: a higher wage lowers job creation through free entry, shrinking the taxed stock and raising the benefit rolls, and each displaced worker costs the budget $\tilde{\tau}\,\tilde{w} + \tilde{z}$, the tax forgone plus the benefit paid. Setting

$\partial T/\partial \tilde{w} = 0$ defines a critical payroll rate

$$\bar{\tau} \;=\; \frac{\tilde{z}\,\partial u/\partial \tilde{w}}{(1-u) - \tilde{w}\,\partial u/\partial \tilde{w}},$$

below which wage increases lower revenue and above which they raise it. Like the markdown, $\bar{\tau}$ does not depend on the wage concept, since a common scaling of $\tilde{w}$ and $\tilde{z}$ leaves its numerator and denominator unchanged. At the calibrated equilibrium $\partial u/\partial \tilde{w} = 0.214$ and $\bar{\tau} = 0.039$. The United States rate, $\tilde{\tau} = 0.077$, is twice as large, so revenue rises with the wage at the monopsonistic equilibrium. The comparison is local, however: as the wage rises, u rises and $\partial u/\partial \tilde{w}$ steepens, so $\bar{\tau}$ rises toward $\tilde{\tau}$ and then exceeds it.

4.2 Quantitative Results

Table 4 evaluates the decomposition along the wage path, and Figure 4 plots revenue next to welfare and unemployment. Revenue is hump-shaped in the wage, with an early and shallow peak: it reaches its maximum 4.6 percent above the monopsonistic wage, where it is 1.35 percent above its baseline. Beyond the peak the employment effect dominates, and at the efficient wage revenue is 3.0 percent below its baseline. Between the monopsonistic and the efficient wage, the expansion of the tax base and the employment loss nearly cancel, and revenue is almost flat (panel c of Figure 4).

Panel B of Table 4 reports the fiscal-adjusted optimal wage $\tilde{w}^{opt}_{\text{fisc}}$, which maximizes (36); the dash-dotted line in panel (c) of Figure 4 marks it. With $\mu = 1.25$, the efficient wage is 9.5 percent above the monopsonistic wage, against 10.1 percent without the budget: since revenue falls with the wage near the efficient wage, internalizing the budget lowers the optimum. The correction grows with μ: as μ rises from 1 to 2, the premium falls from 9.6 to 9.2 percent.

The result could depend on counting only the employer payroll cost as public revenue. A government that also collects employee-side FICA, at 7.65 percent, and income tax, at an average rate of 15 percent, has a tax rate of 0.327 on the cash wage, or $\tilde{\tau}_{\text{broad}} = 0.252$ on total compensation, well above $\bar{\tau}$. For that government a higher wage raises revenue: at the efficient wage revenue is 4.4 percent above its baseline, rather than 3.0 percent below. Yet the fiscal-adjusted optimum barely changes, to 9.8 percent at $\mu = 1.25$ (Panel B, right-hand columns). The optimum depends on the marginal effect of the wage on revenue near the efficient wage, and by (37) this is a difference between two terms that both scale with the tax rate; the level of revenue, which changes a lot, does not enter the first-order condition. Whatever the tax base, the fiscal correction to the efficient wage premium is below one percentage point.

The fiscal externality of monopsonistic wage setting is small in the United States because the statutory payroll rate is low. At $\tilde{\tau} = 0.077$ the rate is about twice the critical rate $\bar{\tau} = 0.039$, so the expansion of the tax base and the employment effect are of similar size and nearly cancel. In an economy with high payroll taxes the statutory rate would be far above $\bar{\tau}$, the expansion of the tax base would dominate over the whole relevant range, and fiscal considerations would strengthen the case for higher wages. The sign of the fiscal correction therefore depends on institutions, not on monopsony.

