Search, Screening and Sorting*

Xiaoming Cai† Pieter Gautier‡ Ronald Wolthoff§

April 3, 2024

Abstract

We examine how search frictions impact labor market sorting by constructing a model consistent with recent evidence that employers collect a pool of applicants before interviewing a subset. We derive the necessary and sufficient conditions for sorting in both applications and matches. Positive sorting is obtained when production complementarities outweigh a force against sorting measured by a (novel) *quality-quantity elasticity*. Interestingly, the production complementarities needed for positive sorting depend on the population fraction of high-type workers and can be *increasing* in the number of interviews. Our model shows how policies like Ban the Box can backfire because when attractive firms can no longer perfectly screen certain workers, they may discourage them from applying.

JEL codes: D82, D83, E24.

Keywords: sorting, complementarity, search frictions, information frictions, heterogeneity.

*We are grateful to the editor (Ayşegül Şahin) and referees for their valuable feedback, and to Jim Albrecht, Axel Anderson and Katka Borovičková for insightful discussions of this paper. We thank Steve Davis, Philipp Kircher, Robert Shimer, and various seminar and conference participants for valuable comments. Wolthoff is grateful for financial support from the Social Sciences and Humanities Research Council.

†Peking University HSBC Business School. E-mail: xmingcai@gmail.com.

‡VU University Amsterdam and Tinbergen Institute. E-mail: p.a.gautier@vu.nl.

§University of Toronto. E-mail: ronald.p.wolthoff@gmail.com.

1 Introduction

One of the most important tasks for any firm is to hire the right workers. A crucial part of this process consists of screening applicants through job interviews.[1] In this paper, we are interested in the question how such screening affects sorting patterns in the labor market. That is, if technological innovations allow firms to screen more applicants with higher precision, does that make sorting more or less likely?[2]

Unfortunately, the economic literature is silent on these questions. The earliest work on assignment problems (Tinbergen, 1956; Shapley and Shubik, 1971; Becker, 1973) considers frictionless environments with no role for screening since there is full information about types. More recent work by Shimer and Smith (2000), Shi (2001, 2002), Shimer (2005) and Eeckhout and Kircher (2010) allows for search frictions but makes particular assumptions about the available information in the matching process and does not explore how outcomes depend on them.

To answer our question, we therefore present a new directed search model of the labor market. In line with recent evidence by Davis and Samaniego de la Parra (2017), we allow firms to interview multiple (but not necessarily all) applicants before making a job offer. We show how the equilibrium allocation of workers to firms in this environment depends on the degree of production complementarities on the one hand and the extent to which firms can interview applicants on the other hand. Perhaps surprisingly, we find that reducing frictions by allowing firms to interview more workers can be a force *against* sorting.

To explain this result, we must first describe our setup in more detail. We consider a static environment in which heterogeneous firms compete for heterogeneous workers by posting menus of type-contingent wages. Workers direct their search to the menu that maximizes their expected payoff. This choice determines the expected number of low- and high-type applicants (the 'queue') at each firm. The realized numbers are stochastic due to coordination frictions. As mentioned, the

[1] See below for some empirical evidence. Note that 'screening' in this context has a different meaning than the homonymous game-theoretic concept. In addition to job interviews, screening workers may involve other instruments like checking references, assessments, and job tests. We use 'interview' as shorthand for the entire collection of instruments.

[2] As an example of such a technological innovation, Hoffman et al. (2018) describe how some firms subject all applicants to an online job test. Based on their answers, every applicant is assigned a score, calculated from correlations between answers and job performance among existing employees.

key innovation is that we allow firms to interview a subset of applicants, which reveals their types. Firms hire the most profitable candidate among their interviewees and the match produces output according to a general production function.

Firms in this environment face a trade-off. Attracting low-type applicants can be beneficial because the search frictions imply that it is always possible that no high type applies, in which case hiring a low type is better than remaining unmatched. However, this kind of insurance comes at a cost, because the presence of low types makes it harder for the firm to identify the high types in the applicant pool. Clearly, the magnitude of the cost is smaller if firms can screen more, so firms' decision what applicant pool to attract ex ante depends on the extent to which they can screen workers ex post.

We start our analysis by considering the problem of a planner who chooses the queue for each firm to maximize the expected total net output (or surplus). At the optimum, the marginal values of applicants of each type are the same across different firms. An applicant directly contributes to surplus if no other applicant with the same or better type is being interviewed. However, when firms cannot screen everyone, an applicant also affects surplus by making it harder for other (potentially more-productive) applicants to be interviewed.

We then turn to sorting. Given the meaningful distinction between applicants and hires in our environment, we analyze sorting along both dimensions. We define *positive assortative matching* (PAM) as first-order stochastic dominance in the distribution of hires, and introduce *positive assortative contacting* (PAC) as the corresponding concept for the distribution of applicants.[3]

To analyze when the planner's solution exhibits positive or negative sorting, it is helpful to focus on the boundary between both cases where the planner's solution exhibits no sorting. At this boundary, complementarities in production imply that more-productive firms have longer queue lengths. This longer queue length reduces the probability that a marginal high-type applicant creates surplus, which discourages more-productive firms from attracting such applicants and therefore forms a force against positive sorting. This force is captured by an elasticity which we label the *quality-quantity elasticity* (and which differs between PAC and PAM). Whether positive sorting is optimal depends on whether the complementarities in

[3]We also provide results for negative assortative contacting (NAC) and matching (NAM). We omit intuition for those results here as it mirrors the intuition for PAC and PAM.

production are large enough to offset this force.

The relevance of production complementarities for sorting has been known since Becker (1973). The quality-quantity elasticity, however, is novel and we view its characterization as one of our main contributions. When a firm attracts more low-type and high-type applicants, an individual applicant's marginal contribution to surplus falls. The quality-quantity elasticity measures how fast the probability that a high-type worker contributes to surplus decreases relative to the same probability for a low-type worker. The larger it is, the stronger the force against positive sorting and the larger production complementarities therefore need to be to offset this force and induce positive sorting.

The quality-quantity elasticity is not only economically intuitive but also simple in the sense that it only depends on the meeting technology (queue length, queue composition and the degree of screening). To understand the dependence, note that there are two scenarios in which a high-type applicant fails to create surplus: (1) he is not interviewed, (2) he is interviewed, but at least one other high-type applicant is interviewed as well. Both scenarios become more likely as the queue length increases. The first scenario is the most relevant one when the applicant pool mainly consists of low-type workers (the effect of a longer queue at more-productive firms predominantly operates by making it less likely for a high-type applicant to be interviewed). The second scenario is the most relevant one when the applicant pool mainly consists of high types (multiple interviews with high-type applicants are a key concern and a longer queue makes this outcome more likely).

To ensure positive sorting for any distribution of agents' types, the infimum of the elasticity of complementarity should exceed the supremum of the quality-quantity elasticity. We show that this bound on the quality-quantity elasticity is attained when high-type workers are abundant, because the probability that a high type creates surplus is most sensitive to the queue length in that case.

Finally, we analyze how the quality-quantity elasticity varies with the degree of screening. Viewing increased screening as a relaxation of the frictions in the environment, one may expect that it must facilitate sorting. We show that while this intuition is correct when high-type workers are scarce, it is wrong when they are abundant. To understand this result, note that an increase in firms' screening ability *mitigates* the force against positive sorting in the first scenario above

(as it becomes easier to identify the rare high-type applicant) but *amplifies* it in the second scenario (as it becomes increasingly likely that multiple high types are interviewed).

When deriving a sorting condition for any distribution of agents' types, the tightest condition matters, which is again the second. The elasticity of complementarity that is necessary and sufficient for sorting in this case is thus *increasing* in the expected number of interviews that firms can conduct, ranging from $\frac{1}{2}$ (square-root-supermodularity) with a single interview to 1 (log-supermodularity) when firms can interview all their applicants.

Our model has important implications for policies that affec the information that is available during the matching process. We illustrate this using the case of "Ban the Box" policies which aim to help ex-offenders find better jobs by revealing their criminal history later in the recruiting process. We find that such policies can backfire because when it becomes harder for productive firms to screen applicants ex post, they may discourage less desirable workers from applying in the first place. In equilibrium, the policy causes ex-offenders to find worse jobs, hurting the workers it was meant to help.

The paper is organized as follows. The remainder of this section discusses related literature. Section 2 introduces the model. Section 3 formulates the planner's problem and shows that the decentralized equilibrium implements the planner's solution. Section 4 derives our main sorting results. In Section 5, we illustrate how our framework yields predictions regarding the effects of the "Ban the Box" (BTB) policy on ex-ante and ex-post screening. Finally, Section 6 concludes, while proofs and additional results can be found in the (online) appendix.

Related Literature. We primarily contribute to the theoretical literature on sorting in markets with search frictions. This literature dates back to Shimer and Smith (2000) who showed that search frictions are a force against positive sorting, because the opportunity cost of remaining unmatched is larger for high types, which makes them more eager to match with a low type rather than run the risk to not match at all. To undo this effect, the production function must exhibit stronger complementarities than the supermodularity condition that prevails in a Walrasian world (Becker, 1973).

Most related to our work, Eeckhout and Kircher (2010) show that under directed

search (but with a single interview per firm) PAM requires that the elasticity of complementarity exceeds the elasticity of substitution of the aggregate meeting function. As mentioned, the relevant threshold for sorting in our environment with simultaneous interviews is the *quality-quantity elasticity*. Like the threshold in Eeckhout and Kircher (2010), this elasticity depends on the properties of the meeting technology only. However, a crucial difference is that the quality-quantity elasticity depends not only on the queue length but also on the queue composition and the degree of screening. It reduces to the threshold in Eeckhout and Kircher (2010) when firms can only screen a single worker, but may increase in magnitude as screening becomes easier.

Some papers have argued that increased sorting of high-type workers at high-wage firms has contributed to the observed increased inequality from the mid-nineties onwards (see e.g. Card et al., 2013; Song et al., 2019).[4] Håkanson et al. (2018) argue that the increased sorting patterns are mainly due to increasing complementarities in production. Our results suggest that if during the same period, new technologies like automated resume screening made it easier to screen workers, then this would require even stronger complementarities in the production technology.

Our results also have important implications for the empirical literature that deals with both the sign and the strength of sorting (Gautier and Teulings, 2006; Eeckhout and Kircher, 2011; Gautier and Teulings, 2015; Lise et al., 2016; Hagedorn et al., 2017; Lopes de Melo, 2018; Bartolucci et al., 2018; Bagger and Lentz, 2018; Borovičková and Shimer, 2020). An important aim of this literature is to identify the shape of the production function from observed matching patterns. In general, a particular meeting technology is assumed and then the strength and sign of sorting are used to identify key parameters of the production function.[5] Our findings imply however that such assumptions are not innocuous and that the meeting technology needs to be identified alongside the production function. Progress along this dimension is facilitated by our theoretical results on PAC/PAM combined with recent empirical work by Banfi et al. (2020) who document evidence for PAC as well

[4] Card et al. (2013) use education and occupational sorting.

[5] Since wages for a given worker type are typically non-monotonic in firm types, the methodology by Abowd et al. (1999) of detecting sorting patterns from simply correlating worker and firm fixed effects fails; the cited papers propose various ways to deal with this.

as PAM using data from a Chilean online job board. In a similar vein, the strength of sorting is often used to estimate how far an economy is from the frontier. Our results show that stronger sorting patterns do not necessary imply lower frictions. Gautier and Teulings (2006, 2015) and Lise et al. (2016) estimate the output loss due to search frictions. In their models, more frictions imply more output loss and more mismatch. In this paper, we show that while more frictions always implies less output, it may sometimes imply less mismatch.

Our paper also adds to a recent macro literature that focuses on information frictions. Both Kurlat (2016) and Board et al. (2019) consider a competitive model with heterogeneity in productivity on the worker side and heterogeneity in screening ability on the firm side; workers essentially apply to every firm, so screening only takes place ex post.[6] Unlike their work, we consider firms that are heterogeneous in productivity, making it possible to analyze varying degrees of complementarity in production and a more conventional notion of sorting. We further emphasize the frictional nature of most labor markets and allow for ex-ante screening through workers' applications decisions in addition to ex-post screening, showing that firms typically use a combination. In appendix A.12 where we endogenize screening, the firms who benefit most from screening (rather than the most productive firms as in Board et al., 2019) invest most in screening.

