An Information-based Theory of Monopsony Power

Anton Cheremukhin*
Paulina Restrepo-Echavarria†

March 22, 2025

Abstract

Workhorse models of monopsony power attribute firm wage-setting ability either to search frictions or to job differentiation. This paper develops an information-based model that unifies these perspectives, incorporating both directed and random search through endogenous information frictions. Our framework provides a tractable closed-form wage equation that identifies four key sources of monopsony power: first-mover advantage of firms, labor market tightness, search cost asymmetries, and productive complementarities. Equilibrium sorting patterns influence monopsony power as firms benefit from reduced wage competition in positively assortative matching environments. Numerical calibration of the model generates realistic wage markdowns of 30-40%, consistent with empirical estimates. A constrained-efficient social planner would prescribe a wage increase of approximately 20%, highlighting the potential role of policy in mitigating monopsony distortions.

*Federal Reserve Bank of Dallas, 2200 N Pearl St, Dallas TX 75201, chertosha@gmail.com
†Federal Reserve Bank of St. Louis, One Federal Reserve Bank Plaza, Broadway and Locust St., St. Louis MO 63166, paulinares@me.com

1 Introduction

Workhorse models of monopsony power derive upward sloping labor supply curves either from search frictions or from idiosyncratic preferences for jobs (see Card (2022)). In this paper, we propose an information-based model of monopsony power that provides a unifying framework, where workers have idiosyncratic preferences over job openings and where one of the characteristics of an opening is the size of the application pool as in the directed search literature. We derive implications for wages posted by firms, shedding new light on the sources of monopsony power.

More specifically, we propose a search model with imperfect information and two-sided heterogeneity. We build on Cheremukhin, Restrepo-Echavarria, and Tutino (2020) by formulating a sequential version of our targeted search model, where firms first post wage menus, workers *choose* where to apply in a probabilistic way, and then firms probabilistically *choose* among the workers that applied and make job offers. The degree of precision in the probabilistic decisions is chosen by the agents rationally, subject to information constraints. Thus, our model derives from fundamentals the decision rules which have the multinomial logit (MNL) form commonly postulated in the literature. Firms set wages strategically, fully taking into account the consequences their decisions will have on the number and composition of applicants and the amount of screening they will have to do.

The model preserves the core features of directed search: 1) there is a submarket for each combination of types of agents, characterized by a submarket-specific matching technology; 2) firms (sellers) strategically post (and commit to) wage (price) menus with the intent of attracting specific types of workers (buyers); 3) Search is sequential, meaning that first firms strategically post wages, then choose which submarket to apply to, and then firms choose among the queue of workers that applied to them.

In the spirit of targeted search, both workers and firms incur search costs related to their imperfect ability to distinguish among potential partners. Even though agents know the distribution and their productivities with different types, they do not know exactly where to find a particular type. To do so, they decide how much effort they want to exert to locate a particular type of partner by trading off the cost of search with the payoff they can achieve if successful in finding their desired match. Therefore, agents choose whom to contact in a probabilistic way, and the strategies chosen are discrete

probability distributions over types. Each element of the distribution represents the probability with which an agent will target (i.e., contact) each potential match based on the agent's expected payoff. Exerting more search effort, which results in a higher search cost, allows agents to spot a particular type more accurately. Given the discrete nature of the probability distributions, we model the search cost as proportional to the distance between an uninformed, uniform strategy, where every type has the same probability of being contacted, and the distribution that is optimally chosen by the agent.[1] Varying costs of search allows our model to span the continuum of possible outcomes between directed search (costs tend to zero) and random search (costs tend to infinity).

We characterize theoretically and numerically the equilibrium properties of the strategies of workers and firms, the posted wages, and the sorting and matching patterns depending on parameters. Importantly, we derive a closed-form expression for wages set by firms in equilibrium, which describes how posted wages as a fraction of the surplus depend on equilibrium labor market tightness in each submarket, on the search strategies of workers and firms, and on the information search costs that firms and workers face. When search costs approach zero, in many cases we obtain the prescription of directed search models that an equal split of the surplus should prevail. However, for positive information costs, firms will typically enjoy higher monopsony power which will allow them to pay the workers less than an equal share of the surplus. As can be inferred from the formula for wages, this increased monopsony power of firms comes from three distinct sources: informational search costs which determine the labor supply and labor demand elasticities, labor market tightness which determines the firms' numerical disadvantage (as in oligopsony models) and attenuates both elasticities, and from firms' first-mover advantage which gives them the ability to strategically manipulate wages. In other words, if there were fewer workers than available jobs, and/or if workers were first to strategically post (and commit to) wages (or if search was simultaneous and wages were bargained ex-post), this would allow workers to get higher equilibrium wages.

We derive an easily interpretable formula for the wage markdown which depends on the replacement ratio, the relative number of firms and workers, on the search costs of

[1]This cost specification in Cheremukhin et al. (2020) is borrowed from the literature on discrete choice under information frictions (Cheremukhin et al. (2015) and Matejka and McKay (2015)).

firms and workers, and can be directly compared to estimates in the literature. Interestingly, in the context of a search model, the commonly derived one-to-one relationship between the markdown and the inverse elasticity of labor supply breaks down. This is because wages are determined by the interaction of labor supply, labor demand and equilibrium sorting. The model approaches the Bertrand wage competition case in the directed search limit as the workers' expected payoffs with different firms need to be equalized. As costs move away from zero this equalization is slowly relaxed.

We find in the calibrated version of the model, that markdowns of 30-40%, consistent with those observed in many empirical studies, indicate strong monopsony power of firms consistent with capturing more than 70% of the surplus. As the equilibrium is not socially efficient, we find that a constrained efficient allocation of workers to jobs would be consistent with markdowns of 10-15%, implying a wage increase of 20%. We find that the empirically observed elasticity of labor supply can be used to narrow down the range of possible information search costs to 0.1-0.3. This range of costs is high enough that we should not expect to observe multiplicity of equilibria in practice.

We recover most of the results of directed search models in the limiting case when search costs approach zero, with a few caveats. First, we find that in the limit of zero costs there are generically multiple equilibria. Directed search models routinely select the positive assortative matching equilibrium. This is a wise choice when the productive complementarities are strong (match surplus is horizontal). When productive complementarities are weak (match surplus is vertical, as commonly assumed in the literature), we show that a different type of equilibrium — the mixing equilibrum — is often better from the social planner's point of view, while the positively assortative equilibrium is socially inefficient. Second, the strength of productive complementarities determines sorting, which in turn affects monopsony power. Positive equilibrium assortativeness reduces competition for workers giving firms more monopsony power and enabling lower wages. A mixing equilibrium, on the contrary, features intense competition for workers, leaves firms less monopsony power and leads to a more equal distribution of the surplus.

We find that under positive information search costs all the equilibria are generically socially inefficient. This inefficiency comes from the fact that sequentiality of search removes the positive search externality present in simultaneous search models, but the negative congestion externality remains. Under simultaneous search, the two externalities can exactly balance out each other, which leads to a socially-optimal equilibrium

outcome. When only the negative externality remains, both workers and firms exhibit too much search effort, so there is a role for a social planner to play in dampening their excessive search efforts by taxing their payoffs. In other words, there exists a tax scheme that improves welfare of all the involved parties and brings a net positive profit for the planner (which can be reverted back to the searchers in a lump-sum way to further improve payoffs).

Thus, we make three key contributions. First, we unify directed and random search models by incorporating information frictions, allowing us to span the full continuum between these two extremes. Second, we provide a closed-form wage equation that directly links monopsony power to labor market tightness and search costs, offering new insights into how firms strategically set wages. Third, we generate predictions that align with observed wage markdowns and monopsony estimates, bridging the gap between theory and empirical studies. These findings contribute to ongoing debates on labor market competition and wage-setting power.

The paper proceeds as follows. Section 2 describes the model and studies its properties. Section 3 describes theoretically and numerically the properties of equilibrium wages, and their relationship to the literature on monopsony power. Section 4 discusses the relationship and parallels to the directed search literature and discusses the importance of productive complementarities and sorting for monopsony power. Section 5 concludes.

Related literature

In Cheremukhin, Restrepo-Echavarria, and Tutino (2020) we developed a theory of targeted search where search was simultaneous and the payoff was set through bargaining, and we analyzed it in the context of the marriage market. In this paper we focus on the labor market and extend our previous setup to a sequential search setting where firms post wages and workers decide where to send their applications. Like in Cheremukhin, Restrepo-Echavarria, and Tutino (2020), our paper effectively blends two sources of randomness used in the literature. The first source is a search friction with uniformly random meetings and impatience, as in Shimer and Smith (2000). The second approach introduces unobserved characteristics as a tractable way of accounting for the deviations of data from the stark predictions of the frictionless model, as in Choo and

Siow (2006) and Galichon and Salanie (2012). We introduce a search friction into the meeting process by endogenizing agents' choices of whom to contact. We build on the discrete-choice rational-inattention literature—i.e., Cheremukhin, Popova, and Tutino (2015) and Matejka and McKay (2015)—that derives multinomial logit decision rules as a consequence of cognitive constraints that capture limits to processing information. Therefore, the equilibrium matching rates in our model have a multinomial logit form similar to that in Galichon and Salanie (2012). Unlike Galichon and Salanie, the equilibrium of our model features strong interactions between agents' contact rates driven entirely by their choices, rather than by some unobserved characteristics with fixed distributions.