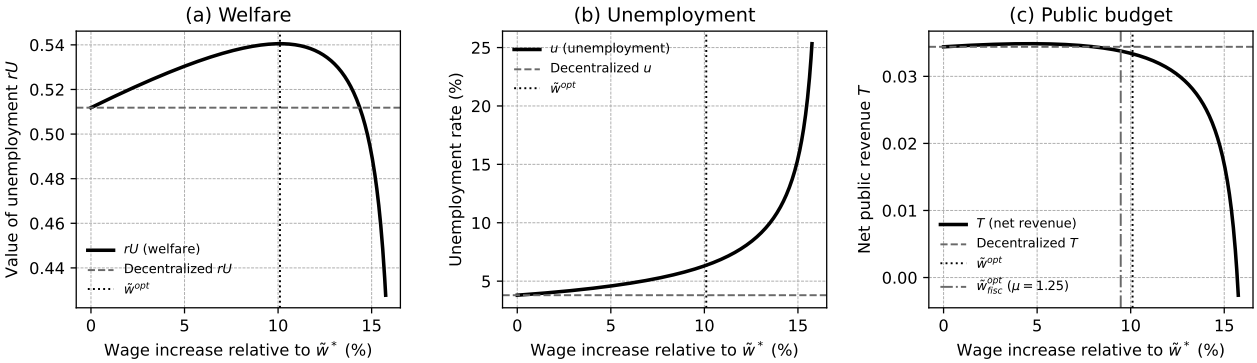
Table 4: Fiscal externalities and the optimal wage, United States

	u	T	$\Delta T/T$	mechanical	employment
Panel A. Decomposition of the revenue effect					
$\tilde{w}^*$ (baseline)	3.80%	0.0344	—	—	—
$\tilde{w}^* \times 1.05$	4.59%	0.0349	+1.33%	+0.0020	−0.0016
$\tilde{w}^{\text{Laffer}} = \tilde{w}^* \times 1.046$	4.50%	0.0349	+1.35%	+0.0019	−0.0014
$\tilde{w}^* \times 1.10$	6.31%	0.0334	−2.82%	+0.0040	−0.0050
$\tilde{w}^{opt} = \tilde{w}^* \times 1.101$	6.37%	0.0334	−3.03%	+0.0041	−0.0051

Panel B. Fiscal-adjusted optimal wage				
μ	payroll base $\tau_{\text{pay}} = 0.100$		broad base $\tau_{\text{broad}} = 0.327$	
	$\tilde{w}^{opt}_{\text{fisc}}$	premium	$\tilde{w}^{opt}_{\text{fisc}}$	premium
1.00	0.5941	+9.58%	0.5955	+9.83%
1.25	0.5935	+9.47%	0.5952	+9.78%
1.50	0.5929	+9.36%	0.5950	+9.73%
2.00	0.5919	+9.16%	0.5945	+9.65%
memo: pure welfare optimum	0.5970		+10.10%	

Baseline revenue is $T(\tilde{w}^*) = 0.0344$ per worker, 3.44 percent of output. Panel A evaluates $T(\tilde{w}) = \tilde{\tau}\,\tilde{w}\,(1-u) - \tilde{z}\,u$ along the counterfactual wage path, holding $(\gamma, \bar{x}, A, c, \lambda)$ fixed and letting free entry determine θ and the Beveridge curve determine u. The *mechanical* column is $\tilde{\tau}(\tilde{w} - \tilde{w}^*)(1-u^*)$, the expansion of the payroll base at unchanged stocks; the *employment* column is $-(\tilde{\tau}\tilde{w} + \tilde{z})(u - u^*)$, the loss of base plus the extra benefit outlay. Panel B reports $\tilde{w}^{opt}_{\text{fisc}} = \arg\max_{\tilde{w}}\{rU(\tilde{w}) + \mu T(\tilde{w})\}$. As in the text, wages are in total compensation, $\tilde{w} = (1+\varrho)w$, the benefit outlay is valued at $\tilde{z}$, and welfare is rU in the same units. The *payroll base* columns count only the statutory employer payroll cost as public revenue; the *broad base* columns add employee-side FICA (0.0765) and the average employee income tax rate (0.150), giving $\tau_{\text{broad}} = 0.327$.