Finally, although our focus is on the labor market, our results are also important for other markets with matching between heterogeneous agents and a role for screening, such as the housing market or the marriage market. Also in trade, there is a growing interest in deriving patterns of international specialization (i.e. under which conditions do exporters hire the most productive workers) from fundamental properties of the production technology, see Costinot (2009). More generally, the interaction between quality (attracting high-type workers) and quantity (attracting low types as well) has been little studied in economics and we expect our analysis to be useful beyond the questions we address here.

[6]The main difference between the two papers is that the screening outcomes of a worker at different firms are independent in Board et al. (2019), whereas in Kurlat (2016) they are conditionally perfectly correlated across firms (if an applicant passes one firm's test, this candidate will pass the test of all firms with worse screening skills).

2 Model

Agents. A static economy is populated by a measure 1 of firms and a measure $L > 0$ of workers. All agents are risk neutral. Each firm demands and each worker supplies a single unit of indivisible labor. Each firm is characterized by a type $y \in \mathcal{Y} = [\underline{y}, \overline{y}] \subseteq \mathbb{R}_+$. The measure of firms with types weakly below y is denoted by $J(y)$, where $J(\overline{y})$ is normalized to one. Similarly, each worker is characterized by a type $x \in \mathcal{X} = [\underline{x}, \overline{x}] \subseteq \mathbb{R}_+$. In particular, a fraction $z \in (0,1)$ of workers has a low type x_1 and the remaining workers have a high type x_2, with $0 < x_1 < x_2$. The distribution of agents' types in the economy is thus $(x_1, x_2, L, z, J(y))$.

Wage Menus and Search. Each firm commits to a wage menu $\boldsymbol{w} = (w_1, w_2)$, where w_i is the wage for a hire of type x_i. Workers observe all wage menus and apply to one, taking into account that there will be more competition at high wages.[7] We initially assume that workers also observe firm types, but then show that this assumption is redundant because workers only care about their expected payoff, which they can infer from the wage menu alone. We capture the anonymity of the large market with the standard assumption that identical workers must use symmetric strategies (see e.g. Shimer, 2005).

A *submarket* $(\boldsymbol{w}, y)$ consists of the firms of type y that post a wage menu $\boldsymbol{w}$ and all workers who apply to such a menu. For each submarket, we denote the ratio of the number of high-type applicants to the number of firms by $\mu(\boldsymbol{w}, y)$, and the ratio of the total number of applicants (regardless of their type) to the number of firms by $\lambda(\boldsymbol{w}, y)$. Naturally, these ratios—or *queue lengths*—are endogenous and satisfy $0 \leq \mu(\boldsymbol{w}, y) \leq \lambda(\boldsymbol{w}, y)$ for all $(\boldsymbol{w}, y)$.

Benchmark Frictions. Our benchmark matching process features two stages (applying and screening) and was introduced by Cai et al. (2022). To understand it, consider a submarket with queues (μ, λ). Workers and firms in the submarket are randomly located on the circumference of a circle according to a uniform distribution. Workers apply clockwise to the nearest firm.[8] The probability that a firm

[7] A single chance to match (per period) is standard and captures the idea that (opportunity) costs are associated with applying. Work relaxing this assumption uses (ex ante) homogeneous agents (Albrecht et al., 2006; Galenianos and Kircher, 2009; Kircher, 2009; Wolthoff, 2018; Albrecht et al., 2019), except Auster et al. (2021) which considers one-sided heterogeneity.

[8] When workers cannot keep track of distance, this is merely a tie-breaking rule.

receives n applications depends only on λ (constant returns to scale), and is given by $\frac{1}{1+\lambda}(\frac{\lambda}{1+\lambda})^n$ for $n = 0, 1, 2, \ldots$, which is a geometric distribution with mean λ.[9] In the screening stage, each firm interviews its applicants in a random order. An interview allows the firm to learn the type of the applicant, which is x_2 with probability μ/λ. After every interview, and conditional on applicants remaining, there is an exogenous probability $\sigma \in [0, 1]$ that the firm can conduct another interview, while interviewing stops with complementary probability.

Our setup nests two common but extreme specifications of the meeting technology as special cases. If $\sigma = 0$, each firm can interview only one applicant, as in the bilateral model of Eeckhout and Kircher (2010). In this case, the presence of low-type applicants makes it harder for firms to identify a high type in their applicant pool. Increasing σ reduces this meeting externality. It disappears entirely when σ reaches 1 and firms can interview all their applicants. As in the urn-ball setup of Shimer (2005), firms' chances of finding a high type in their applicant pool then become independent of the number of low-type applicants—a property known in the literature as *invariance* (see Lester et al., 2015; Cai et al., 2017).

It is worth pointing out that our analysis does not depend on this particular microfoundation; it can be applied to other two-stage matching processes, as long as the first stage treats workers symmetrically, irrespective of their types.[10]

Matching and Production. After the interviews have been conducted, matches are formed. Firms can only hire a worker which they have interviewed.[11] If a firm has interviewed multiple applicants, it hires the most profitable one. A match between a worker of type x and a firm type of y produces net output $f(x, y) > 0$, which is twice continuously differentiable. The partial derivatives $f_x(x, y)$ and $f_y(x, y)$ are strictly positive for all (x, y), and the cross-partial is denoted by $f_{xy}(x, y)$.[12] From the produced output, the firm pays the worker the promised wage w_i and keeps the

[9] Note the subtle difference compared to an equidistant positioning of firms, which yields a Poisson number of applicants with mean λ, as in an urn-ball technology.

[10] That is, the probability that a firm receives at least one applicant depends only on λ, i.e. independent of μ, and the expression of surplus in equation (3) stays valid.

[11] This assumption can easily be rationalized by introducing a small chance that any given worker provides the firm with a sufficiently negative payoff when hired.

[12] Although worker types are binary, our objective to find a sorting condition for any distribution of agents' types requires that f is defined on the full domain $\mathcal{X} \times \mathcal{Y}$ rather than only for given x_1 and x_2.

rest. Firms and workers which fail to match obtain a zero payoff.

Elasticity of Complementarity. For our analysis, a key characteristic of the production function is its *elasticity of complementarity* (Hicks, 1932, 1970), which is the inverse of the elasticity of substitution. It is defined as

$$\rho(x,y) \equiv \frac{f_{xy}(x,y)f(x,y)}{f_x(x,y)f_y(x,y)} \in \mathbb{R}, \tag{1}$$

with extrema $\overline{\rho} \equiv \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \rho(x,y)$ and $\underline{\rho} \equiv \inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \rho(x,y)$. This elasticity is closely related to the notion of n-root-supermodularity, as defined in Eeckhout and Kircher (2010).[13]

Definition 1. *The function $f(x,y)$ is* n-root-supermodular *if and only if $\rho(x,y) \geq 1 - 1/n$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$; special cases include supermodularity ($n = 1$) and log-supermodularity ($n \to \infty$). When $\rho(x,y) \leq 1 - 1/n$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $f(x,y)$ is said to be* n-root-submodular.

In other words, n-root-supermodularity is equivalent to $\underline{\rho} \geq 1-1/n$ and n-root-submodularity is equivalent to $\overline{\rho} \leq 1 - 1/n$.

Special Case. We will sometimes illustrate our results with a CES production function, because it has a constant elasticity of complementarity, $\rho(x,y) = \rho$. That is, $f(x,y) = (\alpha x^{1-\rho} + (1-\alpha)y^{1-\rho})^{\frac{1}{1-\rho}}$ where $\alpha \in (0,1)$. This production function is submodular when $\rho \leq 0$, $\frac{1}{1-\rho}$-root-supermodular when $0 < \rho < 1$, and log-supermodular when $\rho \geq 1$.

3 Planner's Problem and Market Equilibrium

In this section, we first derive surplus within a submarket. We then analyze the problem of a social planner who aims to maximize surplus (net output) subject to the search frictions. Finally, we show that the planner's solution is the equilibrium outcome when firms post wage menus.

[13] Eeckhout and Kircher (2010) define $f(x,y)$ to be n-root-supermodular if $\sqrt[n]{f(x,y)}$ is supermodular. Since $\frac{1}{\partial x \partial y}\sqrt[n]{f} = n^{-2}f^{1/n-2}\left(ff_{xy} - (1-\frac{1}{n})f_x f_y\right)$, our definition is equivalent.

3.1 Surplus Within a Submarket

Interview Probability. A firm in a submarket with queues (μ, λ) hires a high-type worker if and only if it interviews *at least* one such worker. The following lemma, borrowed from Cai et al. (2022), derives the probability of this event.[14]

Lemma 1 (Cai et al., 2022)**.** *In a submarket with queues* (μ, λ)*, the probability that a firm interviews at least one high-type worker equals*

$$\phi(\mu, \lambda) = \frac{\mu}{1 + \sigma\mu + (1 - \sigma)\lambda}. \tag{2}$$

Proof. See Appendix A.1. □

As Cai et al. (2022) show, $\phi(\mu, \lambda)$ is useful for multiple reasons. First, $\phi(\mu, \lambda)$ is sufficient to summarize the meeting process within a submarket. It not only describes the probability that the firm will hire a high-type worker, but—upon evaluation in $\mu = \lambda$—also the firm's overall matching probability (regardless of the hire's type), which we denote by $m(\lambda) \equiv \phi(\lambda, \lambda)$. Hence, the probability that the firm hires a low-type worker is given by $m(\lambda) - \phi(\mu, \lambda)$.

Second, the partial derivatives of $\phi(\mu, \lambda)$ have economically meaningful interpretations. The partial derivative $\phi_\lambda(\mu, \lambda) \leq 0$ captures recruiting externalities as it describes how a firm's chances to hire a high-type worker change if the queue of low-type workers gets longer. As discussed before, these externalities are absent, i.e. $\phi_\lambda(\mu, \lambda) = 0$, if and only if all applicants are interviewed ($\sigma = 1$).

In contrast, $\phi_\mu(\mu, \lambda)$ describes how a firm's probability of hiring a high-type worker changes if the queue of such workers increases, while the total queue remains constant (i.e. changing the composition of the applicant pool). From the perspective of a high-type applicant, this partial derivative represents the probability to be hired and to increase surplus because no other high-type worker was interviewed.[15]

Properties. The expression in (2) has the following intuitive properties:

[14] Cai et al. (2022) study market segmentation in a world with homogeneous firms. Our focus is quite different, so we provide a derivation of $\phi(\mu, \lambda)$ for completeness.

[15] To see this, note that $\phi_\mu(\mu, \lambda)\Delta\mu = \phi(\mu + \Delta\mu, \lambda) - \phi(\mu, \lambda)$ represents the probability that replacing $\Delta\mu$ low-type workers with high types generates additional surplus. Naturally, this is the case if and only if these $\Delta\mu$ workers are the only high types that are interviewed.

A0. $\phi(\mu, \lambda)$ is strictly increasing and concave in μ: replacing low-type workers with high-type workers in a submarket increases a firm's probability of interviewing at least one high-type worker, but at a decreasing rate;

A1. Let ζ be the fraction of high-type workers, then for any given $\zeta \in (0, 1]$, $\phi(\lambda\zeta, \lambda)$ is strictly increasing and strictly concave in λ: holding the fraction of high-type workers constant, adding more workers to the submarket increases a firm's probability of interviewing at least one high type, but at a decreasing rate;

A2. for any given $\zeta \in (0, 1]$, $\phi_\mu(\lambda\zeta, \lambda)$ is strictly decreasing in λ: holding the fraction of high-type workers constant, adding more workers to the submarket reduces the probability that a high-type worker creates surplus.

Surplus. We can now derive expected surplus. With probability $m(\lambda) = \phi(\lambda, \lambda)$, a firm of type y facing queues (μ, λ) receives at least one application, generating at least a surplus $f(x_1, y)$; with probability $\phi(\mu, \lambda)$, the firm interviews at least one high-type worker, generating an additional surplus $f(x_2, y) - f(x_1, y)$. Expected surplus is thus

$$S(\mu, \lambda, y) = m(\lambda) f(x_1, y) + \phi(\mu, \lambda) [f(x_2, y) - f(x_1, y)]. \tag{3}$$

The marginal contributions to surplus by firms and workers can be derived by taking partial derivatives of $S(\mu, \lambda, y)$, and are given in Appendix A.2. The concavity of $S(\mu, \lambda, y)$ is analyzed in Appendix A.3.

