Search, Screening and Sorting*

Xiaoming Cai† Pieter Gautier‡ Ronald Wolthoff§

February 10, 2023

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.

JEL codes: D82, D83, E24.

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

*We are grateful to Jim Albrecht, Axel Anderson and Katka Borovičková for insightful discussions of this paper. We further thank Steve Davis, Philipp Kircher, Robert Shimer, and various seminar and conference participants for valuable comments. Wolthoff gratefully acknowledges 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 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.

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 offers some preliminary characterizations. Section 4 derives our main sorting results. Section 5 considers the market equilibrium and establishes that it is constrained efficient. In Section 6, we consider two extensions: i) noisy signals for every applicant and ii) endogenous choice of screening capacity. Finally, Section 7 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 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

[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.

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.

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.

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

[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).

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 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

[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.

[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.

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 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 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]

[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.

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

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 Social Planner's Problem

In this section, we analyze the problem of a social planner who aims to maximize surplus subject to the search frictions. We first derive surplus within a submarket before considering the optimal allocation of firms and workers.

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 B.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

[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.

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:

- 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

[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.

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 B.2.

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 $\sigma = 1$). To see this, let $\kappa(y)$ be a measure of output dispersion, defined as the relative gain in output for a firm of type y from hiring a high- rather than a low-type worker, i.e.

$$\kappa(y) \equiv \frac{f(x_2, y) - f(x_1, y)}{f(x_1, y)} > 0. \tag{4}$$

The following lemma then presents the planner's second-order condition (SOC).[16]

Lemma 2. *Surplus $S(\mu, \lambda, y)$ is strictly concave at a point (μ, λ) with $0 < \mu < \lambda$ if*

$$\frac{1}{\kappa(y)} > \frac{\phi_{\lambda\lambda} - \phi_{\mu\lambda}^2/\phi_{\mu\mu}}{-m''}. \tag{5}$$

Proof. See Appendix A.1. □

The right-hand side of (5) is a rescaled version of the determinant of the Hessian matrix of $\phi(\mu, \lambda)$. It is zero if $\sigma = 1$, which means that the SOC always holds in that case.[17] It is positive for $0 < \sigma < 1$ and converges to infinity when $\sigma \to 0$. That is, the SOC never holds when meetings are bilateral, as is well-known from

[16] We omit the arguments of the derivatives of $\phi(\mu, \lambda)$ and $m(\lambda)$ for simplicity.

[17] Cai et al. (2017) describe a broader class of meeting technologies for which $\phi(\mu, \lambda)$ is jointly concave in (μ, λ) such that (5) is always satisfied. However, as they show, such technologies feature (weakly) positive meeting externalities, making them unsuitable for our paper.

Eeckhout and Kircher (2010); in what follows, we will therefore focus on the case $\sigma > 0$, but our results extend to the bilateral case by continuity.

If the planner creates a submarket with queues (μ, λ), then (5) must hold, otherwise splitting the submarket increases total surplus. In general, the planner may wish to create multiple submarkets. Let K be this number and let γ_i, μ_i and λ_i be the measures of firms, the queue length of the high-type and of the low-type 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 3 (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 B.3. □

As a result of this lemma, the planner's problem can be rewritten as

$$\widehat{S}\left(Lz, L, y\right) = \max_{\gamma, \Delta} \ \gamma S\left(\frac{Lz}{\gamma}, \frac{L-\Delta}{\gamma}, y\right) + (1-\gamma)\, S\left(0, \frac{\Delta}{1-\gamma}, y\right), \tag{6}$$

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

$$\max_{\overline{\mu}(y),\overline{\lambda}(y)} \int_{\underline{y}}^{\overline{y}} \widehat{S}\left(\overline{\mu}(y),\overline{\lambda}(y),y\right)\, dJ(y), \tag{7}$$

subject to the linear constraints

$$\int_{\underline{y}}^{\overline{y}} \left(\overline{\lambda}(y)-\overline{\mu}(y)\right)\, dJ(y) \leq L(1-z), \tag{8}$$

$$\int_{\underline{y}}^{\overline{y}} \overline{\mu}(y)\, dJ(y) \leq Lz. \tag{9}$$