Figure 4: Welfare, unemployment and the public budget along the wage path



United States calibration. The horizontal axis is the wage increase relative to the decentralized monopsonistic wage $\tilde{w}^*$. In each panel the dashed line marks the decentralized level and the dotted vertical line the welfare-maximizing wage $\tilde{w}^{opt}$, 10.1 percent above $\tilde{w}^*$; in panel (c) the dash-dotted line marks the fiscal-adjusted optimum of Section 4. Computed holding the calibrated primitives $(\gamma, \bar{x}, A, c, \lambda)$ fixed, with free entry determining tightness and the Beveridge curve determining unemployment.

For the main question of the paper, the fiscal correction does not change the answer. The efficient wage premium of Section 3.2 is 10.1 percent before fiscal considerations, 9.5 percent after them at $\mu = 1.25$, and between 9.2 and 9.8 percent for every value of μ and tax base we consider. The distortion we identify lies in entry and wage setting, and the public budget neither aggravates nor offsets it to first order.

5 Conclusion

This paper has developed a search-and-matching model of monopsony. When a worker meets a firm, she draws a match-specific amenity cost that she observes and the firm does not, so a firm that pays more is more likely to have its offers accepted. The model thus delivers an upward-sloping labor supply curve to the firm while keeping a matching function, free entry and a Beveridge curve.

The decentralized equilibrium is inefficient. Firms trade off the probability that a worker accepts against the surplus from a filled job, ignoring both the worker's surplus and the effect of wages on vacancy creation. On balance, wages are too low: a constrained planner facing the same matching technology and the same free-entry condition sets a higher wage. The wedge between the monopsonistic and optimal wages is governed by three elasticities, and it closes only in the limiting case in which the Hosios condition holds.

When amenities follow a constant-elasticity distribution, monopsony power can be measured from aggregate data alone. Calibrated to the United States for 2022–2024, the model implies an inverse acceptance elasticity of 2.9 and a markdown of 14 percent. The welfare-maximizing wage is 10 percent above the market wage and raises welfare by 5.6 percent, but it also raises unemployment from 3.8 to 6.4 percent. Accounting for the effect of wages on the public budget lowers the efficient wage premium to 9.5 percent.

The calibration also shows that the recruiting cost implied by free entry does not depend on how wages are set: fixed wages, Nash bargaining and monopsonistic wage setting deliver the same cost. A large calibrated recruiting cost therefore reflects how long jobs last, not how much market power firms have.

Several extensions are natural. Job-to-job flows would let productive jobs end faster than workers return to unemployment, lowering the calibrated recruiting cost, and comparing the efficiency premium across economies with high and low employment protection would isolate the role of employment protection. Firm heterogeneity would generate a wage distribution and let the model speak to wage inequality. Finally, since the identification requires only aggregate moments, it can be applied across countries and sectors to compare monopsony power across them.

References

Amodio, F. and N. De Roux (2024). Measuring labor market power in developing countries: Evidence from colombian plants. *Journal of Labor Economics 42*, 949–977.

Armangué-Jubert, T., T. Rapone, and A. Ruggieri (2026). Firm dynamics, monopsony, and aggregate productivity differences. *Review of Economic Dynamics 59*, 101231.

Azar, J., S. T. Berry, and I. Marinescu (2022). Estimating labor market power. Technical report, National Bureau of Economic Research.

Azar, J. and I. Marinescu (2024). Monopsony power in the labor market: From theory to policy. *Annual Review of Economics 16*, 491–518.

Bachmann, R., C. Bayer, H. Stüber, and F. Wellschmied (2026). Monopsony makes firms not only small but also unproductive: Why East Germany has not converged. Forthcoming, *Review of Economic Studies*.

Berger, D., K. Herkenhoff, and S. Mongey (2022). Labor market power. *American Economic Review 112*, 1147–1193.