3.2 Optimal Allocation of Workers and Firms

After deriving surplus, we now turn to the allocation of workers and firms. We first consider the case in which firms are homogeneous in productivity, as it provides a helpful building block for the analysis of heterogeneous firms.

Homogeneous Firms and the Concave Envelope. Even when all firms have the same productivity y, the planner's problem is non-trivial because the surplus function $S(\mu, \lambda, y)$ is not globally concave (unless $\sigma = 1$; see Appendix A.3). As a result, the planner may wish to create multiple submarkets. Let K be this number and let γ_i, μ_i and λ_i be the measure of firms, the queue length of the high-type, and the queue length of both types of workers in submarket i, respectively. The

planner's problem is then

$$\widehat{S}(Lz, L, y) \equiv \max_{K \geq 1, \{\gamma_i, \mu_i, \lambda_i\}} \quad \sum_{i=1}^{K} \gamma_i S\left(\mu_i, \lambda_i, y\right),$$

subject to $\sum_{i=1}^{K} \gamma_i = 1$, $\sum_{i=1}^{K} \gamma_i(\lambda_i - \mu_i) \leq L(1-z)$, and $\sum_{i=1}^{K} \gamma_i \mu_i \leq Lz$.

This formulation makes it clear that the maximal surplus $\widehat{S}(\mu, \lambda, y)$ that the planner can create is the concave envelope (or the *least concave majorant*) of $S(\mu, \lambda, y)$, i.e. the smallest concave function that is greater than $S(\mu, \lambda, y)$. In general, finding the concave envelope of a non-concave function is challenging. However, Cai et al. (2022) show that if $\phi(\mu, \lambda)$ satisfies a single-crossing condition, which is the case for (2), the planner's solution is unique and takes a simple form with at most two submarkets. The following lemma presents this result.

Lemma 2 (Cai et al., 2022)**.** *If ϕ is given by (2) and all firms are homogeneous, then the planner's solution is unique and consists of at most two submarkets, one of which contains all high-type workers and has a shorter total queue.*

Proof. See Appendix A.4. □

Therefore, the planner's problem with homogeneous firms is given by

$$\widehat{S}\left(Lz, L, y\right) = \max_{\gamma, \Delta} \ \gamma S\left(\frac{Lz}{\gamma}, \frac{L-\Delta}{\gamma}, y\right) + (1-\gamma)\, S\left(0, \frac{\Delta}{1-\gamma}, y\right), \tag{4}$$

where $\gamma \in (0, 1]$ is the measure of firms in the first submarket and $\Delta \in [0, L(1-z)]$ is the measure of the low-type workers in the second submarket. In the first submarket, the planner aims for *quality* by allocating all high-type workers and limiting the number of low-type applicants to reduce congestion. In the second submarket, the planner goes for *quantity* and aims for a large hiring probability by allocating many low-type workers but no high-type workers. Note that γ can be 1, in which case the second submarket is inactive. Intuitively, if high-type workers are unlikely to be crowded out by low-type workers, then all firms and workers should form one submarket.

Heterogeneous Firms. When firm productivity is distributed according to $J(y)$, the planner's problem can be formulated as

$$\max_{\overline{\mu}(y),\overline{\lambda}(y)} \int_{\underline{y}}^{\overline{y}} \widehat{S}\left(\overline{\mu}(y),\overline{\lambda}(y),y\right)\, dJ(y), \tag{5}$$

subject to the linear constraints

$$\int_{\underline{y}}^{\overline{y}} \left(\overline{\lambda}(y)-\overline{\mu}(y)\right)\, dJ(y) \leq L(1-z), \tag{6}$$

$$\int_{\underline{y}}^{\overline{y}} \overline{\mu}(y)\, dJ(y) \leq Lz. \tag{7}$$

That is, one can think of the planner's problem as a two-step maximization process. First, the planner chooses $(\overline{\mu}(y),\overline{\lambda}(y))$ for each firm type y, which one can interpret as the *average* queue lengths for these firms. Second, for each firm type y, the planner can divide the firms and the average queues across two submarkets, as in (4), if separating some firms and low-type workers increases surplus. So, $(\overline{\mu}(y),\overline{\lambda}(y))$ is not necessarily the queue faced by a firm of type y; it is a convex combination of two different queues faced by different firms of the same type if and only if $S(\overline{\mu}(y),\overline{\lambda}(y),y) < \widehat{S}(\overline{\mu}(y),\overline{\lambda}(y),y)$. However, when $S(\overline{\mu}(y),\overline{\lambda}(y),y) = \widehat{S}(\overline{\mu}(y),\overline{\lambda}(y),y)$, the planner creates only a single submarket for firms of type y and their queues $(\mu(y),\lambda(y))$ must equal $(\overline{\mu}(y),\overline{\lambda}(y))$.

Although $\widehat{S}(\mu,\lambda,y)$ is concave by construction, it is not *strictly* concave (unless $\sigma = 1$). Hence, the solution to the planner's problem (5)–(7) is not necessarily unique.[16] However, we will later show that uniqueness is guaranteed under the sufficient condition for sorting.

Let W_i be the social marginal value of an application by worker of type $i = 1,2$, i.e. the Lagrange multipliers associated with the resources constraints (6) and (7). Since $\widehat{S}(\mu,\lambda,y)$ is concave, Lagrangian duality implies that if $(\overline{\mu}(y),\overline{\lambda}(y))$ (as a function of y) solves the planner's problem in (5), then for any given y, $(\overline{\mu}(y),\overline{\lambda}(y)) \in \mathbb{R}^2_+$ solves the maximization problem $\max_{\mu,\lambda}\ \widehat{S}(\mu,\lambda,y) - \mu W_2 -$

[16] When the number of firm types is finite, the existence of the planner's solution follows from the standard argument that the objective function is continuous and the domain of the choice variables is compact. With a continuum of firm types, the planner's solution exists by continuity.

$(\lambda - \mu)W_1$. Since $\widehat{S}(\mu, \lambda, y)$ is the concave envelope of $S(\mu, \lambda, y)$, the solution to this problem can be obtained from

$$\max_{\mu,\lambda} \; m(\lambda) f(x_1, y) + \phi(\mu, \lambda) [f(x_2, y) - f(x_1, y)] - \mu W_2 - (\lambda - \mu) W_1. \tag{8}$$

If (8) has exactly one solution, all firms of type y are in the same submarket; otherwise, by Lemma 2, (8) has two solutions and $(\overline{\mu}(y), \overline{\lambda}(y))$ is a convex combination of the queues in the two submarkets in which firms of type y are present.[17]

3.3 Market Equilibrium

We now briefly consider the market equilibrium, and show that when firms post wage menus the equilibrium implements the planner's solution. This can be viewed as a generalization of similar results in Shi (2002), Shimer (2005) and Eeckhout and Kircher (2010). In Appendix A.10, we formally define the directed search equilibrium and provide some technical results that are needed to fill logical gaps.

Payoffs. Consider a firm of type y which posts a wage menu $\boldsymbol{w} = (w_1, w_2)$ and attracts a queue (μ, λ), where w_i is the wage for x_i workers. To simplify exposition, assume for now that the wage menu satisfies

$$f(x_2, y) - w_2 > f(x_1, y) - w_1, \tag{9}$$

i.e. more productive workers are more profitable and are therefore preferred by the firm. In Lemma 10 of Appendix A.10, we show that (9) must indeed hold when firms act optimally, making our assumption without loss of generality. The firm then hires a high-type worker if it interviews at least one such worker, which happens with probability $\phi(\mu, \lambda)$. Similarly, the firm hires a low-type worker if it interviews no high-type workers but at least one low-type worker, which happens

[17] By the definition of $\widehat{S}(\mu, \lambda, y)$, the problem $\max_{\mu,\lambda} \; \widehat{S}(\mu, \lambda, y) - \mu W_2 - (\lambda - \mu) W_1$ can be rewritten as $\max_{K, \gamma_i, \mu_i, \lambda_i} \sum_{i=1}^{K} \gamma_i [S(\mu_i, \lambda_i, y) - \mu_i W_2 - (\lambda_i - \mu_i) W_1]$, where $\sum_{i=1}^{K} \gamma_i = 1$, since (μ, λ), which corresponds to $\sum_{i=1}^{K} (\gamma_i \mu_i, \gamma_i \lambda_i)$, can be chosen arbitrarily. The latter maximization problem is then equivalent to (8). This procedure also makes clear that (μ, λ) solves the original problem $\max_{\mu,\lambda} \; \widehat{S}(\mu, \lambda, y) - \mu W_2 - (\lambda - \mu) W_1$ if and only if it is a convex combination of the maximizers in (8).

with probability $m(\lambda) - \phi(\mu, \lambda)$. The expected payoff of the firm therefore equals

$$\pi(\boldsymbol{w}, \mu, \lambda, y) = \phi(\mu, \lambda)[f(x_2, y) - w_2] + [m(\lambda) - \phi(\mu, \lambda)][f(x_1, y) - w_1]. \quad (10)$$

The expected payoff of applicants of type x_i is $V_i(\boldsymbol{w}, \mu, \lambda, y) = \psi_i(\mu, \lambda) w_i$, where, by a simple accounting identity, their matching probability $\psi_i(\mu, \lambda)$ equals

$$\psi_1(\mu, \lambda) = \frac{m(\lambda) - \phi(\mu, \lambda)}{\lambda - \mu} \qquad \text{or} \qquad \psi_2(\mu, \lambda) = \frac{\phi(\mu, \lambda)}{\mu}. \quad (11)$$

The special cases $\mu = 0$ and $\mu = \lambda$ are obtained by taking the corresponding limits, which yields $\psi_1(\lambda, \lambda) = \phi_\mu(\lambda, \lambda)$ and $\psi_2(0, \lambda) = \phi_\mu(0, \lambda)$.

Competitive Market for Queues. Consider a worker of type x_i. Define his *market utility* U_i as the maximum expected payoff that he can obtain in equilibrium. The optimality of workers' application decision implies that

$$\begin{cases} V_1(\boldsymbol{w}, \mu, \lambda, y) \leq U_1, \text{ with equality if } \lambda - \mu > 0, \\ V_2(\boldsymbol{w}, \mu, \lambda, y) \leq U_2, \text{ with equality if } \mu > 0. \end{cases} \quad (12)$$

As standard in the literature, we can use the market utility condition (12) to substitute the wages w_1 and w_2 out of (10) and rewrite the firm's problem with queue lengths as choice variables (see Lemma 9 and 10 in Appendix A.10 for a formal justification). This yields

$$\max_{0 \leq \mu \leq \lambda} m(\lambda) f(x_1, y) + \phi(\mu, \lambda)[f(x_2, y) - f(x_1, y)] - \lambda U_1 - \mu(U_2 - U_1). \quad (13)$$

Equation (13) has a straightforward interpretation: it is the payoff of a firm buying queues of low-type and high-type workers in a competitive market at prices equal to their respective market utilities.[18] This payoff is similar to equation (8) except that the costs that the firm faces are now workers' market utilities instead of their marginal contribution to surplus.

[18] Hence, the difference with a "conventional" competitive market is that the firm buys a distribution of applicants rather than directly hiring a particular type of worker. We have implicitly assumed that $0 < \mu < \lambda$ such that both market utility conditions hold with equality. However, it is easy to see that (13) also holds if $\mu = 0$ or $\mu = \lambda$.

Efficiency. The market equilibrium with wage menus thus coincides with the equilibrium in a competitive market where firms can buy queues directly at prices equal to workers' market utility. Hence, by the first welfare theorem, we obtain the following efficiency result.

Proposition 1. *The market equilibrium is constrained efficient, i.e., the equilibrium outcome solves the planner's problem given by* (5).

4 Sorting

In this section, we analyze under what conditions the planner's solution exhibits sorting. We focus on positive sorting, as the analysis of negative sorting is similar with reversal of the relevant inequalities. We show that in the limit case where $x_2 \to x_1$, the necessary and sufficient condition for sorting is that (the infimum of) the elasticity of complementarity of the production function is greater than (the supremum of) a new *quality-quantity elasticity*. Although it may appear counter-intuitive to think about screening and sorting when $x_2 \to x_1$, we show that the force against positive sorting is largest in this case when the above condition holds, making it sufficient for any given x_1 and x_2.