The search and matching literature has seen multiple attempts to produce intermediate degrees of randomness with which agents meet their best matches. In particular, Menzio (2007) and Lester (2011) nest directed search and random matching to generate outcomes with an intermediate degree of randomness.[2] Our paper produces equilibrium outcomes in between uniform random matching and the frictionless assignment, endogenously, without nesting these two frameworks. One recent paper considering our specification of targeted search with information costs in application to the labor market is Wu (2020).

Also note that although the directed search literature, such as Shimer (2005) and Eeckhout and Kircher (2010), technically involves a choice of whom to meet, the choice is degenerate—directed by signals from the other side. See Chade, Eeckhout, and Smith (2017) for a thorough summary of this literature.

Finally we build on the literature on monopsony power by proposing a model that unifies the three existing theoretical frameworks: oligopsony, job differentiation, and search and matching. See Azar and Marinescu (2024) for a very comprehensive summary of the literature, and Card (2022) for an analysis on the need of this unifying framework.

While existing directed search models assume costless search and efficient wage-setting, our model endogenizes search effort and allows for strategic firm wage-setting. This leads to novel insights: (i) the emergence of multiple equilibria under low search costs, (ii) the endogenous formation of monopsony power through firm-worker coordi-

[2]Also, see Yang's (2013) model of "targeted" search that assumes random search within perfectly distinguishable market segments.

nation, and (iii) a direct link between search costs, labor market tightness and observed wage markdowns. By relaxing the assumption of costless search, our framework provides a richer and more flexible model of labor market frictions.

2 Model

In this section, we present a model where firms are looking to fill a vacancy, and workers—who are either employed or unemployed—are looking to find a job. Each agent chooses a probabilistic search strategy that can be interpreted as a search intensity over types, where each element of this distribution reflects the likelihood of contacting a particular agent on the other side. A more targeted search, or a probability distribution that is more concentrated on a particular group of agents (or agent), is associated with a higher cost, as the agent needs to exert more effort to locate a particular potential match more accurately.

The economy contains a large, finite number of individual agents: workers whose types are indexed by $x \in \{1, ..., W\}$ and firms whose types are indexed by $y \in \{1, ..., F\}$. We denote by μ_x the number of workers of type x and by μ_y the number of firms of type y. We think of workers and firms characterized by a multidimensional set of attributes. Types x and y are unranked indices that aggregate all attributes.

A match between any worker of type x and any firm of type y generates a payoff (surplus) f_{xy}. We do not place any restrictions on the shape of the payoff function, and we normalize the outside option of both the worker and the firm to zero. We denote the payoff (wage) appropriated by the worker ω_{xy} and the payoff appropriated by the firm η_{xy} such that $\eta_{xy} = f_{xy} - \omega_{xy}$.

Agents form a match if they meet, and each agent (weakly) benefits from forming a match; i.e., each agent's payoff is non-negative. Since a negative payoff corresponds to absence of a match, we make the following assumption on the payoffs:

Assumption 1. The payoffs are non-negative:

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

When seeking to form a match, both workers and firms know the number of agents of each type and the characteristics of their preferred types on the other side of the market.

They face a noisy search process where they are uncertain about how to locate their preferred match. In this environment, each agent's action is a probability distribution over agents on the other side of the market. Since the number of potential matches is finite, the strategy of each agent is a discrete probability distribution. Let $\bar{p}_x(y)$ be the probability that a worker of type x targets or sends an application to a firm of type y. Similarly, we denote by $\bar{q}_y(x)$ the probability that a firm of type y targets or considers the application of a worker of type x.

Reducing the noise to locate a potential match more accurately is costly: It involves a careful analysis of the profiles of potential matches, with considerable effort in sorting through the multifaceted attributes of each firm and candidate. When seeking to form a match, agents rationally weigh costs and benefits of targeting the type of characteristics that result in a suitable match. A worker rationally chooses their strategy $\bar{p}_x(y)$ by balancing the costs and benefits of targeting a given firm. A strategy $\bar{p}_x(y)$ that is more concentrated on a particular firm of type y affords them a higher probability to be matched with their preferred firm. However, it requires more effort to sort through profiles of all the firms in the market to locate their desired match and exclude the others. So locating a particular firm or worker more accurately requires exerting more search effort, and it is costlier.

We assume that agents enter the search process with a uniform prior of whom to target, $\tilde{p}_x(y)$ and $\tilde{q}_y(x)$. Choosing a more targeted strategy implies a larger distance between the chosen strategy and the uniform prior and is associated with a higher search effort. A natural way to introduce this feature into our model is the Kullback-Leibler divergence (relative entropy),[3] which provides a convenient way of quantifying the distance between any two distributions, including discrete distributions as in our model. We assume that the search effort of worker i of type x is defined as follows:

$$\kappa_x = \sum_{y=1}^{F} \mu_y \bar{p}_x(y) \ln \frac{\bar{p}_x(y)}{\tilde{p}_x(y)}. \tag{2.1}$$

We assume that the search costs $c_x(\kappa_x)$ are a function of the search effort κ_x. Note

[3]In the model of information frictions used in the rational inattention literature, κ_x represents the relative entropy between a uniform prior and the posterior strategy. This definition is a special case of Shannon's channel capacity, where information structure is the only choice variable (See Thomas and Cover (1991), Chapter 2). See also Cheremukhin, Popova, and Tutino (2015) for an application to stochastic discrete choice with information costs.

that κ_x is increasing in the distance between a uniform distribution over firms and the chosen strategy, $\bar{p}_x(y)$. If an agent does not want to exert any search effort, she can choose a uniform distribution over types and meet firms randomly. As she chooses a more targeted strategy, the distance between the uniform distribution and her strategy $\bar{p}_x(y)$ grows, increasing search effort κ_x and the overall cost of search. By increasing the search effort, agents bring down uncertainty about locating a prospective match, which allows them to target their better matches more accurately.

Likewise, a firm's cost of search $c_y(\kappa_y)$ is a function of the search effort defined as:

$$\kappa_y = \sum_{x=1}^{F} \mu_x \bar{q}_y(x) \ln \frac{\bar{q}_y(x)}{\tilde{q}_y(x)}. \tag{2.2}$$

Furthermore, we assume the following:

Assumption 2. The search costs of agents $c_x(\kappa)$ and $c_y(\kappa)$ are strictly increasing, twice continuously differentiable and (weakly) convex functions of search effort.

As a special case, we consider a linear cost of search. Then, the total costs of search for a worker of type x are given by $c_x = \theta_x \kappa_x$ and for a firm of type y by $c_y = \theta_y \kappa_y$, where $\theta_x \geq 0$ and $\theta_y \geq 0$ are the marginal costs of search.

For convenience in comparing wage posting and bargaining setups, we introduce a new notation for the strategies of the workers and firms. We define the workers' and firms' search intensities as the ratios of their posterior and prior: $p_x(y) = \frac{\bar{p}_x(y)}{\tilde{p}_x(y)}$ and $q_y(x) = \frac{\bar{q}_y(x)}{\tilde{q}_y(x)}$, respectively.

The meeting rate depends on the strategies of each agent, $p_x(y)$ and $q_y(x)$, and a congestion function $\phi(p_x(y), q_y(x), \mu_x, \mu_y)$, which depends in some general way on the strategies of all other agents as well as the number of agents of each type. Given this, the total number of matches formed between workers of type x and firms of type y is given by

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

Assumption 3. The congestion function is twice continuously differentiable in each p and q.

We introduce this congestion function following Shimer and Smith (2001) and Mortensen (1982), who assume a linear search technology. Note that if $\phi(...) = 1$, then a match takes place if and only if there is mutual coincidence of interests; i.e.,

both agents draw each other out of their respective distribution of interests. By introducing a congestion function we are allowing for matches to depend in some general way on both an agent's search intensity[4] for a specific agent (p and q) and on the number of agents taking part. We think of this assumption as representing the matching technology in a separate submarket for each combination of x and y.

Note that when setting up the congestion function we implicitly assume that there are no direct inter-type congestion externalities. However, our model still features strong indirect equilibrium interactions between the strategies of agents that work akin to inter-type congestion by attracting or deterring agents.

2.1 Sequential targeted search

To initiate the search and matching process, firms start by posting vacancies. Each posted vacancy includes a wage menu, and the firm commits to paying a type-dependent wage in the case of matching. After the vacancies are posted, and because workers cannot perfectly distinguish which firm is of which type despite learning the wage menus of each firm, they choose a distribution of search intensities that determines the likelihood of contacting a particular firm and choose one firm from this distribution to send an application. Finally, once firms have received worker's applications, each firm chooses the worker to which it will extend a job offer from the set of workers that applied to that particular firm.

When workers decide where to send their applications, they take as given the (posted or bargained) wages of firms, such that the set of strategies of workers $p_x(y) \in S_x$ is given by:

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

where $p_x(y) = \frac{\bar{p}_x(y)}{\tilde{p}_x(y)}$, and $\tilde{p}_x(y) = 1/\sum_{y=1}^{F} \mu_y = 1/\delta_x$ is the worker's uniform prior over

[4]Note that here, search intensity refers to how concentrated the distribution of interests of an agent is. A higher search intensity results in assigning higher probability to one or several agents within an agent's distribution of interests.

the whole set of firms $\left(\delta_x = \sum_{y=1}^{F} \mu_y\right)$.