Berger, D., K. Herkenhoff, and S. Mongey (2025). Minimum wages, efficiency, and welfare. *Econometrica 93*(1), 265–301.

Berger, D., K. Herkenhoff, and S. Mongey (2026). Labor market power: From micro evidence to macro consequences. *Journal of Economic Perspectives 40(1)*, 93–114.

Bloesch, J., B. Larsen, and A. Yding (2024). Monopsony with recruiting. Available at SSRN.

Bontemps, C., J.-M. Robin, and G. J. Van den Berg (2000). Equilibrium search with continuous productivity dispersion: Theory and nonparametric estimation. *International Economic Review 41*, 305–358.

Brooks, W. J., J. P. Kaboski, Y. A. Li, and W. Qian (2021). Exploitation of labor? Classical monopsony power and labor's share. *Journal of Development Economics 150*, C.

Burdett, K. and D. Mortensen (1998). Wage differentials, employer size, and unemployment. *International Economic Review 30*, 257–73.

Card, D. (2022). Who set your wage? *American Economic Review 112*(4), 1075–1090.

Card, D. and A. Krueger (1994). Minimum wage, and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review 84 (4)*, 772–93.

Chan, M., S. Salgado, and M. Xu (2020). Heterogeneous passthrough from TFP to wages. Available at SSRN: 3538503.

Deb, S., J. Eeckhout, A. Patel, and L. Warren (2022). What drives wage stagnation: monopsony or monopoly? *Journal of the European Economic Association 20*(6), 2181–2225.

Deb, S., J. Eeckhout, A. Patel, and L. Warren (2024). Walras–Bowley lecture: Market power and wage inequality. *Econometrica 92*(3), 603–636.

Diamond, P. (1982a). Aggregate demand management in search equilibrium. *Journal of Political Economy 90*(5), 881–994.

Diamond, P. (1982b). Wage determination and efficiency in search equilibrium. *Review of Economic Studies 49*(-), 217–227.

Diamond, P. A. (1971). A model of price adjustment. *Journal of Economic Theory 3(2)*, 150–168.

Falch, T. (2010). The elasticity of labor supply at the establishment level. *Journal of Labor Economics 28*, 237–266.

Gottfries, A. and G. Jarosch (2023). Dynamic monopsony with large firms and non-competes. Technical Report 31965, National Bureau of Economic Research.

Gouin-Bonenfant, Émilien. (2022). Productivity dispersion, between-firm competition, and the labor share. *Econometrica 90*, 2755–2793.

Hosios, A. J. (1990). On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies 57*, 279–98.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2024). Granular search, market structure, and wages. *Review of Economic Studies 91*(6), 3569–3607.

Kline, P. (2025). Labor market monopsony: fundamentals and frontiers. In C. Dustmann and T. Lemieux (Eds.), *Handbook of Labor Economics*, Volume 6 of *Handbook of Labor Economics*, Chapter 8, pp. 655–728. Elsevier.

Kroft, K., Y. Luo, M. Mogstad, and B. Setzler (2025). Imperfect competition and rents in labor and product markets: The case of the construction industry. *American Economic Review 115*, 2926–2969.

Lamadon, T., M. Mogstad, and B. Setzler (2022). Imperfect competition, compensating differentials, and rent sharing in the US labor market. *American Economic Review 112*, 169–212.

Malgieri, C. and M. Souchier (2026). Private preferences for jobs and labor market dynamics. Working paper.

Manning, A. (2003). *Monopsony in motion*. Princeton, NJ: Princeton University Press.

Manning, A. (2004). Monopsony and the efficiency of labour market interventions. *Labour Economics 11*, 145–163.

Manning, A. (2021). Monopsony in labor markets: A review. *ILR Review 74(1)*, 3–26.

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

Moen, E. R. and Å. Rosén (2011). Incentives in competitive search equilibrium. *Review of Economic Studies 78*(2), 733–761.