4.1 Definition of Sorting

Following Shimer and Smith (2000) and Shimer (2005), we define sorting as first-order stochastic dominance (FOSD) in firms' distributions of hires.[19] With two worker types, this definition can be expressed in terms of the probability that a firm hires a high-type worker, conditional on hiring someone,

$$h(\zeta, \lambda) \equiv \frac{\phi(\zeta\lambda, \lambda)}{m(\lambda)}, \tag{14}$$

where $\zeta \equiv \mu/\lambda$ is the fraction of high-type applicants in submarket (μ, λ).

[19] Strictly speaking, Shimer and Smith (2000) use a *weaker* notion of sorting, based on the bounds of the support of the distribution of hires; however, their definition is equivalent to FOSD of this distribution in their random-search environment. In contrast, Shimer (2005) proves a *stronger* sorting result (high-type workers are more likely to be employed in high- than in low-type jobs) for a special case ($f(x, y) = xy$ and urn-ball meetings); however, he acknowledges that the data demands to test this result "may be unrealistic" and suggests FOSD of the distribution of hires as a "more easily testable" alternative.

A subtlety in our environment is that firms of the same type may locate in two submarkets. Let $(\overline{\mu}(y), \overline{\lambda}(y))$ be the planner's solution to (5), and $\mathcal{Q}(y)$ be the set of queues that firms of type y face in that solution. As discussed before, if $\mathcal{Q}(y)$ contains a single element, it must be $\{(\overline{\mu}(y), \overline{\lambda}(y))\}$, otherwise $\mathcal{Q}(y)$ is of the form $\{(0, \lambda_0(y)), (\mu_1(y), \lambda_1(y))\}$, where subscript 0 and 1 represent the two submarkets and submarket 0 contains no high-type workers. The following definition of sorting accounts for either possibility.

Definition 2. *The planner's solution exhibits positive assortative matching (PAM) if $h(\zeta(y), \lambda(y))$ is (weakly) increasing in y for any selection $(\mu(y), \lambda(y)) \in \mathcal{Q}(y)$ where $\zeta(y) = \mu(y)/\lambda(y)$. Negative assortative matching (NAM) is defined similarly with $h(\zeta(y), \lambda(y))$ being (weakly) decreasing in y.*

Since firms with the same productivity may belong to multiple submarkets, this definition requires that the minimum conditional probability of hiring high-type workers among firms with a certain productivity is greater than the maximum conditional probabilities for firms with lower productivity. An implication of this definition is that when PAM holds in our environment, there exists at most one firm type y that is active in two submarkets. To see this, note that if $\mathcal{Q}(y)$ contains two elements, then Lemma 2 implies that $h(\zeta(y), \lambda(y))$ is 0 for one element and positive for the other. PAM then requires that $\mathcal{Q}(y')$ contains a single element with $\zeta(y') = 0$ for all $y' < y$ and with $\zeta(y') > 0$ for all $y' > y$, otherwise we can find a violation of the definition. Similar logic applies to NAM.

While the literature has traditionally restricted attention to sorting patterns in matches, our environment yields additional predictions. After all, given that firms may interview multiple applicants and subsequently select the most desirable one, there is a meaningful distinction between an application on the one hand and a match on the other hand. Hence, in addition to assortativeness of matches, we can also analyze the assortativeness of applications (or 'contacts'), i.e. whether the fraction of high-type applicants $\zeta(y)$ increases or decreases in y.

Definition 3. *The planner's solution exhibits positive assortative contacting (PAC) if $\zeta(y) = \mu(y)/\lambda(y)$ is (weakly) increasing in y for any selection $(\mu(y), \lambda(y)) \in \mathcal{Q}(y)$. Negative assortative contacting (NAC) is defined similarly with $\zeta(y)$ being (weakly) decreasing in y.*

As above, PAC/NAC requires that there exists at most one firm type which is active in two submarkets.

Illustration. Figure 1a illustrates a generic case where both PAC and PAM hold.[20] It shows how $\lambda(y)$ (right vertical axis), $\zeta(y)$, and $h(\zeta(y), \lambda(y))$ (left vertical axis) vary with y for a given distribution of agents. In this example, firms of type $y = 0.6$ are present in two submarkets: one with $\zeta = 0.2$ and $\lambda = 1.52$ and the other with $\zeta = 0$ and $\lambda = 1.95$. When $y \in (0, 0.6)$, $\zeta(y)$ and hence $h(\zeta(y), \lambda(y))$ are equal to zero, and $\lambda(y)$ is increasing. When $y \in (0.6, 1)$, $\zeta(y)$, $h(\zeta(y), \lambda(y))$ and $\lambda(y)$ are increasing. Note that if we adjust the distribution of agents such that $\underline{y} > 0.6$ but W_1 and W_2 remain unchanged, then all firm types have a unique queue.

Figure 1b illustrates a generic case where both PAC and PAM fail. Firms of type $y = 0.6$ are again present in two submarkets: one with $\zeta = 0.2$ and $\lambda = 1.48$ and the other with $\zeta = 0$ and $\lambda = 1.92$. PAC fails for two reasons: 1) when $y < 0.6$, $\zeta(y)$ is not monotonically increasing, and 2) at $y = 0.6$, the optimal $\zeta(y)$ jumps down. In contrast, $h(\zeta(y), \lambda(y))$ is strictly increasing when $y < 0.6$, yet PAM still fails because $h(\zeta(y), \lambda(y))$ jumps down at $y = 0.6$. Note that if we adjust the distribution of agents such that $\overline{y} < 0.6$ but W_1 and W_2 remain unchanged, then PAC fails whereas PAM holds at the planner's solution.

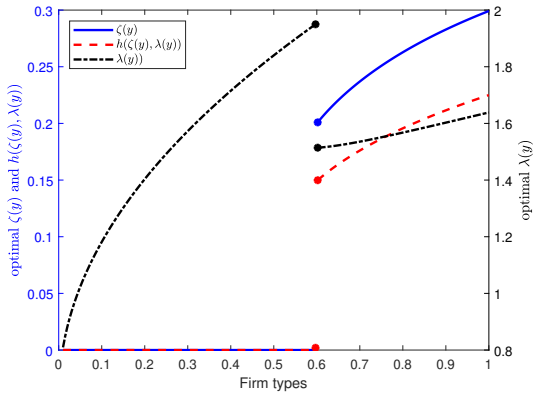
4.2 Quality vs Quantity

Tradeoff Between Quality and Quantity. We now heuristically discuss the tradeoff between quality and quantity faced by the planner. To simplify exposition, we consider the case where $f_{xy}(x, y) > 0$ (strict supermodularity).

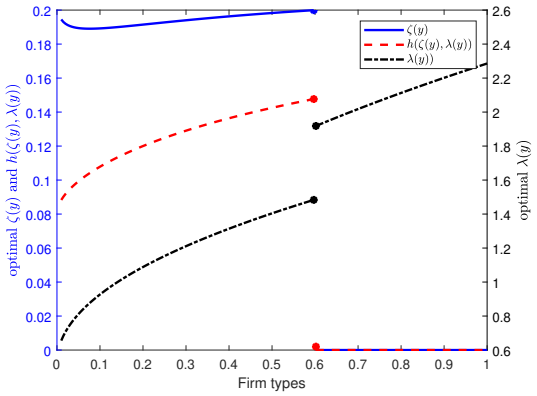
Since $\phi(\mu, \lambda)$ is strictly increasing in μ, the second term in (8) is strictly supermodular in (μ, y) for any given λ. Because of supermodularity, $\mu(y)$ is increasing in y for any given λ. This is the key feature in the model that promotes positive sorting, which we refer to as the desire for *match quality*.

At the same time, the second term in (8) is strictly submodular in (λ, y) for

[20] Figure 1a and 1b are generated as follows. We first set $x_1 = 1$, $x_2 = 3$, $y \in [0, 1]$, and ρ around 0.5 (either 0.485 or 0.515). Next, we create two submarkets for firms of type $y = 0.6$, one with $\zeta = 0$ and one with $\zeta = 0.6$. Given this information, we can compute the queue lengths in these two submarkets (see (39) and (40) in Appendix A.7) and hence W_1 and W_2. Finally, given W_1 and W_2, we can compute the optimal queues for other firm types and the distribution of worker types (L and z) that are consistent with W_1 and W_2 (which requires that the demand for both types of workers is positive).



(a) PAC/PAM holds: $\rho = 0.515$, $W_1 = 0.090237$, and $W_2 = 0.45016$



(b) PAC/PAM fails: $\rho = 0.485$, $W_1 = 0.092623$, and $W_2 = 0.46246$

Figure 1: Illustration of the planner's solution assuming the benchmark search technology ($\sigma = 0.4$) and a CES production function ($\alpha = 0.5$; ρ differs between the two subfigures). Furthermore, $x_1 = 1$, $x_2 = 3$, $\underline{y} = 0$, and $\overline{y} = 1$.

any given μ when $\sigma < 1$, since $\phi(\mu, \lambda)$ is strictly decreasing in λ. This feature also contributes to positive sorting, as it is a force for the optimal $\lambda(y)$ to be decreasing in y, and thus for $\zeta(y) = \mu(y)/\lambda(y)$ or $h(\zeta(y), \lambda(y))$ to increase in y. Intuitively, longer queues reduce the expected marginal contribution of high-type workers and this is a force that makes it relatively more attractive for high-y firms to go for good (quality) rather than for many applicants (quantity). The counterforce comes from the term $m(\lambda)f(x_1, y)$, which is strictly supermodular in (λ, y). This force tends to require the optimal $\lambda(y)$ to be increasing in y, and thus $\zeta(y)$ or $h(\zeta(y), \lambda(y))$ to decrease in y, because for high-type firms the opportunity costs of remaining unmatched are greater. We refer to this counterforce as the desire for *match quantity or match likelihood*.

First-Order Conditions. To make further progress, we now derive the first-order conditions (FOCs) of the planner's problem. Given that we are interested in how ζ and $h(\zeta, \lambda)$ vary with firm types, it simplifies exposition to rewrite (8) in terms of a choice of queue length λ and queue composition $\zeta = \mu/\lambda$, i.e.

$$\max_{\zeta,\lambda} \Pi(\zeta, \lambda, y) = m(\lambda) f^1 + \phi(\lambda\zeta, \lambda)\Delta f - \zeta\lambda W_2 - (1 - \zeta)\lambda W_1, \tag{15}$$

where $f^1 \equiv f(x_1, y)$ and $\Delta f \equiv f(x_2, y) - f(x_1, y)$ to reduce notation.

Consider first the choice of the queue length λ for a given $\zeta \in [0, 1]$. Since $\phi(\zeta\lambda, \lambda)$ is strictly concave in λ for all $\zeta > 0$ and $m(\lambda) = \phi(\lambda, \lambda)$, it follows that $\Pi(\zeta, \lambda, y)$ is *strictly* concave in λ for a given $\zeta \in [0, 1]$. Thus, if firms of type y are active in hiring, their optimal queue is unique and determined by the FOC

$$m'(\lambda) f^1 + \frac{\partial \phi(\zeta\lambda, \lambda)}{\partial \lambda} \Delta f = W_1 + \zeta(W_2 - W_1), \tag{16}$$

where $\partial \phi(\zeta\lambda, \lambda) / \partial \lambda = \zeta \phi_\mu(\zeta\lambda, \lambda) + \phi_\lambda(\zeta\lambda, \lambda)$. The first term on the left-hand side of (16) denotes the marginal contribution to surplus of a low-type applicant when all applicants are of a low type. The second term captures the fact that a fraction ζ of applicants actually has high productivity. The above condition concerns *quantity*: optimality of the queue length $\lambda(y)$ means that the marginal contribution to surplus of an extra worker in the queue is equalized across firms.