The firms will strategically choose a wage menu $\omega_{x,y}$ and screening strategy $q_y(x)$. The other difference between the problem of workers and firms is that firms do not sort through all the workers that are looking for a job; they only sort through those that send an application to their firm, and when doing so, firms do not know the types of the workers that applied, but they know the length and expected composition of the queue. In expectation, the queue of firm y contains $\mu_x p_x(y)\,\delta_x/\mu_y$ workers of type x.

We define the set of strategies available for firms as:

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

where $q_y(x) = \frac{\bar{p}_y(x)}{\tilde{p}_y(x)}$, and $\tilde{q}_y(x) = 1/\sum_{x=1}^{W} (\mu_x p_x(y)\,\delta_x/\mu_y)$ is the firm's uniform prior over their own queue. Here we define new variables for queue weights $a_{xy} = \frac{\mu_x p_x(y)}{\Sigma_{x=1}^{W} \mu_x p_x(y)}$, and queue length $\delta_y = \Sigma_{x=1}^{W} \mu_x p_x(y)$.

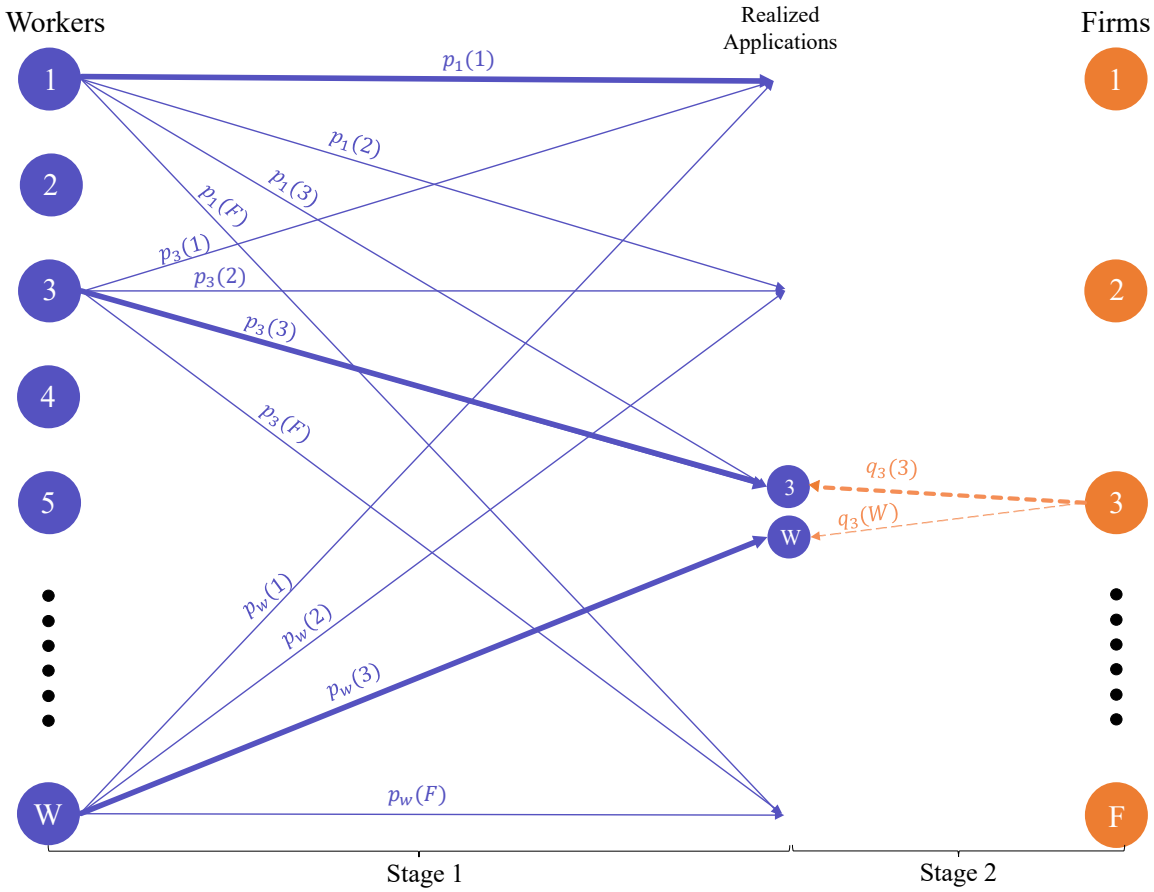
The set of actions $s \in S$ is given by the cartesian product of the sets of strategies of workers $s_x \in S_x$ and firms $s_y \in S_y$.

Figure 2.1 illustrates the interactions and search strategies of workers and firms. The solid arrows show the intensity $p_x(y)$ that a worker of type x assigns to targeting a firm of type y. Similarly, dashed arrows show the intensity $q_y(x)$ that a firm of type y assigns to targeting a worker of type x. Once these are selected, both workers and firms make one draw from their respective distributions to determine where to send an application and which applications to inspect (denoted by bold arrows).

Although applications and/or job offers are not lost in the mail, there is still a coordination problem: $\mu_x p_x(y)$ workers applied to type y firms, and firms sent $\mu_y q_y(x)\,\mu_x p_x(y)$ job offers, but they did not necessarily send all of those to different workers. Several firms might contact the same worker, and some workers may not get any offers. We assume that $\mu_x p_x \mu_y q_y \phi_{xy}$ matches are created, where the coordination problem between type x workers and type y firms is captured by the congestion function/meeting technology $\phi_{xy}(p_x, q_y, \mu_x, \mu_y)$ described earlier.

Both firms and workers choose their optimal strategies, and if a firm and a worker match, the payoff f_{xy} is split between them according to the commitment whereby firms

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



posted type-dependent wage menus in the first stage of the game.

The game is sequential as in Stackelberg in that when firms post wages and choose their search effort, they internalize the best response strategies of workers. Firms behave like leaders and workers behave like followers. However, consistent with the assumptions of the simultaneous model (see Cheremukhin, Restrepo-Echavarria, and Tutino (2020)), neither the workers nor the firms internalize the effects of their strategies on the congestion function and take matching rates in each submarket as given. This is because there are a large number of individuals of each type, so a change in an individual firm's or worker's strategy will not have a noticeable aggregate effect on the number of matches. This assumption of large number of identical agents of each type which all play identical strategies is reminiscent of "competitive" search.

Assumption 4. Agents take the meeting rates they face as given, disregarding the dependence of the congestion function on agents' own search intensities.

Definition. A *matching equilibrium* is a set of admissible strategies for workers $s_x \in S_x$, firms $s_y \in S_y$, and meeting rates, such that the strategies solve the problems for each individual firm and worker given the meeting rates, which are consistent with the strategies of the agents.

2.2 The problem of the worker

We start by describing the problem of the worker. Workers take as given $q_y(x)\,\phi_{xy}$—the probability of forming a match with type y firms. The worker receives a wage ω_{xy} in the case of matching and bears a linear cost of search $\theta_x \kappa_x(p_x(y))$. The goal of type x workers is to maximize surplus subject to a constraint on strategies (with renormalized Lagrange multiplier λ_x):

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

Since the objective function of workers is twice continuously differentiable and concave in their own strategies, first-order conditions are necessary and sufficient conditions for equilibrium. Using the necessary first-order conditions we can derive a closed-form

solution for the optimal strategy of workers:

$$p_x^*(y) = \frac{\exp\left(\frac{q_y(x)\phi_{xy}\omega_{xy}}{\theta_x/\delta_x}\right)}{\Sigma_{y'=1}^{F}\frac{\mu_{y'}}{\delta_x}\exp\left(\frac{q_{y'}(x)\phi_{xy'}\omega_{xy'}}{\theta_x/\delta_x}\right)}. \tag{2.3}$$

2.3 The problem of the firm

The goal of type y firms is to choose wages and search intensities over their queue of workers to maximize their expected match payoffs $f_{xy} - \omega_{xy}$, net of linear search costs $\theta_y \kappa_y(q_y(x))$ and subject to a constraint on strategies (with renormalized Lagrange multiplier λ_y):

$$\begin{aligned} Y_y &= \Sigma_{x=1}^{W}\mu_x p_x(y)\phi_{xy}q_y(x)(f_{xy}-\omega_{xy}) - \theta_y\Sigma_{x=1}^{W}\frac{\mu_x p_x(y)}{\Sigma_{x=1}^{W}\mu_x p_x(y)}q_y(x)\ln q_y(x) \\ &\quad +\theta_y\lambda_y\left(1-\Sigma_{x=1}^{W}\frac{\mu_x p_x(y)}{\Sigma_{x=1}^{W}\mu_x p_x(y)}q_y(x)\right). \end{aligned}$$

The firm internalizes the best responses of the workers (Equation 2.3). To internalize the responses, we need to take derivatives of $p_x(y)$ with respect to the wage ω_{xy} set by the firm and with respect to the firm's search strategy $q_y(x)$. If we introduce new notation $z_{xy} = \frac{\phi_{xy}q_y(x)}{\theta_x/\delta_x}\left(1-\frac{\mu_y}{\delta_x}p_x(y)\right)$, then the partial derivatives of (2.3) are conveniently given by: $\frac{\partial p_x(y)}{\partial q_y(x)}\frac{q_y(x)}{p_x(y)} = \omega_{xy}z_{xy}$ and $\frac{\partial p_x(y)}{\partial \omega_{xy}}\frac{1}{p_x(y)} = z_{xy}$. In addition, note that the derivatives of queue weights $a_{xy} = \frac{\mu_x p_x(y)}{\Sigma_{x=1}^{W}\mu_x p_x(y)}$ can be computed as $\frac{\partial a_{xy}}{\partial X} = a_{xy}(1-a_{xy})\frac{\partial p_x(y)}{\partial X}\frac{1}{p_x(y)}$.