Mortensen, D. T. (1982). The matching process as a noncooperative bargaining game. In *The economics of information and uncertainty*, pp. 233–258. University of Chicago Press.

Mortensen, D. T. and C. A. Pissarides (1994). Job creation and job destruction in the theory of unemployment. *The Review of Economic Studies 61*(3), 397–415.

Mottironi, B. (2025). Labour market power and aggregate productivity. Working paper, Bank of Italy and Centre for Economic Performance, February 2025.

Pissarides, C. (1985). Short-run equilibrium dynamics of unemployment, vacancies, and real wages. *American Economic Review 75*, 676–690.

Pissarides, C. (2000). *Equilibrium Unemployment Theory*. MIT Press.

Pissarides, C. A. (1984). Efficient job rejection. *Economic Journal 94*, 97–108.

Ransom, M. R. and D. P. Sims (2010). Estimating the firm's labor supply curve in a "new monopsony" framework: Schoolteachers in Missouri. *Journal of Labor Economics 28*, 331–355.

Robinson, J. (1969). *The Economics of Imperfect Competition* (2nd ed.). London: McMillan.

Sorkin, I. (2018). Ranking firms using revealed preference. *The Quarterly Journal of Economics 133*(3), 1331–1393.

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*, 211–236.

Veres-Ferrer, E.-J. and J. M. Pavia (2022). The elasticity of a random variable as a tool for measuring and assessing risks. *Risk 10*(68), 1–38.

Wright, R., P. Kircher, J. Benoit, and V. Guerrieri (2021). Directed search and competitive search equilibrium: a guided tour. *Journal of Economic Literature 59*(1), 90–148.

Yeh, C., C. Macaluso, and B. Hershbein (2022). Monopsony in the US labor market. *American Economic Review 112*(7), 2099–2138.

A Properties of Constant-Elasticity Random Variables

Let X have the cumulative distribution function

$$\mathbb{P}(X \leq x) = F(x) = \left(\frac{x}{\bar{x}}\right)^{\gamma} \ , \quad x \in [0, \bar{x}] \ .$$

Its mean is

$$\mathbb{E}[X] = \frac{\gamma}{\gamma + 1}\bar{x} \ ,$$

which is strictly increasing in γ, and conditional expectations satisfy

$$\mathbb{E}[X \mid X \leq x] = \frac{\gamma}{\gamma + 1}x \ .$$

The variance of the random variable is

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left[\frac{\gamma}{\gamma + 2} - \left(\frac{\gamma}{\gamma + 1}\right)^2\right]\bar{x}^2 \ .$$

The variance converges to zero as $\gamma \to 0$ and as $\gamma \to \infty$. In between it first increases and then decreases, with a maximum where

$$\gamma^2(\gamma + 2) = 1 \ ,$$

that is, at $\gamma = (\sqrt{5} - 1)/2 \approx 0.62$.

B Welfare Properties in Dynamic Search

B.1 The Envelope Theorem

Free entry reads

$$(r\kappa + c)(r + \lambda) = q(\theta)F(w - rU)(y - w - r\kappa) \ ,$$

which in turn implies

$$\frac{\eta}{\theta}\frac{\partial \theta}{\partial w} = \frac{f(w - rU)}{F(w - rU)}(1 - r\partial_w U) - \frac{1}{y - w - r\kappa} \ ,$$

where η is the elasticity of $q(\theta)$ with respect to θ, in absolute value. At $w = w^*$ the firm's first-order condition gives $f/F = 1/(y - w - r\kappa)$, so that

$$\frac{\eta}{\theta}\frac{\partial \theta}{\partial w} = -\frac{f(w - rU)}{F(w - rU)}r\partial_w U = -\frac{\epsilon_{F, w - rU}}{w - rU}r\partial_w U = -\frac{r\partial_w U}{y - w - r\kappa} \ .$$

B.2 Proof of Theorem 1

We can use the envelope theorem to prove that it is optimal to raise wages above the monopsonistic level.