Note an optimal ζ for firm y can be either interior or at a corner. For example, in Figure 1a, the optimal $\zeta(y)$ equals 0 for $y < 0.6$ but is interior for $y > 0.6$. When an optimal ζ for firm y is interior ($0 < \zeta < 1$), it must satisfy the FOC

$$\phi_\mu(\zeta\lambda, \lambda) \Delta f = W_2 - W_1, \tag{17}$$

while the appropriate complementary slackness condition must hold if an optimal ζ is at a corner, i.e. $\zeta = 0$ or 1. Condition (17) concerns *quality*: optimality of the queue composition ζ requires that the marginal contribution to surplus from replacing a low-type worker in the queue by a high-type worker is equalized across firms. The left-hand side of (17) is exactly the difference between the marginal contribution to surplus of high-type and low-type workers, while the right-hand side is the difference in their cost. Intuitively, a larger ζ increases the firm's probability of matching with a high-type worker, but comes at a cost as these workers are more expensive.[21]

[21] The firm can increase ζ by $\Delta\zeta$ while keeping λ the same by increasing the queue length of high-type workers by $\lambda\Delta\zeta$ and decreasing the queue length of low-type workers by $\lambda\Delta\zeta$.

4.3 Quality-Quantity Elasticities

To analyze sorting, it is helpful to first consider the limit case in which $x_2 \to x_1 = x$. While it may appear counterintuitive to think about sorting and screening when worker heterogeneity vanishes, this case is particularly instructive for understanding the forces at play. Moreover, we will later show that the sorting condition for the limit case provides a sufficient condition for the general case.

As discussed, the planner may create two submarkets for certain firm types to reduce the extent to which low-type workers crowd out high-type workers. However, when x_2 is sufficiently close to x_1, the planner prioritizes matching probability over match quality, which implies a unique optimal queue for each firm type (see the proof of Proposition 2 for details). Further, as $x_2 \to x_1 = x$, the queue faced by firms of type y converges to a limit $(\zeta^*(y), \lambda^*(y))$, which is determined by the FOCs (16) and (17) evaluated at the limit.

We first characterize the boundary between PAC and NAC, where the optimal $\zeta^*(y)$ is constant across firm types, i.e $\zeta^*(y) = z \in (0,1)$, while the queue length $\lambda^*(y)$ may vary. Perturbing parameters away from this boundary can then be used to generate regions with positive or negative sorting.

Evaluating (16) at the limit reveals that $m'(\lambda^*(y))f(x,y)$ must be constant across firm types. This means that the elasticities of $m'(\lambda^*(y))$ and $f(x,y)$ with respect to y must exactly offset each other, i.e.

$$\frac{d\log f(x,y)}{d\log y} = -\frac{d\log m'\left(\lambda^*(y)\right)}{d\log \lambda^*(y)}\frac{d\log \lambda^*(y)}{d\log y}, \tag{18}$$

which requires that firms with higher productivity have longer queue lengths.

At the same time, it follows from (17) that for constant ζ to be optimal in the limit, the elasticity of $f_x(x,y)$ with respect to y must equal

$$\frac{d\log f_x(x,y)}{d\log y} = -\frac{\partial\log \phi_\mu(\zeta\lambda^*(y), \lambda^*(y))}{\partial\log \lambda^*(y)}\frac{d\log \lambda^*(y)}{d\log y}. \tag{19}$$

The right-hand side of this expression is positive. Intuitively, the longer queue at firms with higher productivity reduces the probability ϕ_μ that a high-type applicant creates surplus at those firms. This is a force against positive sorting. So, for constant ζ to be optimal, $f_x(x,y)$ must increase across firm types to offset this effect.

That is, the production function must exhibit complementarities. The required magnitude of these complementarities follows from combining (18) and (19), which yields

$$\rho(x,y) = \frac{\partial \log \phi_\mu(\zeta\lambda^*(y), \lambda^*(y))}{\partial \log m'(\lambda^*(y))}, \tag{20}$$

where $\rho(x,y)$ is the elasticity of complementarity defined by equation (1).

Contact Quality-Quantity Elasticity. We denote the elasticity at the right-hand side of (20) by $a^c(\zeta,\lambda)$ and refer to it as the *contact* quality-quantity elasticity, because it holds constant the fraction of high-type workers contacting (applying to) the firm. That is,

$$a^c(\zeta,\lambda) \equiv \frac{\partial \log \phi_\mu(\zeta\lambda,\lambda)}{\partial \log m'(\lambda)} > 0. \tag{21}$$

Recall that $\phi_\mu(\zeta\lambda,\lambda)$ represents the probability that a high-type applicant turns out to be the only high-type worker that the firm interviews and $m'(\lambda)$ describes the change in firms' matching probability. Thus the above elasticity measures the tradeoff between match quality and match likelihood when changing the queue length λ but keeping the queue composition ζ fixed. It is strictly positive because $m(\lambda)$ is strictly concave and $\phi_\mu(\zeta\lambda,\lambda)$ is strictly decreasing in λ. A large value means that the longer queue at firms with higher productivity results in a relatively large drop in the probability ϕ_μ that an extra high-type worker creates surplus, which is a force for negative sorting. To nevertheless obtain constant ζ, this force must be offset by the complementarities in production, as measured by $\rho(x,y)$.

As we will prove in Lemma 3, $a^c(\zeta,\lambda)$ is strictly increasing in ζ. Intuitively, when a larger fraction of the applicants is of high type, an increase in the queue leads to a more rapid decline in the probability that a high-type applicant creates surplus, creating a stronger force against positive sorting.

Match Quality-Quantity Elasticity. For PAM/NAM, the logic is similar, except that the boundary between the two cases is now the curve $h(\zeta,\lambda) = \overline{h}$, where all firms have the same conditional probability of hiring a high-type worker. This curve is downward sloping: as the queue length λ increases, the planner must reduce the fraction of high-type worker ζ to keep $h(\zeta,\lambda)$ constant. Analogous to the

above, this is optimal in the limit if $\rho(x,y) = a^m(\zeta(y),\lambda(y))$, where

$$a^m(\zeta,\lambda) \equiv \left.\frac{d\log\phi_\mu(\zeta\lambda,\lambda)}{d\log m'(\lambda)}\right|_{h(\zeta,\lambda)=\overline{h}} = a^c(\zeta,\lambda)\left(1-\frac{\partial\phi_\mu/\partial\zeta}{\partial\phi_\mu/\partial\lambda}\frac{\partial h/\partial\lambda}{\partial h/\partial\zeta}\right) > 0, \tag{22}$$

denotes the *match* quality-quantity elasticity, which holds constant the conditional probability that a firm matches with a high-type worker. The factor in parenthesis in (22) represents the relative effect of adjusting ζ so that $h(\zeta,\lambda)$ stays constant; as we will prove in Lemma 3, it is always between 0 and 1. Intuitively, as the queue length increases, the associated decrease in the fraction of high-type workers ζ mitigates the drop in ϕ_μ that high-type firms experience.

Summary. To summarize, as $x_2 \to x_1 = x$, the queue faced by firms of type y converges to a limit $(\zeta^*(y),\lambda^*(y))$. For the limit $\zeta^*(y)$ to be constant across firm types, the condition $\rho(x,y) = a^c(\zeta^*(y),\lambda^*(y))$ must hold for each y. Similarly, for $h(\zeta^*(y),\lambda^*(y))$ to be constant across firm types, the condition $\rho(x,y) = a^m(\zeta^*(y),\lambda^*(y))$ must hold for each y. Therefore, if $\rho(x,y) > a^i(\zeta^*(y),\lambda^*(y))$ for each y, then PAC (when $i=c$) and PAM (when $i=m$) hold in the limit allocation $(\zeta^*(y),\lambda^*(y))$ and, by continuity, whenever x_2 is sufficiently close to x_1.

The condition $\rho(x,y) \geq a^i(\zeta^*(y),\lambda^*(y))$ for positive sorting in the limit depends on the queues $(\zeta^*(y),\lambda^*(y))$, which generally are difficult to characterize explicitly. Clearly, a sufficient condition is that

$$\underline{\rho} \equiv \inf_{x,y}\ \rho(x,y) \geq \sup_{\zeta,\lambda}\ a^i(\zeta,\lambda) \equiv \overline{a}^i. \tag{23}$$

In fact, (23) guarantees positive sorting in the limit for any firm distribution $J(y)$, common worker type x, measure of workers L, and fraction of high-type workers z. To see that it is also necessary to guarantee positive sorting for any distribution of agents in the limit, consider the special case where firm heterogeneity is sufficiently small ($\underline{y}$ and $\overline{y}$ are close). At the planner's solution, queues faced by different firms are then approximately constant and the sorting condition for $x_2 \to x_1$ becomes $\rho(x,y) > a^m(\zeta,\lambda)$, where ζ and λ are population averages, i.e., $\zeta = z$ and $\lambda = L$. Therefore, the sufficient condition (23) is also necessary for PAC/PAM to always occur in the limit ($x_2 \to x_1$). The following proposition formalizes this idea. Of course, we are not particularly interested in the limit case where worker het-

erogeneity disappears. The advantage of the following proposition is that it only relies on minimal assumptions on the meeting technology, and provides a useful starting point for the general case with arbitrary x_1 and x_2, where, to prove that condition (23) is sufficient, we have to rely on the functional form of $\phi(\mu, \lambda)$.

Proposition 2. *Given a distribution of agents, if x_2 is sufficiently close to x_1, then at the planner's solution, firms of the same type must belong to the same submarket, i.e., the optimal queue faced by firms of the same type must be unique. Furthermore, as $x_2 \to x_1 = x$, the queue faced by firms of type y converges to a limit $(\zeta^*(y), \lambda^*(y))$.*

The necessary and sufficient condition to obtain PAC (resp. PAM) in the limit for any $J(y)$, x, L and z is that (23) holds for $i = c$ (resp. $i = m$). Similarly, the necessary and sufficient condition to obtain NAC (resp. NAM) in the limit for any $J(y)$, x, L and z is that for $i = c$ (resp. $i = m$), we have

$$\overline{\rho} \equiv \sup_{x,y} \rho(x, y) \leq \inf_{\zeta,\lambda} a^i(\zeta, \lambda) \equiv \underline{a}^i. \tag{24}$$

Proof. See Appendix A.5. □

Note that as $x_2 \to x_1$, i.e. the difference between worker types disappears, the queue composition converges to a limit $\zeta^*(y)$, which can be interior or at a corner. Although match quality is of second-order importance when $x_2 \to x_1$, conditional on the optimal queue length each firm has a unique optimal queue composition.

Inspecting the proof shows that the above proposition does not rely on the functional form of $\phi(\mu, \lambda)$; it only needs to satisfy regularity conditions A0 and a weaker version of A1 ($m(\lambda)$ is strictly concave). The following lemma establishes however that this functional form yields very simple expressions for $\underline{a}^i$ and $\overline{a}^i$.

Lemma 3. *If ϕ is given by (2), then i) $a^c(\zeta, \lambda)$ and $a^m(\zeta, \lambda)$ are strictly increasing in ζ; ii) $a^m(\zeta, \lambda) < a^c(\zeta, \lambda)$ when $\zeta \in (0, 1)$ and $\sigma > 0$; and iii)*

$$\overline{a}^c = \overline{a}^m = \frac{1+\sigma}{2} \quad \text{and} \quad \underline{a}^c = \underline{a}^m = \frac{1-\sigma}{2}. \tag{25}$$

Furthermore, i) $a^m(1/2, \lambda) = 1/2$ for any λ and σ, ii) $a^m(\zeta, \lambda)$ is strictly increasing in σ when $\zeta > 1/2$, iii) $a^m(\zeta, \lambda)$ is strictly decreasing in λ when $\sigma \in (0, 1)$ and

$\zeta > 1/2$. *The reverse comparative statics hold when* $\zeta < 1/2$. *Finally,* $a^c(\zeta, \lambda)$ *is strictly increasing in* σ *if and only if* $\lambda\zeta\sigma/(1+\lambda(1-\sigma)) > \sqrt{2(1-\zeta)} - 1$.

Proof. See Appendix A.6. □

By the above Lemma, the infimum (resp. supremum) of a^c and a^m can be reached or approached with $\zeta = 0$ (resp. $\zeta = 1$). Note that $a^m(\zeta, \lambda)$ reduces to $a^c(\zeta, \lambda)$ in those cases, i.e. $a^m(\zeta, \lambda) = a^c(\zeta, \lambda)$ when $\zeta = 0$ or $\zeta = 1$.[22] Hence, $\overline{a}^c = \overline{a}^m$ and $\underline{a}^c = \underline{a}^m$, which means that the conditions for PAC/NAC will coincide with those for PAM/NAM.