The problem can be rewritten as:

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

and we can write the first-order condition of the firm with respect to search intensities as follows:

$$\frac{\partial Y_y}{\partial q_y} = \mu_x p_x(y)\frac{\theta_y}{\delta_y}\left[\begin{array}{c} \frac{\phi_{xy}}{\theta_y/\delta_y}(f_{xy}-\omega_{xy})(1+z_{xy}\omega_{xy})-1 \\ -(\ln q_y(x)+\lambda_y)(1+(1-a_{xy})z_{xy}\omega_{xy}) \end{array}\right] = 0.$$

Strategies of firms then satisfy:

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

Firms' strategies must therefore satisfy the following necessary condition for equilibrium:

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

Firms also optimally choose wage menus in the first stage. We can write the first-order condition with respect to wages as follows:

$$\frac{\partial Y_x}{\partial \omega_{xy}} = \mu_x p_x(y) q_y(x) \frac{\theta_y}{\delta_y} \left[\begin{array}{c} \frac{\phi_{xy}}{\theta_y/\delta_y} \left((f_{xy} - \omega_{xy}) z_{xy} - 1 \right) \\ - (\ln q_y(x) + \lambda_y)(1 - a_{xy}) z_{xy} \end{array} \right] = 0,$$

and the second-order derivatives as:

$$\frac{\partial^2 Y_x}{\partial q_{xy}^2} = -\frac{1}{q_y(x)}, \qquad \frac{\partial^2 Y_x}{\partial \omega_{xy}^2} = -\frac{\phi_{xy}}{\theta_y/\delta_y} z_{xy}.$$

Since the objective function of firms is twice continuously differentiable and strictly concave with respect to their own strategies, the first-order conditions are necessary and sufficient conditions for equilibrium. Furthermore, we can combine the two optimality conditions to eliminate $q_y(x)$ and obtain a simple expression for an interior solution $0 \leq \omega_{xy} \leq f_{xy}$ for the wage:

$$\omega_{xy}^* = \left[a_{xy} f_{xy} + (1 - a_x) \frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{z_{xy}} \right]_0^{f_{xy}}. \tag{2.5}$$

Wages stay at the limits because beyond the limits there is no match and the decision-maker is strictly worse off (as reflected in the constraints on the strategy space). In this case we can also substitute the (interior) optimal wage to obtain optimal search

intensities of firms:

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

The properties of the equilibrium, fully characterized by necessary conditions 2.3, 2.4 and 2.5 critically depend on the assumptions regarding the congestion function, in other words, the matching technology.

The matching technology we introduce is a standard symmetric constant returns to scale matching technology that combines the number of participants in each submarket. The number of agents entering each submarket (x, y) are $c_{x,y} = \mu_x p_x(y) \frac{\mu_y}{\delta_x}$ and $d_{x,y} = \mu_y q_y(x) a_{x,y}$. We assume that the matching technology is described by a symmetric CES function $M(c,d) = \left(\frac{1}{2} c^{\frac{\sigma-1}{\sigma}} + \frac{1}{2} d^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}$, with $\sigma > 0, \sigma \neq 1$, with special cases for Cobb-Douglas when $\sigma = 1$ and Leontief when $\sigma = 0$. In this case, the congestion function is defined as $\phi_{x,y} = M(c_{x,y}, d_{x,y}) / \mu_x \mu_y p_x(y) q_y(x)$. This assumption for various parameter choices encompasses most of the interesting cases studied in the literature. It is also directly comparable to our simultaneous targeted search model as it gives the same first best allocation when search costs approach zero.

Proposition 1. *Under assumptions 1- 4, there exists $\underline{\theta}$ such that for high enough costs relative to the number of agents $\left(\frac{\theta_x}{\delta_x}, \frac{\theta_y}{\delta_y}\right) > \underline{\theta}$ a matching equilibrium exists and is unique.*

Proof. The equilibrium of the matching model can be interpreted as a pure-strategy Nash equilibrium of a strategic form game among first-stage decisions of firms. Since the strategy space is a simplex and, hence, a non-empty, convex, compact set, sufficient conditions for the existence of the equilibrium require us to check whether the payoff functions are super-modular on the whole strategy space as in Tarski (1955). Supermodularity can be proven by showing negativity of diagonal elements and non-negativity of the off-diagonal elements of the Hessian matrix.

Let $J_y = \left[\begin{array}{cc} \frac{\partial Y_y}{\partial q_{yx}} & \frac{\partial Y_y}{\partial \omega_{xy}} \end{array} \right]$ be the Jacobian matrix collecting the set of first-order conditions for all firms $y \in \{1, ..., M\}$, and let H be the corresponding Hessian matrix. To derive the Hessian matrix, note that under A.1, strategies of each firm are non-cooperative, i.e., independent of the strategies of other types as well as the strategies of the other agents of their own type. Note also that we have assumed no direct inter-type

congestion externalities. These assumptions produce a Hessian matrix with a block-diagonal structure, which greatly simplifies the analysis. The Hessian consists of 2x2 blocks along the diagonal of the form:

$$H_{xy} = \begin{bmatrix} \frac{\partial^2 Y_y}{\partial q_{yx} \partial q_{yx}} & \frac{\partial^2 Y_y}{\partial \omega_{xy} \partial q_{yx}} \\ \frac{\partial^2 Y_y}{\partial q_{yx} \partial \omega_{xy}} & \frac{\partial^2 Y_y}{\partial \omega_{xy} \partial \omega_{xy}} \end{bmatrix}.$$

All the remaining off-diagonal elements are zero. The derivatives of interest are quite cumbersome to compute. However, we can express the elements of the Hessian as follows (where F and G are some positive functions):

$$\frac{\partial^2 Y_y}{\partial q_{yx} \partial q_{yx}} = -\frac{1}{q_{xy}} + \frac{\delta_x \delta_y}{\theta_x \theta_y} F\left(f_{xy}, \omega_{xy}, q_{xy}, a_{xy}\right) \leq 0,$$

$$\frac{\partial^2 Y_y}{\partial q_{yx} \partial \omega_{xy}} = \frac{\delta_x \delta_y}{\theta_x \theta_y} G\left(f_{xy}, \omega_{xy}, q_{xy}, a_{xy}\right) \geq 0,$$

$$\frac{\partial^2 Y_y}{\partial \omega_{xy} \partial \omega_{xy}} = -\frac{\delta_x \delta_y}{\theta_x \theta_y} \phi_{xy} \phi_{xy} q_{yx} \leq 0.$$

From this structure, it is clear that if costs of search are large enough (separately or in combination) relative to the number of agents, then all of these inequalities hold, while if costs are very small (or number of agents large) the first inequality is violated. For uniqueness, we need diagonal dominance of the form:

$$\left| \frac{\partial^2 Y_y}{\partial \omega_{xy} \partial \omega_{xy}} \right| \left| \frac{\partial^2 Y_y}{\partial q_{yx} \partial q_{yx}} \right| > \left(\frac{\partial^2 Y_y}{\partial q_{yx} \partial \omega_{xy}} \right)^2.$$

If costs are large enough (or number of agents small enough), then the diagonal terms dominate the off-diagonal terms. On the contrary, when costs are small (or numbers of agents large), then diagonal dominance may well be violated. We observe important cases of multiplicity numerically and discuss these in Section 3.1. □

In practice, we find that the threshold $\underline{\theta}$ is quite low, allowing meaningful computations under most parameterizations of interest.