Differentiating the expression for rU in the constant-elasticity case of Section 2.8 and using free entry, we find

$$[(r+\lambda)(\gamma+1)+\theta q(\theta)F(w-rU)]r\partial_w U = \theta q(\theta)F(w-rU)+ \\ +(w-rU)\left[\frac{\partial\theta}{\partial w}(1-\eta)q(\theta)F(w-rU)+\theta q(\theta)f(w-rU)(1-r\partial_w U)\right] \tag{38}$$

At $w=w^*$, $\partial_w\theta \propto -\partial_w U$, so that

$$\partial_w U(w^*) > 0$$

since only positive terms are left on the right-hand side once the terms in $\partial_w U$ are collected on the left-hand side. This proves the theorem.

Solving for $r\partial_w U$ gives

$$r\partial_w U = \frac{\theta q(\theta)F(w-rU)}{r+\lambda+\frac{1+\epsilon_{F,w-rU}/\eta}{1+\epsilon_{F,w-rU}}\theta q(\theta)F(w-rU)} \in (0,1) \ .$$

Using the free-entry condition away from the monopsonistic wage gives

$$r\partial_w U = \theta q(\theta)F(w-rU)\frac{\eta+\epsilon_{F,w-rU}-(1-\eta)\frac{w-rU}{y-w-r\kappa}}{\eta(r+\lambda)(\epsilon_{F,w-rU}+1)+(\eta+\epsilon_{F,w-rU})\theta q(\theta)F(w-rU)}$$

and setting it to zero gives the optimal wage

$$w = \frac{\epsilon_{F,w-rU}+\eta}{\epsilon_{F,w-rU}+1}(y-r\kappa)+\frac{1-\eta}{\epsilon_{F,w-rU}+1}rU \ .$$

B.3 Proof of Theorem 2

Unemployment is

$$u = \frac{\lambda}{\lambda+\theta q(\theta)F(w-rU)} \ .$$

The derivative of unemployment with respect to the wage reads

$$\frac{\partial u}{\partial w} = -\frac{u}{\lambda+\theta q(\theta)F(w-rU)}\left[(1-\eta)q(\theta)F(w-rU)\partial_w\theta+\theta q(\theta)f(w-rU)(1-r\partial_w U)\right] \ . \tag{39}$$

A higher wage has three effects: (i) a direct effect that raises acceptance; (ii) an indirect effect that lowers acceptance by raising the value of unemployment, which the first one dominates; and (iii) a negative effect through lower market tightness, which lowers the probability of a match per unemployed worker. At the decentralized wage, the elasticity of unemployment with respect to the wage is

$$\epsilon_{u,w}(w^*) = -\epsilon_{F,w-rU}\frac{w}{w-rU}(1-u)\left[1-\frac{r\partial_w U}{\eta}\right]$$

so the acceptance effect dominates, and a higher wage raises employment, if and only if

$$\frac{r\partial_w U}{\eta} < 1 \ .$$

Using the expression for $r\partial_w U$ above,

$$\frac{r\partial_w U(w^*)}{\eta} = \frac{(\epsilon_{F,w-rU}+1)\theta q(\theta)F(w-rU)}{\eta(r+\lambda)(\epsilon_{F,w-rU}+1)+(\eta+\epsilon_{F,w-rU})\theta q(\theta)F(w-rU)} \ ,$$

which can be larger or smaller than one, depending on parameter values.