It is noteworthy that although the definition of $a^m(\zeta, \lambda)$ seems complicated, its explicit expression is simple and is given by equation (34) in Appendix A.6. It satisfies $a^m(\zeta, \lambda) = 1/2$ when $\sigma = 0$ while $a^m(\zeta, \lambda) = \zeta$ when $\sigma = 1$.

Together with Proposition 2, Lemma 3 implies that $\underline{\rho} \geq (1+\sigma)/2$ is necessary for PAC/PAM to hold for any distribution of agents' types. Similarly, $\overline{\rho} \leq (1-\sigma)/2$ is necessary for NAC/NAM. We are of course not particularly interested in the sorting condition for the limit $x_2 \to x_1$. When x_1 and x_2 can take any value, deriving queue lengths across firm types is more complicated. However, below we show that (23) and (24) are sufficient for sorting for any x_1 and x_2.

4.4 Sorting Condition for any Distribution

We now consider the general case where x_1 and x_2 can take any value. Assume that condition (23) holds. We show that the planner's solution always exhibits PAC/PAM, since the degree of complementarity required for positive sorting is larger when worker heterogeneity is smaller. Intuitively, when x_1 and x_2 are close, firms do not care much which type they hire and match likelihood is much more important than match quality. When x_1 and x_2 are far apart and hence match quality is important, high-productivity firms are willing to substitute match likelihood for match quality because of production complementarities and negative externalities in the meeting process. That is, high-y firms would reduce their queue length (by offering low-type workers a worse deal) relative to the case where x_1 and x_2 are close to each other. This is a force towards positive sorting. Hence, it is perhaps not surprising that if for any x_1 (and $(L, z, J(y))$, positive sorting always holds when

[22] To see this, note that $\phi(0, \lambda) = 0$ and $\phi(\lambda, \lambda) = m(\lambda)$ for any λ. Both imply $\partial h/\partial\lambda = 0$.

$x_2 \to x_1$, then it also holds for any x_1 and x_2. The following proposition formally establishes this result.

Proposition 3. *Assume that ϕ is given by (2) with $\sigma > 0$. The planner's solution then exhibits PAC/PAM (resp. NAC/NAM) for any distribution of agents' types if and only if $\underline{\rho} \geq (1+\sigma)/2$ (resp. $\overline{\rho} \leq (1-\sigma)/2$). Furthermore, the planner's solution is unique if $\underline{\rho} \geq (1+\sigma)/2$ or $\overline{\rho} \leq (1-\sigma)/2$.*

Proof. See Appendix A.7. □

The above proposition establishes that the necessary condition identified in Proposition 2 is also sufficient. Unlike Proposition 2, its proof uses i) the functional form of $\phi(\mu, \lambda)$ in equation (2) and ii) that there are only two worker types. However, generalization is possible. For example, the proof does not actually require the particular functional form of $\phi(\mu, \lambda)$ in equation (2), but simply that $\phi(\mu, \lambda)$ satisfies certain properties. Verification of those conditions is possible for other specifications, although not always analytically.[23] Further, we assume two types of workers only to simplify the analysis for the case where the planner finds it optimal to open multiple submarkets for certain types of firms. If at the optimum firms of the same type are in the same submarket, our sufficiency condition extends to the case of an arbitrary number of worker types (see Appendix A.11 for an example).

To derive the necessary condition for sorting in Proposition 2, we used the limit case $x_2 \to x_1$ (for which sorting has the least benefits), where firms of the same type must be in the same submarket, and hence $(\mu(y), \lambda(y))$ is continuous in y. Because of this continuity, there must exist firm types that attract both types of workers in equilibrium, i.e., $0 < \mu(y) < \lambda(y)$. In this case, PAC (resp. PAM) holds if and only if $\zeta'(y) \geq 0$ (resp. $\frac{dh(\zeta(y),\lambda(y))}{dy} \geq 0$) whenever $\zeta(y)$ is interior. However, to show the sufficiency of the necessary condition, we need to consider arbitrary distributions of agents (in particular, arbitrary x_1 and x_2). One can construct examples where $\sigma > 0$ and all firms have a corner solution under the necessary and sufficient condition for PAC/PAM (or for NAC/NAM), in which case firms with types greater than some threshold all attract only high-type workers and firms with types smaller than the

[23] For example, we can show this numerically for the case in which the number of applications follows a Poisson distribution (urn-ball) as common in the literature (see Wolthoff, 2018).

threshold all attract only low-type workers (thus PAC/PAM holds). Figure 1a plots a generic case where PAC/PAM holds at the planner's solution. Note that if firms of some type y_m have two submarkets (for example, $y_m = 0.6$ in both Figure 1a and 1b), then all firms of types below y_m have a corner solution and attract only low-type workers. Of course, it can be the case that all firms of the same type are in the same submarket and y_m does not exist. Here we do not derive conditions for which those different scenarios arise; we just show that PAC/PAM (and similar NAC/NAM) holds for all those scenarios.

Given Definition 1, we can alternatively state Proposition 3 as follows.

Corollary 1. *If ϕ is given by (2) with $\sigma > 0$, the planner's solution exhibits PAC/PAM (resp. NAC/NAM) for any distribution of agents' types if and only if $f(x, y)$ is $2/(1-\sigma)$-root-supermodular (resp. $2/(1+\sigma)$-root-submodular).*

As mentioned, some firms may have multiple optimal queues in the planner solution, because the maximization problem (15) is nonconcave. A standard approach to analyze sorting in such a case, which at first may look simpler, would be to use (16) to obtain the optimal λ as a function of ζ and y, denoted by $\lambda^o(\zeta, y)$. Plugging it into (15) gives firms' expected profit as a function of ζ and y only: $\widetilde{\Pi}(\zeta, y) = \Pi(\zeta, \lambda^o(\zeta, y), y)$. PAC would then hold if $\widetilde{\Pi}(\zeta, y)$ is strictly supermodular in (ζ, y). This approach fails however because $\widetilde{\Pi}(\zeta, y)$ is not strictly supermodular in our model. Our proof of Proposition 3 circumvents this issue by separately considering firm types for which a unique optimal queue exists and firm types for which multiple optimal queues exist, taking into account that in the latter case one solution must be $\zeta = 0$.

4.5 Effect of Screening

We can now consider how screening affects sorting. It follows from Proposition 3 and Corollary 1 that the magnitude of the production complementarities required to obtain PAC/PAM for any distribution of agents is *increasing* in the degree of screening. In particular, when $\sigma \to 0$ and meetings are bilateral, PAC/PAM requires square-root supermodularity, in line with Eeckhout and Kircher (2010).[24]

[24] When $\sigma = 0$, we obtain $a^c(\zeta, \lambda) = a^m(\zeta, \lambda) = m'(\lambda)(\lambda m'(\lambda) - m(\lambda))/(\lambda m(\lambda) m''(\lambda)$, which is independent of ζ and which is precisely the elasticity of substitution of the total number of matches that Eeckhout and Kircher (2010) show to be important for bilateral technologies.

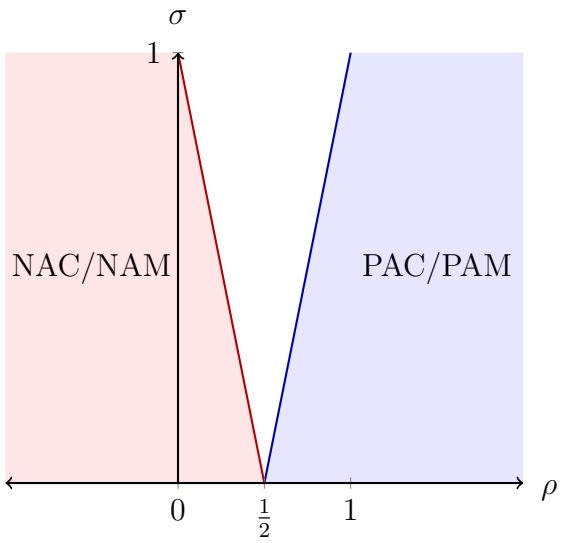


Figure 2: Combinations of ρ and σ that give rise to PAC/PAM (blue) or NAC/NAM (red) for any distribution of agents' types, assuming a CES production function.

At the other extreme, log-supermodularity is required for PAC/PAM when $\sigma = 1$ and firms can interview all their applicants. In contrast, an increase in the expected number of interviews raises the degree of substitutability required for NAC/NAM from square-root-submodularity if $\sigma = 0$ to submodularity when $\sigma = 1$. Figure 2 illustrates these results.

Intuition. To understand how screening affects sorting, consider first the special case $\sigma = 0$ as in Eeckhout and Kircher (2010). Since the crowding-out effect is strong, the planner will assign to a firm either a long queue with only low-skill workers or a short queue with only high-skill workers. The former option reduces the probability of being unmatched, whereas the latter makes it possible to take advantage of production complementarity. Positive sorting arises when complementarity dominates this complementarity-insurance tradeoff.

For general σ, an optimal queue can contain both types of workers, and the above tradeoff is more complicated. To see how a longer queue affects the complementarity channel, note that there are two distinct cases in which a high-type applicant *fails* to create surplus: 1) he is not interviewed, or 2) he is interviewed, but at least one other high-type applicant is interviewed as well. Each of these two

cases becomes more likely as the queue length increases, i.e. ϕ_μ is decreasing in λ, which is a force against sorting.

However, the exact impact of an increase in the queue length depends on whether primarily low types or high types are being added (as measured by ζ) as well as whether types can easily be distinguished (as measured by σ). After all, when the queue mainly consists of low types (ζ is low), multiple interviews with high types are unlikely and the effect of a longer queue predominantly operates by making it less likely for a high-type applicant to be interviewed. Clearly, this force is *mitigated* by an increase in firms' screening ability σ: When σ is high, a high-type applicant is likely to be interviewed regardless of whether there are many or few other applicants.

In contrast, when the queue mainly consists of high types (ζ is high), multiple interviews with high-type applicants are a key concern. A longer queue makes this outcome more likely and this force is *amplified* by an increase in firms' screening ability σ, since it increases every applicant's interviewing probability.

Sorting with Few High-Type Workers. Since Proposition 3 derives a sorting condition for any distribution of agents' types, the required degree of complementarity in production for positive sorting is increasing in σ. The above logic also suggests that for a given distribution of agents with relatively few high-type workers, the required degree of complementarity in production is not necessarily increasing in σ, since the first case above is the relevant one. We offer some results on this issue in Proposition 4. To simplify the exposition, we assume that the production function is CES with $\rho(x, y) = \rho < 1$.

Recall that for any given distribution of agents, when $\sigma = 0$, as long as $\rho > 1/2$ we have PAC/PAM at the planner' solution (when $\rho < 1/2$, we have NAC/NAM, and when $\rho = 1/2$, the results are indeterminate: the planner's solutions are not unique; they can exhibit PAC/PAM or NAC/NAM or no sorting.)

Next, suppose $\sigma = 1$. Our next result shows that for any positive ρ (no matter how small it is), if high-skilled workers are sufficiently scarce, then the planner's solution exhibits PAC/PAM (while the planner imposes NAC/NAM for $\rho \leq 0$).[25]

[25]Its proof also shows how the analysis becomes much simpler when $\sigma = 1$ and the planner's problem is strictly concave (such that the optimal queue is unique for each firm type y).

Proposition 4. *Assume that $\sigma = 1$ and the production function is CES with $\rho \in (0,1)$. For any given firm type distribution $J(y)$, worker skills x_1 and x_2, and aggregate worker-firm ratio L, there exist two thresholds $\overline{z}^c$ and $\overline{z}^m$ with $0 < \overline{z}^c < \overline{z}^m$ such that PAC (resp. PAM) holds at the planner' solution if and only if the fraction of high-type workers $z \leq \overline{z}^c$ (resp. $z \leq \overline{z}^m$). Furthermore, $\zeta(y) < \rho$ for each y in both cases.*

Proof. See Appendix A.8. □

Therefore, by setting $\rho < 1/2$, the required degree of supermodularity can decrease when σ changes from 0 to 1 for a given distribution of agents when the fraction of high-type workers is sufficiently small. The intuition for this result is similar as before. The force against positive sorting is measured by $a^c(\zeta, \lambda)$ and $a^m(\zeta, \lambda)$, which are both increasing in ζ. Hence, when the fraction of high-type workers is small, the force against positive sorting is also small. Furthermore, when $\zeta < \rho < 1/2$, both $a^c(\zeta, \lambda)$ and $a^m(\zeta, \lambda)$ are decreasing in σ.[26] Thus, in this case the force against positive sorting is weaker at $\sigma = 1$ relative to $\sigma = 0$. Note that if $\sigma = 1$ and $z \to 0$, then the sufficient condition for positive sorting becomes $\rho > 0$ (strict supermodularity), which is the same as in Becker (1973) despite the presence of search frictions. Intuitively, when $\zeta \to 0$, then $a^c(\zeta, \lambda), a^m(\zeta, \lambda) \to 0$, i.e. the effect of a longer queue on the probability that a high-skill worker increases surplus becomes negligible.