2.4 Social planner's solution

We solve the social planner's problem for the sequential model assuming an utilitarian welfare function. Interestingly, the wage decision disappears from the social planner's problem altogether. We can write social welfare as the sum of objective functions of all the agents in the model, as the planner takes into account all the same benefits and costs of the matching process as the agents, subject to the same constraints on search intensities as individual agents. The social welfare function is then:

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

The wages cancel out from the problem, and hence the planner's solution only describes allocations of search effort, but does not place restrictions on wage determination. The first-order conditions for the planner's problem can be written as follows:

$$\frac{\partial\Omega}{\partial p_x(y)} = \mu_x\mu_y\left(\begin{array}{c} q_y(x)f_{xy}\phi_{xy}\left(1+\varepsilon_{\phi,p}\right) - \frac{\theta_x}{\delta_x}\left(\ln p_x(y) + \lambda_x + 1\right) \\ -\frac{\theta_y}{\delta_y}\left(1-a_{xy}\right)q_y(x)\left(\ln q_y(x) + \lambda_y\right)\end{array}\right) = 0,$$

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

where we denote $\varepsilon_{\phi,q} = \frac{\partial\phi_{xy}}{\partial q_y(x)}\frac{q_y(x)}{\phi_{xy}}$ and $\varepsilon_{\phi,p} = \frac{\partial\phi_{xy}}{\partial p_x(y)}\frac{p_x(y)}{\phi_{xy}}$. We can deduce that the search intensities prescribed by the planner satisfy:

$$\frac{\theta_y}{\delta_y}\left(\ln q_y(x) + \lambda_y\right) = \phi_{xy}f_{xy}\left(1+\varepsilon_{\phi,q}\right) - \frac{\theta_y}{\delta_y},$$
$$\frac{\theta_x}{\delta_x}\left(\ln p_x(y) + \lambda_x + 1\right) = q_y(x)\left(f_{xy}\phi_{xy}\left[\left(1+\varepsilon_{\phi,p}\right) - \left(1-a_{xy}\right)\left(1+\varepsilon_{\phi,q}\right)\right] + \left(1-a_{xy}\right)\frac{\theta_y}{\delta_y}\right).$$

Now, let's compare these expressions with those of the competitive equilibrium:

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

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

Comparing the conditions for the workers, to implement the strategies proposed by the social planner, workers should be promised a wage:

$$\omega_{xy}^{PO,W}=f_{xy}\left(1+\varepsilon_{\phi,p}-\left(1-a_{xy}\right)\left(1+\varepsilon_{\phi,q}\right)\right)+\left(1-a_{xy}\right)\frac{\theta_y/\delta_y}{\phi_{xy}}.$$

Interestingly, under our calibration of the congestion function, $\varepsilon_{\phi,q}=-\frac{1}{2}$ and $\varepsilon_{\phi,p}=-\frac{1}{2}a_{xy}$. Substituting these expressions gives:

$$\omega_{xy}^{PO,W}=\frac{1}{2}f_{xy}+\left(1-a_{xy}\right)\frac{\theta_y/\delta_y}{\phi_{xy}}.$$

The planner promises the worker half the surplus plus a positive term which vanishes as firms' search costs approach 0. In the limit, workers should receive exactly half the surplus. Comparing with the wage prevailing in competitive equilibrium given by (2.5), we observe that the workers are promised a fraction a_{xy} of the surplus instead of half, and the firms charge an additional monopsony discount $1/z_{xy}$ reflecting their first mover advantage.

Comparing the conditions for the firms, to implement the socially optimal strategies, firms should be promised a wage that satisfies:

$$\left(\phi_{xy}\left(f_{xy}-\omega_{xy}\right)\left(1+z_{xy}\omega_{xy}\right)-\frac{\theta_y}{\delta_y}\right)=\left(\phi_{xy}f_{xy}\left(1+\varepsilon_{\phi,q}\right)-\frac{\theta_y}{\delta_y}\right)\left(1+\left(1-a_{xy}\right)z_{xy}\omega_{xy}\right),$$

which boils down to a quadratic equation with respect to wages with one positive solution

$$\omega_{xy}^{PO,F}=\frac{A}{2}+\sqrt{\frac{A^2}{4}-\frac{1}{z_{xy}}f_{xy}\varepsilon_{\phi,q}},$$

where we denote $A = a_{xy}f_{xy} - (1-a_{xy})\, f_{xy}\varepsilon_{\phi,q} + \frac{\theta_y/\delta_y}{\phi_{xy}}\,(1-a_{xy}) - \frac{1}{z_{xy}}$. Note that in our calibration when $\varepsilon_{\phi,q} = -\frac{1}{2}$, we can approximate the wage the planner would prescribe for firms to give away as follows:

$$\omega_{xy}^{PO,F} \approx \frac{1}{2}f_{xy} + (1-a_x)\,\frac{\theta_y/\delta_y}{\phi_{xy}} + a_{xy}\frac{1}{2}f_{xy} - \frac{1}{z_{xy}}\left(1 - \frac{\frac{1}{2}f_{xy}}{\frac{1}{2}f_{xy}\,(1+a_{xy}) + (1-a_x)\,\frac{\theta_y/\delta_y}{\phi_{xy}} - \frac{1}{z_{xy}}}\right).$$

For most parameters of interest, for low values of search costs, the firms should give away noticeably more than half of the surplus, leaving less than half for themselves, while workers should be getting exactly half. This demonstrates the fact that in the presence of negative externalities coming from congestion, both workers and firms jointly over-supply search effort in equilibrium, while the planner would promise them together less than the whole surplus in an attempt to dis-incentivize them from putting excessive effort into search.

More generally, for the special (Cobb-Douglas) congestion function described earlier (implying a constant returns-to-scale matching function) for low enough costs of search we have $\omega_{xy}^{CE} < \omega_{xy}^{PO,W} < \omega_{xy}^{PO,F}$. Because of strong negative congestion externalities, both workers and firms need to be dis-incentivized from putting inefficiently high search efforts by the planner promising lower payoffs in the case of matching. Implementation of this solution looks very much like a tax scheme that benefits workers and hurts firms yet obtains a better matching outcome at a lower search cost and on top generates extra revenue for society.

3 Properties of wages and monopsony power

Having established the equilibrium structure of the model, we now explore how posted wages are determined. The key questions we seek to answer are: (i) How do search costs shape equilibrium wages? (ii) What are the sources of monopsony power in this framework? (iii) How does labor market tightness interact with firm wage-setting strategies? This section presents analytical results that shed light on these questions, followed by numerical simulations that illustrate and validate our findings.

We start the discussion of properties of equilibrium from equation 2.5 which de-

scribes how the equilibrium posted wage is determined.

$$\omega_{xy}^{*} = \left[\underbrace{a_{xy} f_{xy}}_{\substack{\text{Direct incentive} \\ \text{to self-select}}} + \underbrace{(1 - a_x) \frac{\theta_y / \delta_y}{\phi_{xy}}}_{\text{Competitive premium}} - \underbrace{\frac{1}{z_{xy}}}_{\text{Monopsony discount}} \right]_{0}^{f_{xy}}.$$

Let us unpack each of these terms. The first term promises the workers of type x a share of the suplus $a_{xy} = \frac{\mu_x p_x(y)}{\Sigma_{x=1}^{W} \mu_x p_x(y)}$ equal to their fraction in the queue of workers applying to positions at firms of type y. This term reflects two mechanisms. The workers are given an incentive to search harder so that if they are able to better self-select into this type of job, they will get a higher wage. This term also reflects the fact that if other types of workers do not apply to this job, then this type of workers faces less direct wage competition from other types of workers and can expect a higher wage.

The second term adds on top a premium proportional to the marginal search cost faced by the firms. As it gets harder for the firm to screen workers in their queue, they prefer to delegate some of that self-selection to the workers by promising a higher wage. The firms' marginal search cost acts as the inverse of the elasticity of labor demand in conventional models (see below for more on this intuition). If we open up the congestion term ϕ_{xy}, e.g. under the assumption that it produces a symmetric Cobb-Douglas matching function we described earlier, it will simplify to $(1 - a_{xy})\, \theta_y \sqrt{\frac{1}{a_{xy}} \frac{\delta_x}{\delta_y}}$. This derivation shows that the competitive premium also depends positively on the ratio of the overall number of firms ($\delta_x = \Sigma_y \mu_y$) to the total number of workers that apply to firms of type y ($\delta_y = \Sigma_x \mu_x p_x(y)$). If firms are at a numerical disadvantage (δ_x is small), this increases the firms' monopsony power and decreases the equilibrium wage.

The third term represents the firms' monopsony discount (we deliberately use a different word to distinguish it from the wage markdown). This term is reminiscent of the literature on monopsony power where the wage markdown is often shown to be proportional to the inverse of the labor supply elasticity. Recall that the variable z_{xy} was introduced as the semi-elasticity measuring the effect of an increase in the wage on the number of workers applying to the firm: $z_{xy} = \frac{\partial p_x(y)}{\partial \omega_{xy}} \frac{1}{p_x(y)}$. The more conventional labor supply elasticity in this case can be computed as follows: $\epsilon_{p,\omega} = \frac{\partial p_x(y)}{\partial \omega_{xy}} \frac{\omega_{xy}}{p_x(y)} = \omega_{xy} z_{xy}$.

The monopsony discount is driven by competition with other firms. It works towards

equalizing the expected payoffs faced by the worker from this type of firm compared with other types of firms. If we recall that from the strategy of the worker we earlier derived $z_{xy} = \frac{\phi_{xy} q_y(x)}{\theta_x/\delta_x}\left(1 - \frac{\mu_y}{\delta_x} p_x(y)\right)$, we can further substitute the (Cobb-Douglas) congestion function to deduce that $\epsilon_{p,\omega} \propto \frac{1}{\theta_x}\sqrt{\frac{1}{a_{xy}}\frac{\delta_x}{\delta_y}}$. The interpretation of this result is similar to that in the literature deriving a multinomial logit form for the firms' labor supply from a distribution of idiosyncratic tastes. In our specification labor supply also has the multinomial logit (MNL) form (2.3) derived from a micro-founded information search friction. The elasticity of labor supply in this MNL form is inversely proportional to the marginal search costs faced by workers, θ_x. As it gets harder for workers to distinguish between firms, labor supply becomes less elastic which lowers the equilibrium wage. The elasticity also depends on the ratio of the total number of firms and the total number of workers that apply to each firm. As the firms' relative numerical disadvantage increases (μ_y decreases), the elasticity of labor supply decreases, the monopsony discount increases, and the equilibrium posted wage declines.