B.4 Proof of Proposition 1

At the optimal wage, the derivative of welfare with respect to the wage is zero, $r\partial_w U = 0$, which by (38) implies

$$\frac{\partial\theta}{\partial w} = -\frac{\theta}{w-rU}\frac{1+\epsilon_{F,w-rU}}{1-\eta} \ .$$

Substituting into (39) gives

$$\frac{\partial u}{\partial w} = \frac{u(1-u)}{w-rU} > 0 \ .$$

C The Model with Perfect Information

Suppose that the firm, as well as the worker, observes the cost x. The take-it-or-leave-it offer then leaves the worker no rent,

$$w(x) = rU + x \ ,$$

so that $Q^W(w(x),x) = 0$. The firm hires a worker only if $J(w(x)) \geq V$, so it accepts every cost up to

$$x^* = y - r\kappa - rU$$

The reduced form of the model is

$$\begin{aligned} x^* &= y - r\kappa - rU \\ w(x) &= rU + x \\ r\kappa &= -c + q(\theta)\int_0^{x^*}\left[\frac{y-w(s)-r\kappa}{r+\lambda}\right]dF(s) \\ rU &= z \end{aligned}$$

Since $rU = z$, the cutoff $x^* = y - r\kappa - z$ is the upper bound of the support, every contact is accepted, and free entry becomes

$$r\kappa + c = \frac{q(\theta)}{r+\lambda}\left(y - z - r\kappa - \mathbb{E}[x]\right)$$

We call this result the *information-congestion paradox*: with perfect information the monopsonist leaves workers no rent, so welfare is at its lowest, $rU = z$, while with private information $rU > z$.

A perfectly informed planner, who sets a wage schedule $w(x)$ subject to the matching technology, chooses $x \leq w(x) \leq y$ for every x, so that every contact is accepted, and solves

$$\max_{w} rU = z + \theta q(\theta) \frac{\mathbb{E}[w] - \mathbb{E}[x] - z}{r + \lambda} \ , \quad s.t. \quad r\kappa + c = q(\theta) \frac{y - r\kappa - \mathbb{E}[w]}{r + \lambda}$$

whose solution is

$$\mathbb{E}[w] = \eta(\theta)(y - r\kappa) + (1 - \eta(\theta))(z + \mathbb{E}[x]) \ ,$$

where $\eta(\theta)$ is the elasticity of $q(\theta)$: the planner gives workers the Hosios share of the surplus.

D The Recruiting Cost and the Wage Protocol

This appendix establishes the three claims made in Section 3.3 about the recruiting cost implied by (34).

How to read a large calibrated value of c, and where to place the rent π of (28), depends on what kind of cost c is. We treat c as a *search cost*—an effort the firm expends to fill a vacancy—not as goods used up in hiring. The textbook alternative makes c a flow of output consumed in recruiting, subtracted from GDP as an intermediate use; we do not adopt it. Here c is the firm-side counterpart of the worker's flow value of non-employment z: both are real costs, an effort or a forgone value of leisure or home production, that shape behavior and that the planner internalizes, but neither enters accounting output.

Two things follow. The ratio c/y_{net} is a comparison of *scale*, not a claim that c is deducted from output; the question it poses is whether a search effort worth several quarters of net output per hire is plausible. And because the search cost sits outside GDP, the firm's operating surplus is *recorded* income rather than a recruiting expense netted away. The rent $\pi = 0.095$ is measured output, which the accounts attribute to capital; on the value ledger free entry still binds, $c = q(\theta)F(x^*)J$, so the capitalized rent equals the recruiting effort the firm sinks and the net value of a hire is zero. Income and effort balance because utility is transferable, without either passing through GDP. One caution matters for the welfare exercise: since the rent is the return to a matching effort that the planner also incurs, π is not the welfare cost of monopsony. The distortion is the effect of the wage wedge on entry, which Section 3.2 measures by the efficient wage premium of 10.1 percent.

We now show that the recruiting cost is the same under three wage protocols: a fixed wage, Nash bargaining and monopsonistic wage setting.

Under a fixed wage, with $\tilde{w}$ taken from the data, (34) closes immediately:

$$c \ = \ \frac{p}{\theta} \frac{y_{net} - (1 + \tilde{\tau})\tilde{w}}{r + \lambda}.$$

One equation, one unknown; no wage theory has been used.