5 Empirical Illustration: Ban the Box

In this section, we illustrate the empirical relevance of our model by applying it to the case of "Ban the Box" (BTB) policies.

Background. In recent years, 36 US states and over 150 US cities and counties (jointly having a population of more than 260 million people) have introduced BTB laws and policies, which aim to reduce barriers to employment for people with criminal records (Avery and Lu, 2021). In particular, these policies prohibit employers from asking about criminal history early in the recruiting process (e.g., through a checkbox on the application form), although background checks may still

[26] By Lemma 3, $a^m(\zeta, \lambda)$ is strictly decreasing in σ when $\zeta < 1/2$, and for any given λ, $a^c(\zeta, \lambda)$ is strictly decreasing in σ when ζ is small enough (more precisely, when $\lambda\zeta < \sqrt{2(1-\zeta)} - 1$).

be conducted before making a job offer. The idea is that it may help ex-offenders find better jobs if their criminal history is revealed later in the process.

Given the high incidence of criminal records among unemployed men in the US, the potential of BTB to affect labour market outcomes is substantial.[27] Estimating these effects has been the aim of a sizeable empirical literature, including the audit study by Agan and Starr (2018) which analyzes how the policies affect callback rates of (fictious) applicants. Despite their strengths, audit studies are limited in that they cannot speak to how BTB affects either the conversion of callbacks (interviews) into matches or workers' application strategies. Such questions benefit from a structural model. To illustrate how our framework might contribute, we demonstrate how it captures the key features of BTB and we provide a simple calibration to show that the policies may push former convicts towards *worse* jobs.[28]

Mapping to the Model. Our model provides a natural framework to analyze policies like BTB. Low-type and high-type workers can be interpreted as those with and without criminal records, respectively.[29] The assumption of type-dependent wages is consistent with the fact that neither the Civil Rights Act (Equal Employment Opportunity Commission, 2012) nor BTB policies (Avery and Lu, 2021) prohibit employers from conditioning wages on criminal history. Similarly, the assumption that an interview reveals an applicant's true type is consistent with the fact that employers can do background checks before hiring.

BTB may cause applicants without criminal records to lose interview opportunities due to the presence of applicants with criminal records. This effect resembles a decrease in σ in our model but is not exactly the same, since the policy does not affect the number of interviews that firms can do but rather makes it harder to distinguish between different types of applicants. To capture this idea, we extend our model by allowing firms to costlessly observe a binary signal for every appli-

[27] Bushway et al. (2022) estimate for the US that 53% of the unemployed men aged 30 to 38 have been arrested, around 40% have been convicted, and slightly more than 20% have been incarcerated (see their Table 1).

[28] Of course, our model abstracts from various aspects that might be important for a full analysis. For example, Agan and Starr (2018) present evidence that the policy causes employers to statistically discriminate against applicants with distinctly black names. Although this effect is certainly important, omitting race allows us to emphasize a new unintended consequence of BTB which is present even if statistical discrimination is not feasible.

[29] A positive recidivism rate could be one reason for the productivity difference.

cant. For applicants without criminal records, the signal is positive with certainty. In contrast, a former convict generates a negative signal with probability $\tau \in [0,1]$. Hence, the signal is perfect if $\tau = 1$, but pure noise if $\tau = 0$. BTB then naturally corresponds to a reduction in τ.

In this extended environment, firms first interview applicants with positive signals and proceed with those with negative signals only if interview capacity remains. The following proposition establishes that the environment with signals is isomorphic to our baseline model, as long as we transform the parameter σ.

Proposition 5. *In our environment with signals, consider a firm with queues* (μ, λ). *Let* $\widehat{\sigma} = 1 - (1-\tau)(1-\sigma) \in [0,1]$, *then the probability that the firm interviews at least one high-type worker equals*

$$\phi(\mu, \lambda) = \frac{\mu}{1 + \widehat{\sigma}\mu + (1-\widehat{\sigma})\lambda}.$$

Proof. See Appendix A.9. □

As a corollary, all our earlier results continue to hold, except that they apply to $\widehat{\sigma}$ instead of σ to account for the fact that the signal precision τ is a substitute for the screening intensity σ.

Empirical Predictions. As mentioned, BTB reduces the signal precision τ and thus $\widehat{\sigma}$. Our theoretical results then imply that the policy may backfire: it may help induce sorting and therefore push workers with criminal records towards *worse* jobs. This is particularly true when the fraction of workers with criminal records is relatively small. The logic is the same as before: because the policy makes it harder to distinguish between different types of applicants during the recruiting process, high-type firms may decide to discourage workers with criminal records from applying by reducing their compensation or, in the most extreme case, committing to not hire them (i.e., posting a zero wage).

To demonstrate this effect, we conduct a simple quantitative exercise in which we calculate the equilibrium for reasonable parameter values that are roughly in line with the US labor market. We show that for these parameter values BTB indeed pushes workers with criminal records towards worse jobs. The exact parameter values are irrelevant for the qualitative effect.

For simplicity, we assume that there are only two types of firms with productivities $y_L < y_H$. For the production function, we use $f(x, y) = xy$ such that PAC/PAM is obtained for any value of $\widehat{\sigma}$. The multiplicative structure allows us to normalize both $x_2 = 1$ and $y_H = 1$. We define low-type workers as workers who have been convicted, and set their fraction equal to 0.4. This number is consistent with Bushway et al. (2022) who, using the 1997 National Longitudinal Survey of Youth, estimate that 40% of the unemployed white males between age 30 and 38 have been convicted, while the corresponding number is 38% for black males. Finally, we set $L = 1$, which is roughly the number of unemployed persons per job opening in the JOLTS data for 2017 and 2018, which is the time frame of the income data in Finlay and Mueller Smith (2021) and the BLS data on employment and wages by occupation that we use below.

We have no direct information on x_1, y_L and the fraction of high-type jobs. We set them at values that are consistent with three data moments. Specifically, the first moment is a wage ratio of 80% between employed low-type and high-type workers, based on Finlay and Mueller Smith (2021); in our model, this moment is driven primarily by the value of x_1.[30] For the second and third moment, we equate the low-type jobs in our model with occupations that employ a larger number of convicted workers. Based on BLS data for May 2018, we then target an employment share of 38% for these occupations and a wage ratio of 0.58 for workers employed in these jobs relative to other workers; in the model these moments are primarily driven by the fraction of high-type jobs and the value of y_L, respectively.[31] When we match those targets, we set $\widehat{\sigma} = 1$ capturing the idea that before the introduction of BTB, employers could distinguish between the different types of applicants at

[30] Finlay and Mueller Smith (2021) compare convicted felons with a control group of similar workers in the same age group who have not more than a high school degree. For 2018, they report yearly employment probabilities of 50% for the convicted workers and 60% for the control group (see Panel A of their Figure 1). They also find an income ratio of 65% ($13K for the convicted workers and $20K for the control group; see Panel B of their Figure 1). Accounting for the difference in employment probabilities, this corresponds to a wage ratio of 80%. The advantage of the ratio is that we do not have to take a stance on the period length.

[31] Among the large occupations, we define the y_L occupation as consisting of "Office and Administrative Support", "Food Preparation and Serving Related", "Transportation and Material Moving", "Production", which have employment shares of 15.1%, 9.2%, 7.1%, and 6.3% and mean hourly wage equal to $18.75, $12.30, $18.41, and $18.48, respectively ("Occupational Employment and Wage Statistics", BLS, May 2018), see Carson et al. (2021) for evidence on occupations that employ convicted workers. The sum of employment shares is 37.7%, and the employment weighted average wage is $17.12. Note that the national average wage is $24.98.

essentially zero cost. The resulting parametrization is summarized in Table 1.[32]

parameter	value	source
x_2	1	normalization
y_H	1	normalization
% x_2 workers	0.6	% unemployed with conviction record (Bushway et al., 2022)
L	1.0	JOLTS (2017 and 2018)
x_1	0.6	Finlay and Mueller Smith (2021), BLS
y_L	0.4	idem
% y_H jobs	0.5	idem

Table 1: Parametrization of the model

To remain agnostic about the value of $\widehat{\sigma}$ after the introduction of BTB, we plot outcomes for all $\widehat{\sigma} \in [0, 1]$. Figure 3 shows that a naive analysis with an exclusive focus on matching probabilities would conclude that the policy indeed helps convicted workers: their matching probability increases substantially from 33% to 55% as $\widehat{\sigma}$ is reduced from 1 to 0 (red dashed line). This reduction naturally comes at the expense of the workers without conviction whose matching probability decreases (blue dashed line).

However, Figure 3 also shows that the increase in the matching probability of convicted workers is the result of increased sorting: reducing $\widehat{\sigma}$ from 1 to 0 leads to fewer convicted (x_1) workers applying to (and accordingly being matched at) high-type jobs and more at low-type jobs (red solid line), while the reverse holds for workers without prior conviction (blue solid line). Basically, the forced reduction in ex-post screening by a BTB policy makes firms screen more ex ante: the good jobs discourage the convicted workers from applying by offering them lower wages. As a result, convicted workers apply more often to (and match with) the less-productive jobs. Note that when $\widehat{\sigma}$ is smaller than 0.5, all convicted workers apply to low-type jobs and all workers without conviction apply to high-type jobs, so that the market equilibrium features complete separation; further reduction of $\widehat{\sigma}$ has no effects on the equilibrium.[33]

[32] The parameter values that match the targets precisely are $x_1 = 0.5597$, $y_L = 0.4416$, and the fraction of y_H jobs being 0.5128. Since our targets are rough measures, we round these values.

[33] For the given parameter values, the high-type firms always stay in one submarket for any $\widehat{\sigma} \in [0, 1]$, and the same is true for the low-type firms. When $\widehat{\sigma} \leq 0.5$, the high-type firm submarket contains only workers without conviction. When $\widehat{\sigma} \leq 0.6$, then low-type firms attract only convicted workers.

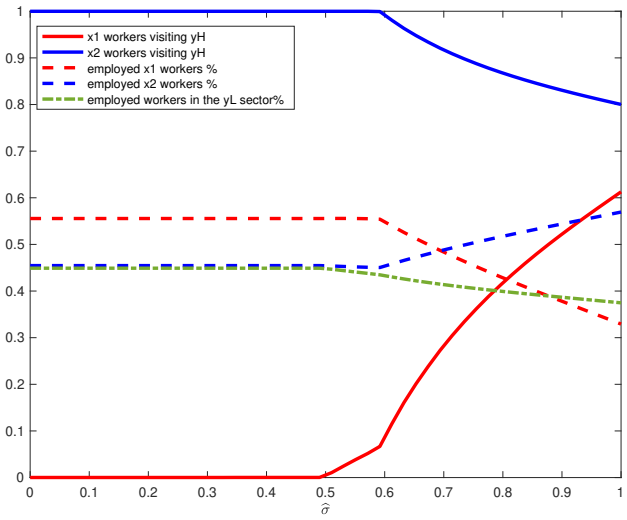


Figure 3: The effect of Ban the Box on employment and sorting

Convicted workers are hurt by the lower quality of their jobs through the wages that they earn. Figure 4 shows that wages of ex-offenders fall after BTB since they work more often in the low-productivity sector (red solid line). We see that workers without conviction experience a small wage increase (from 0.36 to 0.45) because they are more likely to be employed in the more productive sector (blue dashed line).[34]

In order to quantify the total effect of the policy on workers, we consider their expected payoffs, which is the product of their matching probability and wage given by Figure 3 and 4, respectively. We find that the market utility for convicted workers falls by 26% as we reduce σ from 1 to 0, while market utility for workers without conviction *increases* by 1.3%. So, ironically, BTB hurts convicted workers and can potentially benefit the other workers.[35] Total surplus drops monotonically

[34] Rose (2021) gives direct evidence that a 2013 BTB law passed in Seattle had a negligible impact on ex-offenders' employment and that the ex-offenders shifted away from retail and into food services. He reports a modest effect on the average wages of ex-offenders. This could mean that BTB had a small effect on σ. Alternatively, if σ was close to 0.5 before BTB was introduced, lowering it to 0 has very little impact on wages.