For comparison with the literature, we can also derive the equilibrium markdown. Here we need to note that we have assumed, without loss of generality, that the outside option of workers is normalized to 0. For the definition and appropriate computation of the markdown (and the labor supply elasticity) we need to assume (and calibrate) the size of the outside option b relative to the size of the surplus f_{xy}. Taking into account the outside option, the equilibrium markdown can be computed as:

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

Note, that there is no simple one-to-one mapping between the labor supply elasticity $(\omega_{xy} + b) z_{xy}$ and the markdown. This is because multiple factors simultaneously determine the equilibrium markdown. The main factors affecting the equilibrium wage and wage markdown are:

1. The workers' marginal costs of search determine the labor supply elasticity and the monopsony discount.

2. The firms' marginal costs of search determine the labor demand elasticity and the competitive premium.

3. The numerical (dis)advantage each firm and worker face in the market (also, labor market tightness) attenuate both elasticities.

4. The strength and pattern of equilibrium sorting (through a_{xy}) determines the starting point for the equilibrium wage and markdown.

5. The first-mover advantage (the ability of firms to announce and pre-commit to a wage menu) gives the firms the ability to strategically manipulate wages and extract the monopsony discount. Unlike this model, a model of simultaneous search with bargaining can achieve the constrained social optimum.

This theoretical analysis highlights several key drivers of monopsony power, including search costs, labor market tightness, and equilibrium sorting patterns. However, the strength of these effects in realistic labor markets remains an open question. To quantify these mechanisms and compare them to empirical benchmarks, we now turn to numerical simulations using calibrated parameter values.

3.1 Numerical results

In this subsection, we illustrate the theoretical results of the model with some numerical examples. In particular, we explore the effects that sorting, search costs and relative numbers of workers and firms can have on the monopsony power of the firm, the labor supply elasticity, and equilibrium markdowns. We first discuss the calibration of some core parameters (f_{xy}, b, ϕ_{xy}), and then compute the competitive equilibria for various combinations of other parameters $(\theta_x, \theta_y, \mu_x, \mu_y)$.

We assume $W = 2$ types of workers and $F = 2$ types of firms. We consider two opposite shapes that the surplus function f_{xy} can take, representing horizontal and vertical preferences. Thus, we consider $f_{xy} = 2\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $f_{xy} = 2\begin{bmatrix} 2 & 1 \\ 1 & 0.4 \end{bmatrix}$ for horizontal and vertical preferences respectively. We set the congestion function to its symmetric Cobb-Douglas form $\phi_{xy} = \left(\mu_x p_x(y) \frac{\mu_y}{\delta_x} \mu_y q_y(x) a_{xy}\right)^{-\frac{1}{2}}$. We calibrate the outside option of the workers b to approximately 70% of their product of labor $f_{xy} + b$, in the middle of the range of calibrations of DMP models. This implies a value of $b/f_{xy} = 2.5$ and a maximum markdown of 40%.

We consider a two-dimensional subset of all combinations of $\{\theta_x, \theta_y, \mu_x, \mu_y\}$ by assuming equal costs $\theta_x = \theta_y = \theta$ and by varying the ratio of the number of jobs to

Figure 3.1: Equilibrium monopsony power and wage markdowns for vertical case

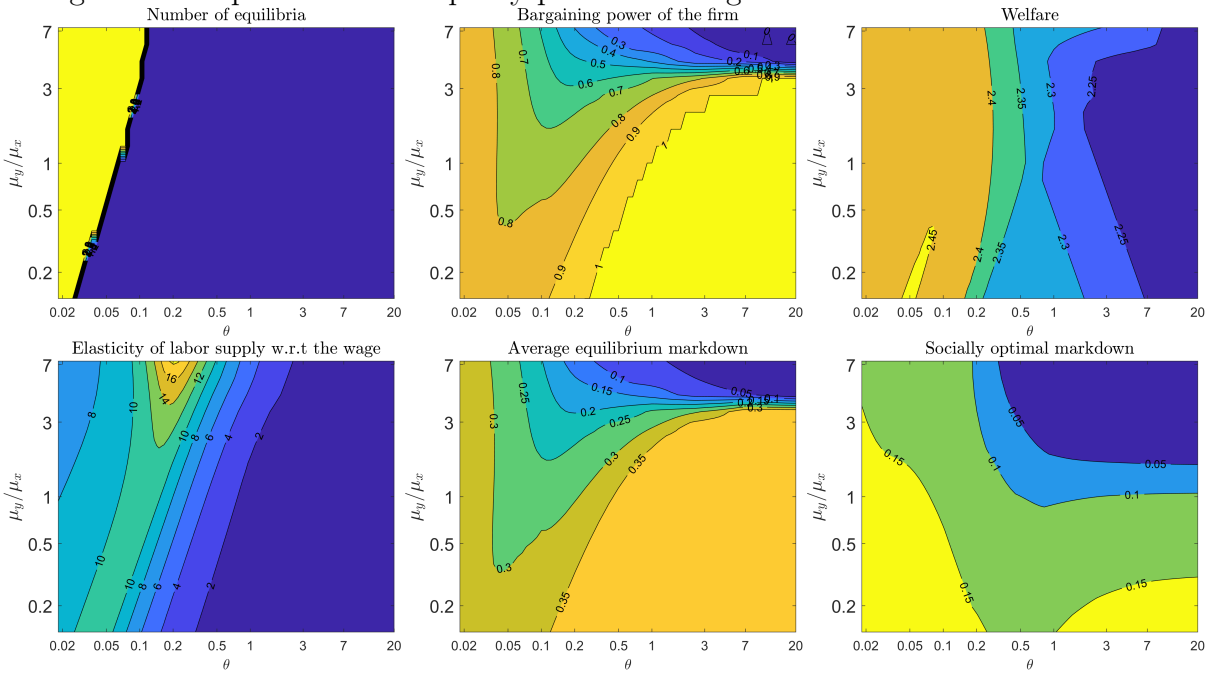
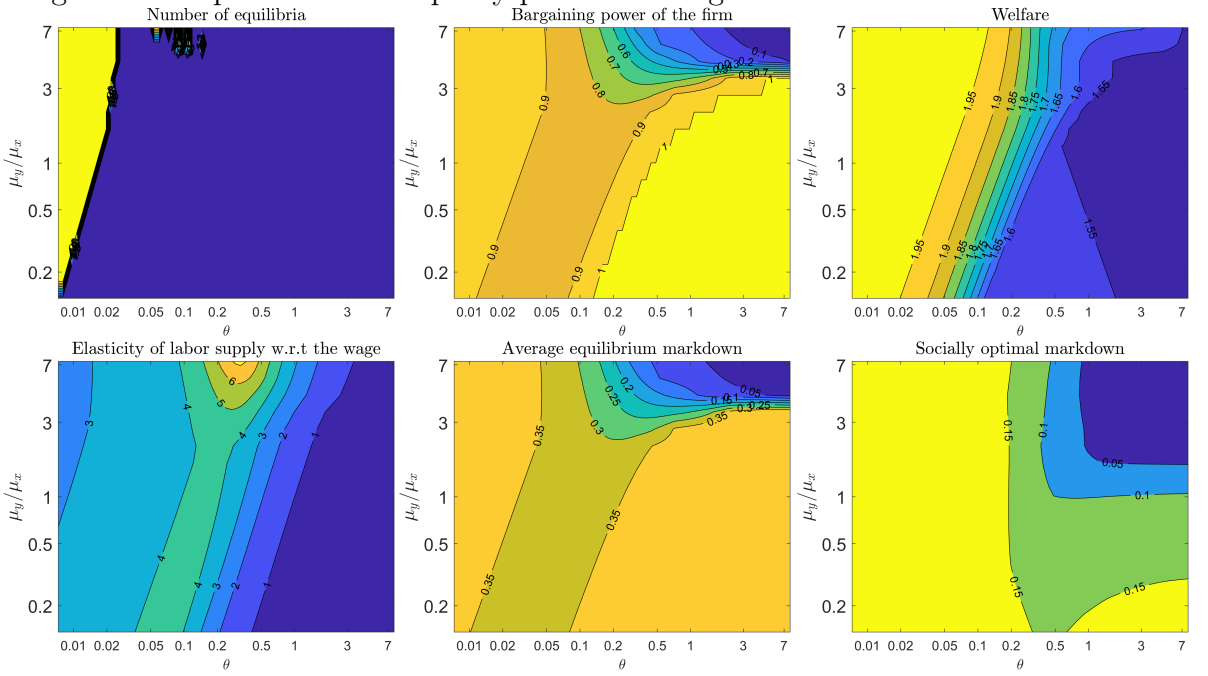


Figure 3.2: Equilibrium monopsony power and wage markdowns for horizontal case



workers μ_y/μ_x, a variable also known as labor market tightness. For each combination of $\{\theta, \mu_y/\mu_x\}$, we compute the equilibrium by making an initial guess for the strategies of the workers and firms, computing the equilibrium posted wage, and then checking if the optimality conditions for the remaining strategies are satisfied. We vary the vector of strategies until we find a fixed point.