Under Nash bargaining, the wage is instead the Nash split with free entry,

$$\tilde{w} \;=\; (1-\beta)\,\tilde{z} + \beta\big(\tilde{S} + c\,\theta\big),$$

with $\tilde{S}$ the break-even wage of (31), $\tilde{z}$ the threat point and β bargaining power. There are now two unknowns and the calibration is *block-recursive*. Stage one pins c by (34) exactly as under a fixed wage, using the observed wage; the bargaining story is not used. Stage two pins β *linearly*, because the threat point is data that does not depend on β:

$$\beta \;=\; \frac{\tilde{w} - \tilde{z}}{\tilde{S} + c\,\theta - \tilde{z}} \;=\; 0.196. \tag{40}$$

A large recruiting cost inflates the bargaining surplus $\tilde{S} + c\theta = 2.14$ in the denominator, forcing a small β to match the observed wage. The small share cannot be repaired by raising the outside option: since the surplus is dominated by $c\theta$, the share is bounded above by $\beta \le \tilde{w}/(\tilde{S} + c\theta) = 0.253$, attained as $\tilde{z} \to 0$, and falls toward zero as $\tilde{z} \to \tilde{w}$. Recovering a conventional $\beta \approx 0.5$ would require a c about five times smaller. The diagnostic is moreover invariant to market tightness: since $c\theta = (p/\theta)J \cdot \theta = pJ$, neither the hiring-cost term nor β depends on θ.

Under monopsonistic wage setting, stage one is again unchanged: c is pinned by (34) from the observed wage, independently of γ. Stage two is not linear, because the outside option is now rU, itself endogenous to γ. The model imposes three relations at the data wage—the first-order condition $x^* = \gamma(\tilde{S} - \tilde{w})$, the value of unemployment $rU = \tilde{z} + p\,x^*/[(\gamma+1)(r+\lambda)]$, and the reservation cost $x^* = \tilde{w} - rU$—and the factor $1/(\gamma+1)$ in rU multiplies the γ in x^*, so eliminating rU and x^* yields the quadratic (26) of Section 2.8.3. Compared with Nash, it is the endogeneity of the outside option, not market power, that makes the second stage quadratic. The recruiting cost c is unchanged, since $J = \pi/(r+\lambda)$ uses the observed wage and $F(x^*)$ cancels against $q(\theta)$ in $c = q(\theta)F(x^*)J = (p/\theta)J$.

Table 5 puts the three protocols side by side on identical data. The recruiting cost is identical across them because each uses the same observed profit flow and the same data-pinned filling rate; they differ only in the second parameter each must impute, and in whether that stage is linear or quadratic. The wage protocol therefore cannot make the recruiting cost larger or smaller, as long as the average wage is matched.

Table 5: Invariance of the calibrated recruiting cost to the wage protocol

	Fixed wage	Nash	Monopsony
c (flow recruiting cost)	0.989	0.989	0.989
c/y_{net}	1.46	1.46	1.46
J/y_{net} (per hire, quarters)	2.99	2.99	2.99
second parameter	—	$\beta = 0.196$	$\gamma = 0.344$
identified by	—	linear (40)	quadratic (26)
implied elasticity	—	—	$1/\gamma = 2.91$

United States calibration, 2022–2024. All three columns use the same data: $y_{net} = 0.679$, $\tilde{w} = 0.542$, $\tilde{z} = 0.153$, $\hat{\theta} = 1.530$, $\hat{p} = 0.745$, $q(\theta)F(x^*) = p/\theta = 0.487$, $\hat{\lambda} = 0.0294$, $\tilde{\tau} = 0.077$, $\kappa = 0$, with wages in total compensation as in the text. Because the profit flow $\pi = y_{net} - (1+\tilde{\tau})\tilde{w}$ uses the *observed* wage, the recruiting cost $c = q(\theta)F(x^*)\,\pi/(r+\lambda)$ is identical across protocols. The protocols differ only in the second parameter each must impute, and in whether that second stage is linear or quadratic.