[35] The effect of lowering $\widehat{\sigma}$ on the expected payoff of workers without conviction depends on the parameter values that we use. For example, for $y_L = 0.5$, as we reduce $\widehat{\sigma}$ from 1 to 0, the expected payoff of workers without conviction record decreases by 5.7%, whereas convicted workers experience a decrease of 18% in their expected payoff.

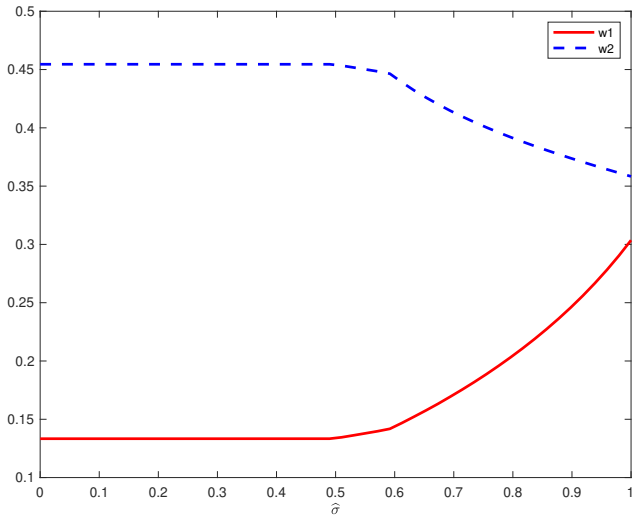


Figure 4: The effect of Ban the Box on wages

by 2.3% (from 0.334 to 0.326) as we lower $\widehat{\sigma}$ from 1 to 0.

Summarizing, our model suggests that policies like BTB may not achieve their objective because discouraging firms from screening workers through interviews may cause firms to induce more self-selection. These findings operate through workers' application decisions. As a result, they cannot readily be identified by an audit study in which applications are sent randomly, which highlights the importance of a micro-founded search-theoretic model for analyzing policies like BTB.

6 Conclusion

A firm with a vacancy typically has multiple instruments to screen applicants. By announcing the terms of trade ex ante, it can discourage certain workers from applying, while ex post—after receiving applications—it can interview applicants in an attempt to identify the most profitable hire. In this paper, we show how these instruments jointly determine equilibrium outcomes, including sorting patterns.

Perhaps surprisingly, we find that if firms can interview more applicants, sorting may be harder in the sense that stronger complementarities in production are necessary to get positive assortative matching. The more workers a firm can screen, the stronger the incentives for high-type workers are to avoid ending up in the same pool of applicants and this is a force against sorting which is by itself efficient (a

planner also wants to reduce the probability that resources are wasted because they end up in the same pool). We show that our results have important implications for policies that change the amount of information that is available during the recruiting process, like Ban the Box.

For simplicity, we have treated screening capacity as exogenous. We endogenize it at a linear cost in Appendix A.12. Interestingly, production complementarity must be even stronger for positive sorting to arise in this case, because firms with intermediate types usually have the strongest incentives to invest in (ex-post) screening: the most-productive firms find it optimal to attract high-type applicants only and avoid the screening costs. Similarly, the least-productive firms will choose to attract low-type applicants. Firms in the middle of the distribution would also like to take advantage of the production complementarity, but they cannot afford attracting long queues of high-type workers; they therefore pay the screening costs and attract both types of workers. In Appendix A.11, we further generalize our results to an arbitrary number of worker types for invariant meeting technologies (which include urn-ball and geometric; see Lester et al., 2015; Cai et al., 2017).

There are several promising avenues for future research. In many markets, workers have incentives to send multiple applications simultaneously. In recent work, Birinci et al. (2023) demonstrate how an increase in the number of simultaneous applications interacts with firms' interviewing decisions and can explain aggregate trends in job finding and separation rates. Extending our framework along these dimensions makes it possible to explore implications for sorting. An increase in the number of simultaneous applications reduce the cost for high-type workers to end up in the same queue as other high-type workers. However, even then, high-type workers have incentives to diversify and not only apply to the top places.

Further, in recessions, when firms are flooded with applicants, firms may shift their hiring strategy more towards ex-ante sorting by discouraging certain types while in booms, when workers are scarce, firms may encourage a wider variety of applicants and screen more ex-post. This would lead to higher unemployment and more sorting (less mismatch) in recessions. Baley et al. (2022) and Crane et al. (2022) give evidence that mismatch is counter cyclical.

On the empirical side, an important implication of our model is that sorting patterns are driven both by the production function and the meeting process. To

identify complementarities in production, we may therefore need—besides data on matches—additional information on the entire pool of applicants. This way, we can first identify the parameters of the meeting technology (i.e. how many workers of each type and how many are screened) and then, conditional on the meeting technology, matching patterns are informative on production complementarities.

Finally, our framework can be used for more in-depth analysis of policies that affect the amount of information that is available during the recruiting process. For example, Agan and Starr (2018) show how Ban the Box has lead to discrimination against minorities in callbacks. Enriching our model to account for minority-status would make it possible to analyze to what extent such discrimination translates into differences in hiring rates and wages. Alternatively, an extended version of our framework could also contribute to the literature on the question of whether firms should be prohibited from screening workers based on the duration of their unemployment spell (see, for example, Jarosch and Pilossoph, 2018).

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Agan, A. and Starr, S. (2018). Ban the Box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, 133(1):191–235.

Albrecht, J. W., Cai, X., Gautier, P. A., and Vroman, S. B. (2019). Multiple applications, competing mechanisms, and market power. mimeo.

Albrecht, J. W., Gautier, P. A., and Vroman, S. B. (2006). Equilibrium directed search with multiple applications. *Review of Economic Studies*, 73(4):869–891.

Auster, S., Gottardi, P., and Wolthoff, R. (2021). Simultaneous search and adverse selection. mimeo.

Avery, B. and Lu, H. (2021). Ban the Box. U.S. Cities, Counties, and States Adopt Fair-Chance Policies to Advance Employment Opportunities for People with Past Convictions.

Bagger, J. and Lentz, R. (2018). An empirical model of wage dispersion with sorting. *Review of Economic Studies*, forthcoming.

Baley, I., Figueiredo, A., and Ulbricht, R. (2022). Mismatch cycles. *Journal of Political Economy*, 130(11):2943–2984.

Banfi, S., Choi, S., and Villena-Roldán, B. (2020). Sorting on-line and on-time. mimeo.

Bartolucci, C., Devicienti, F., and Monzón, I. (2018). Identifying sorting in practice. *American Economic Journal: Applied Economics*, 10(4):408–438.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846.

Birinci, S., See, K., and Wee, S. L. (2023). Job applications and labor market flows. *Review of Economic Studies*, forthcoming.

Board, S., Meyer-ter Vehn, M., and Sadzik, T. (2019). Recruiting talent. mimeo.

Borovičková, K. and Shimer, R. (2020). High wage workers work for high wage firms. mimeo.

Bushway, S., Cabreros, I., Welburn Paige, J., Schwam, D., and Wenger, J. B. (2022). Barred from employment: More than half of unemployed men in their 30s had a criminal history of arrest. *Science Advances*, 8(7).

Cai, X., Gautier, P., and Wolthoff, R. (2017). Search frictions, competing mechanisms and optimal market segmentation. *Journal of Economic Theory*, 169:453–473.

Cai, X., Gautier, P., and Wolthoff, R. (2022). Meetings and mechanisms. *International Economic Review*, forthcoming.

Card, D., Heining, J., and Kline, P. (2013). Workplace heterogeneity and the rise of West German wage inequality. *The Quarterly Journal of Economics*, 128(3):967–1015.

Carson, E. A., Bhaskar, R., and Leticia Fernandez, S. P. (2021). Employment of Persons Released from Federal Prison in 2010. Technical report, Bureau of Justice Statistics.

Costinot, A. (2009). An elementary theory of comparative advantage. *Econometrica*, 77(4):1165–1192.

Crane, L. D., Hyatt, H. R., and Murray, S. M. (2022). Cyclical labor market sorting. *Journal of Econometrics*.

Davis, S. J. and Samaniego de la Parra, B. (2017). Application flows. mimeo.

Eeckhout, J. and Kircher, P. (2010). Sorting and decentralized price competition. *Econometrica*, 78:539–574.

Eeckhout, J. and Kircher, P. (2011). Identifying sorting - in theory. *Review of Economic Studies*, 78(3):872–906.

Equal Employment Opportunity Commission (2012). Enforcement Guidance on the Consideration of Arrest and Conviction Records in Employment Decisions under Title VII of the Civil Rights Act.

Finlay, C. and Mueller Smith, M. (2021). Justice-involved Individuals in the Labor Market since the Great Recession. *ADEP Working paper, US Census*.

Galenianos, M. and Kircher, P. (2009). Directed search with multiple job applications. *Journal of Economic Theory*, 114:445–471.

Gautier, P. A. and Teulings, C. N. (2006). How large are search frictions? *Journal of the European Economic Association*, 4(6):1193–1225.

Gautier, P. A. and Teulings, C. N. (2015). Sorting and the output loss due to search frictions. *Journal of the European Economic Association*, 13(6):1136–1166.

Hagedorn, M., Law, T. H., and Manovskii, I. (2017). Identifying equilibrium models of labor market sorting. *Econometrica*, 85(1):29–65.

Hicks, J. (1932). *The Theory of Wages*. Macmillan, London.

Hicks, J. (1970). Elasticity of substitution again: Substitutes and complements. *Oxford Economic Papers*, 22(3):289–296.

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *Quarterly Journal of Economics*, 133(2):765–800.

Håkanson, C., Lindqvist, E., and Vlachos, J. (2018). Firms and skills: The evolution of worker sorting. mimeo.

Jarosch, G. and Pilossoph, L. (2018). Statistical Discrimination and Duration Dependence in the Job Finding Rate. *The Review of Economic Studies*, 86(4):1631–1665.

Kircher, P. (2009). Efficiency of simultaneous search. *Journal of Political Economy*, 117:861–913.

Kurlat, P. (2016). Asset markets with heterogeneous information. *Econometrica*, 84(1):33–85.

Lester, B., Visschers, L., and Wolthoff, R. (2015). Meeting technologies and optimal trading mechanisms in competitive search markets. *Journal of Economic Theory*, 155:1–15.

Lise, J., Meghir, C., and Robin, J.-M. (2016). Matching, sorting and wages. *Review of Economic Dynamics*, 19(1):63–87.

Lopes de Melo, R. (2018). Firm wage differentials and labor market sorting: Reconciling theory and evidence. *Journal of Political Economy*, 126(1):313–346.

Rose, E. K. (2021). Does banning the box help ex-offenders get jobs? evaluating the effects of a prominent example. *Journal of Labor Economics*, 39(1):79–113.

Shapley, L. and Shubik, M. (1971). The assignment game I: The core. *International Journal of Game Theory*, 1(1):111–130.

Shi, S. (2001). Frictional assignment I: Efficiency. *Journal of Economic Theory*, 98:232–260.

Shi, S. (2002). A directed search model of inequality with heterogeneous skills and skill-biased technology. *Review of Economic Studies*, 69(2):467–491.

Shimer, R. (2005). The assignment of workers to jobs in an economy with coordination frictions. *Journal of Political Economy*, 113(5):996–1025.

Shimer, R. and Smith, L. (2000). Assortative matching and search. *Econometrica*, 68(2):343–369.

Song, J., Price, D., Guvenen, F., Bloom, N., and von Wachter, T. (2019). Firming up inequality. *Quarterly Journal of Economics*, 134(1):1–50.

Tinbergen, J. (1956). On the theory of income distribution. *Weltwirtschaftliches Archiv*, 77:155–175.

Wolthoff, R. P. (2018). Applications and interviews: Firms' recruiting decisions in a frictional labor market. *Review of Economic Studies*, 85(2):1314–1351.