In Figures 3.1 and 3.2 we show in six panels how the number of equilibria, monopsony power of (share of the surplus going to) the firm, the elasticity of labor supply with respect to the wage, the average equilibrium markdown, the socially optimal markdown, and welfare — vary with search costs and labor market tightness. The first Figure shows the results for a vertical structure of preferences which produces mixed sorting in equilibrium. The second Figure shows the results for a horizontal structure of preferences which produces positive assortative matching in equilibrium. In both cases there is an area of low costs producing 3 equilibria: PAM, mixed, and NAM — with only the mixed equilibrium surviving for higher costs in the vertical case, and only PAM in the horizontal case.

Overall, the patterns of monopsony power and markdowns are not very different between the two surplus calibrations. Outcomes corresponding to labor market tightness in the range from 0.5 to 1.2, which are routinely observed in the U.S. labor market, produce markdowns on the order of 30%-40%. This corresponds to extremely high levels of monopsony power on the part of firms, capturing between 70% and 100% of match surplus. A constrained social planner would prescribe a tax and redistribution scheme equivalent to increasing wages by about 20%, lowering wage markdowns towards the 10-15% range. Consistent with our theoretical derivations, the elasticity of labor supply is tightly linked with the inverse of marginal search costs faced by workers. Therefore, the empirical estimates of the labor supply elasticity which are usually in the 3-10 range, put a relatively tight bound on the value of search costs θ which should be in the 0.1-0.3 range. As we can see from the numerical simulations, this is well above the level of costs, below which the model produces multiple equilibria.

Figures 3.3 and 3.4 show the equilibrium strategies of workers and firms, the surplus split and numbers of matches — for three equilibria (PAM, Mixing, NAM) and for the planner's solution (PO) — allowing us to compare them for vertical and horizontal structure of the surplus for a very low level of search costs. In the vertical case, the mixing equilibrium is qualitatively the closest to the planner's solution, although it

Figure 3.3: Equilibrium strategies and wages for vertical case

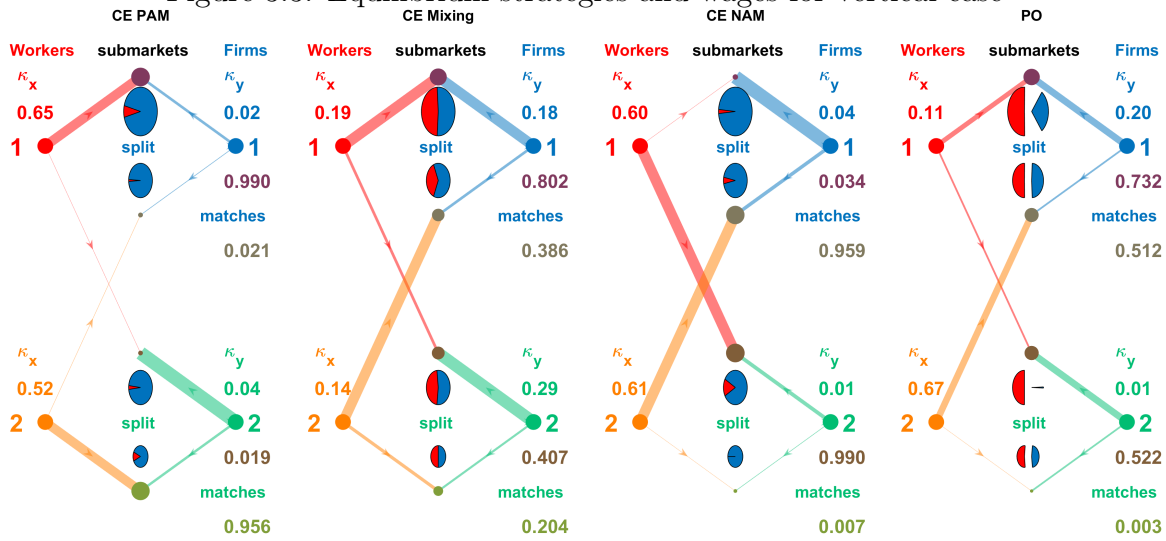
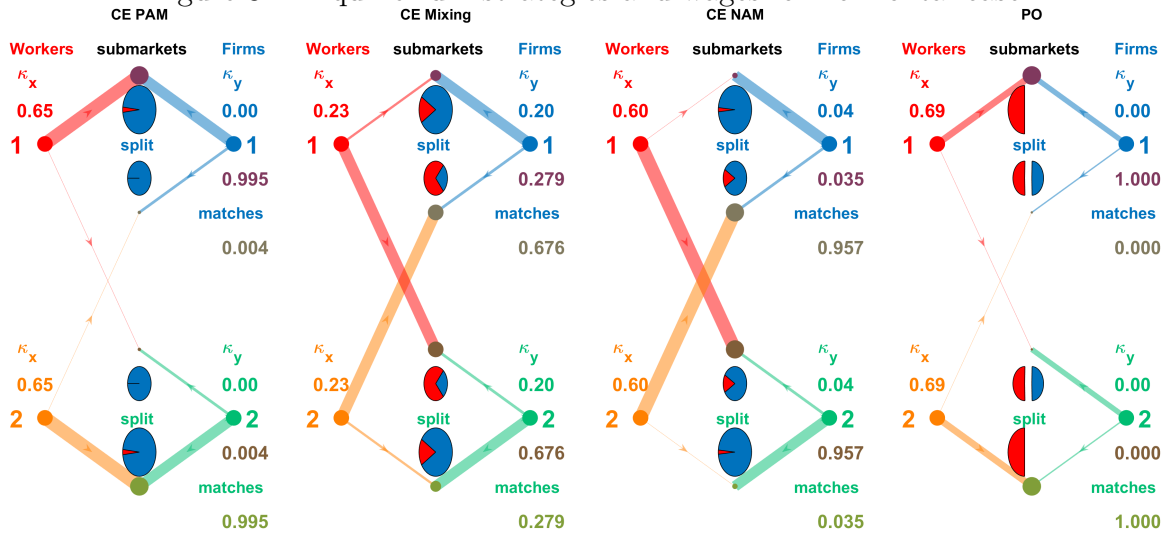


Figure 3.4: Equilibrium strategies and wages for horizontal case



cannot fully achieve it as that would require firms to get substantially less than half the surplus, while workers would still get half. In the horizontal case, the positively assortative equilibrium comes qualitatively closest to the planner's solution, but it promises workers extremely low wages, while the planner still promises workers half the surplus. Notably, in both cases, in each assortative equilibrium, positive or negative, workers are promised extremely low wages. We observe that assortativeness reduces competition for workers and substantially increases the firms' monopsony power.

4 Comparison with the directed search literature

Our numerical findings suggest that firm monopsony power depends crucially on the interplay between search frictions and labor market tightness. This has important implications for the directed search literature, which has traditionally assumed zero search costs. In this section, we situate our model within this broader literature and highlight the novel contributions.

The baseline assumptions of our model mimic all standard assumptions of directed search, as described in the survey of the directed search literature by Guerrieri et al (2021). One caveat is that our model operates under the assumption of an exogenously given number of agents of each type on both sides of the market, but it can be easily extended to allow free entry.

The most standard setup, described in Section 2 of the survey, assumes homogeneous agents and CRS matching probabilities. In our model, this setup could be captured by assuming two identical types of firms and workers, each with measure 1 of agents, a symmetric Cobb-Douglas congestion function, and a constant surplus $f_{xy} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. This case differs from the previously calibrated models only in the shape of the surplus.

The model of Burdett, Shi and Wright (2001) also considers homogeneous workers and firms, but postulates increasing returns to scale in matching rates — by assuming that the probability of a match equals the product of the probabilities with which agents look at each other. This model is represented by assuming two identical types of firms and workers, each with measure 1 agents, a constant surplus $f_{xy} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, and assuming no congestion $\phi_{xy} = const$.

The two numerically calibrated models that we computed in the previous section represent the case of heterogeneity on the side of both workers and firms, under two different assumptions regarding their productive complementarities. These models mimic the assumptions of models in Section 6 of the survey, coming particularly close to Shi (2001) and Eeckhout and Kircher (2010). We assume a standard symmetric Cobb-Douglas matching function, while assuming different other congestion functions would correspond to alternative assumptions about the matching probabilities, such as those assumed in these papers. In particualr, Shi (2001) deviates from the most natural Cobb-Douglas assumption because in this case the model has multiplicity of equilibria

which are hard to fully characterize. Changing the matching function helps side-step this problem. Eeckhout and Kircher (2010) consider various forms of congestion for a continuum of infinitesimal agents, another way around the problem. Our numerical results demostrate sorting patterns and properties of equilibria broadly consistent with their findings. Our model allows consideration of the full set of equilibria under each different specification of congestion, of complementarities in production, under different levels of awareness of agents (abilities to process and use information).

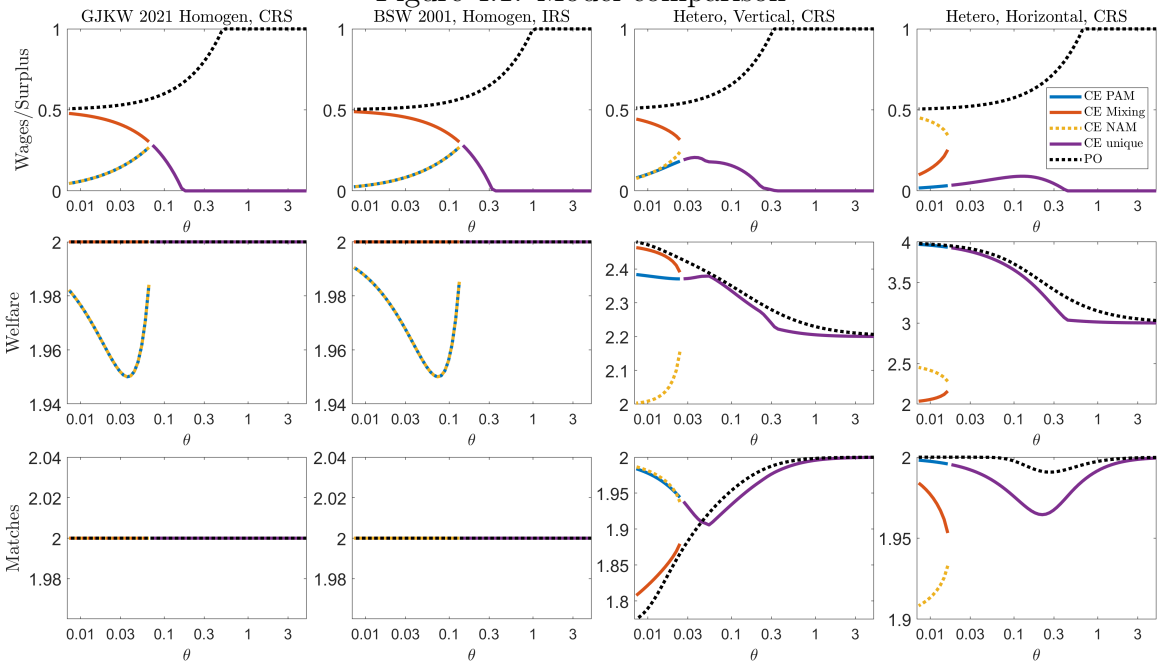
Figure 4.1 shows how the average posted wages, the number of matches, and welfare vary with search costs for various equilibria, and for the social planner's solution. We report results on four models: 1) the directed search model from the survey (homogeneous workers and firms, Cobb-Douglas matching); 2) the model of Burdett et al (2001) (homogeneous workers and firms, increasing returns to scale in matching); 3) the two-sided heterogeneous agents model with vertical preferences (Cobb-Douglas matching, log-submodular production); 4) two-sided heterogeneous agents model with horizontal preferences (Cobb-Douglas matching, log-supermodular production).

First, consistent with Burdett et al (2001), the second model for low costs of search (directed search is obtained in the zero-cost limit) gives three equilibria which we label positive assortative matching (PAM), negative assortative matching (NAM) and a mixing equilibrium. Consistent with Burdett et al (2001), the symmetric mixed-strategy equilibrium produces wages which split the surplus equally between workers and firms. In both pure-strategy equilibria (PAM and NAM) firms take advantage of workers and leave most of the surplus for themselves. In the homogeneous case this has no effect on the number of matches and a small negative effect on welfare. Notably, the mixing equilibrium is socially preferable and survives for higher levels of costs.

Second, unlike the standard homogeneous-agents directed search model with a CRS matching function, which routinely derives a single socially optimal PAM equilibrium, our model produces the same three types of equilibria as in Burdett et al (2001), except the change in the curvature of congestion affects the level of costs at which multiplicity arises. As we shall see from the following discussion, this is not surprising given the absense of production complementarities in this case.

Third, the two outcomes of the heterogeneous-agents directed search model with a CRS matching function are consistent with the derivation of Eeckhout and Kircher (2010) whereby prevalence of sorting depends on the interplay of production comple-

Figure 4.1: Model comparison



mentarity and matching complementarity. Eeckhout and Kircher (2010) predict that for positive assortative matching to prevail, the strength of supermodularity of production needs to exceed the strength of matching complementarity. In the cases we consider, the strength of matching complementarity equals 1, while the production complementarity index $\frac{f_{xy}f}{f_x f_y}$ equals 0 for the first two cases, is between 0 and 1 for the vertical surplus case, and exceeds 1 for the horizontal surplus case. While we get three equilibria for low search costs, for higher costs, consistent with the prediction, the PAM equilibrium prevails in the horizontal case, but the mixing equilibrium invariably survives and leads to highest welfare for the remaining three cases.

To summarize, the wage-posting competitive equilibrium in the sequential targeted search model fills in the continuum between random matching (when $\theta \to \infty$) and directed search (when $\theta \to 0$). This model features all the defining assumptions of directed search models: a) search is sequential: firms post wages, workers apply, firms choose among those that applied; b) firms post type-specific wages strategically such that they attract specific kinds of workers; c) after deciding direction of search, workers and firms meet in submarkets each featuring a matching technology which determines the number of matches. The novel feature of our model is that we fill in this continuum

by varying the degree to which firms and workers are able to inform themselves about the available options.

We present several novel findings that shed light on properties of directed search models. First, for low enough levels of search costs, there are generally multiple equilibria. These are likely assumed away in the literature either through focusing only on specific types of equilibria, or due to the method of constraining agents to deliver a certain utility value to the other side. Second, while we find competitive equilibria to be generically constrained-inefficient, in the zero-cost limit, one of the equilibria typically approaches the planner's solution. Third, whether the best competitive equilibrium exhibits positive assortative matching depends on whether the strength of production complementarities exceeds matching complementarities, consistent with Eeckhout and Kircher (2010). Fourth, when production complementarities are strong, positive assortativeness implies that firms face little competition for workers, and therefore gain monopsony power and use it to reduce promised wages. Fifth, when production complementarities are weak, the best competitive equilibrium tends to exhibit a mixed sorting pattern, rather than negative assortative matching. We find that mixed sorting patterns are typically characterized by intense competition for workers which leads to a more even split of the surplus, consistent with constrained-efficiency of equilibria in the directed search limit.

5 Conclusion

This paper develops an information-based model of monopsony power that unifies directed and random search frameworks, offering a novel perspective on wage-setting in imperfectly competitive labor markets. By incorporating information frictions into workers' and firms' search behaviors, our model captures the full spectrum of search outcomes, from competitive wage-setting under minimal frictions to strong monopsony power when search costs are high. The closed-form wage solution derived in this framework highlights four distinct sources of monopsony power: firms' first-mover advantage, labor market tightness, search cost asymmetries, and productive complementarities. These findings challenge the traditional one-to-one relationship between wage markdowns and labor supply elasticity, showing that monopsony power is a function of strategic wage-setting rather than just labor supply responsiveness.

Numerical simulations further illustrate the model's implications, demonstrating that empirically observed wage markdowns of 30-40% can be explained by plausible search cost levels and labor market tightness conditions. The model also reveals that multiple equilibria may arise when search frictions are low, with the equilibrium selection playing a critical role in determining wage outcomes and monopsony power. Importantly, a constrained-efficient social planner would prescribe a reallocation of surplus that reduces wage markdowns and raises worker pay, suggesting a potential role for policy interventions. By bridging gaps between theoretical monopsony models and empirical wage-setting patterns, this framework provides a richer understanding of labor market frictions and offers new avenues for research on wage determination, market structure, and the role of policy in mitigating monopsony distortions.

References

[1] **Azar, J. and I. Marinescu (2024)** "Monopsony Power in the Labor Market" Chapter 10 in *Handbook of Labor Economics*, Volume 5, pp. 761-82. https://doi.org/10.1016/bs.heslab.2024.11.007

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

[3] **Card, D. (2022)** "Who Set Your Wage?" *American Economic Review*, 112(4), pp. 1075-1090. https://doi.org/10.1086/322835

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

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

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

[7] **Choo, E. and A. Siow (2006).** "Who Marries Whom and Why." *Journal of Political Economy*, 114(1), pp. 175-201. https://doi.org/10.1086/498585

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

[9] **Galichon, A. and B. SalaniÃ© (2012).** "Cupid's Invisible Hand: Social Surplus and Identification in Matching Models." Working Paper hal-01053710

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

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

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

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

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

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

[16] **Shimer, R. and L. Smith (2000).** "Assortative Matching and Search." *Econometrica*, 68(2), pp. 343-369. https://doi.org/10.1111/1468-0262.00112

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

[18] **Shimer, R. (2005).** "The Assignment of Workers to Jobs in an Economy with Coordination Frictions." *Journal of Political Economy*, 113, pp. 996-1025. https://doi.org/10.1086/444551

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

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

[21] **Wu L. (2020).** "Partially Directed Search in the Labor Market." mimeo.

[22] **Yang, H. (2013).** "Targeted Search and the Long Tail Effect." *The RAND Journal of Economics*, 44(4), pp. 733-756. https://doi.org/10.1111/1756-2171.12036