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 (6), 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 can instead be a convex combination of two different queues faced by different firms of the same type. However, if the planner creates only a single submarket for firms of type y, then these firms' 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 (7)–(9) is not necessarily unique. 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 (8) and (9). 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 (7), then for any given y, $(\overline{\mu}(y),\overline{\lambda}(y)) \in \mathbb{R}^2_+$ solves the maximization problem $\max_{\mu,\lambda} \widehat{S}(\mu,\lambda,y) - \mu W_2 - (\lambda-\mu)W_1$. Since $\widehat{S}(\mu,\lambda,y)$ is the concave envelope of $S(\mu,\lambda,y)$, the solution to

this problem can be obtained from

$$\max_{\mu,\lambda} \; m(\lambda) f(x_1, y) + \phi(\mu, \lambda) [f(x_2, y) - f(x_1, y)] - \mu W_2 - (\lambda - \mu) W_1. \tag{10}$$

If (10) has exactly one solution, all firms of type y are present in the same submarket; otherwise, by Lemma 3, (10) 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.[18]

In Section 5, we will show that the decentralized equilibrium is constrained efficient, so that workers' marginal contribution to surplus W_i equals their market utility. Thus, (10) corresponds to firms' profit maximization problem (26).

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

[18] 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 (10). 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 (10).

[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-

worker types, this definition can be expressed in terms of the probability that a firm hires a high-type worker, conditional on hiring someone,

$$h(\zeta, \lambda) \equiv \frac{\phi(\zeta\lambda, \lambda)}{m(\lambda)}, \tag{11}$$

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

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 (7), 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 3 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,

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.

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.

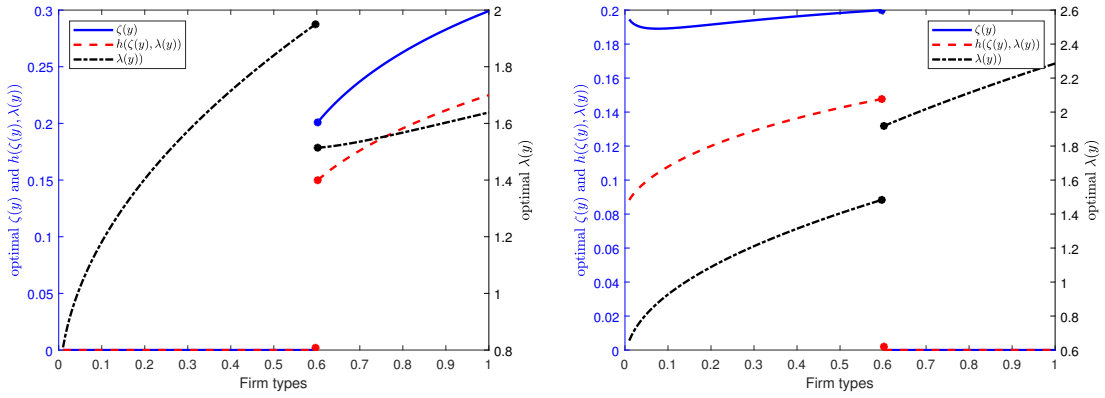
Definition 3. *The planner's solution exhibits positive assortative contacting (PAC) if $\zeta(y) = \mu(y)/\lambda(y)$ is (weakly) increasing in y for any selection $(\mu(y), \lambda(y)) \in \mathcal{Q}(y)$. Negative assortative contacting (NAC) is defined similarly with $\zeta(y)$ being (weakly) decreasing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

[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), so that output dispersion $\kappa(y)$ is between 2 and (around) 0.87. 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 (42) and (43) in Appendix B.4) 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 $\overline{y} = 1$.

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 (10) is strictly supermodular in (μ, y) for any given λ. Because this term is the only one in which μ and y interact, well-known results in monotone comparative statics imply that every selection of the optimal choice, $\mu(y)$, is strictly increasing in y for any given λ (see e.g. Milgrom and Shannon, 1994).[21] 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 (10) is strictly submodular in (λ, y) for 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

[21] For monotone comparative statics applied to sorting, see Chade et al. (2017).

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 (10) in terms of a choice of queue length λ and queue composition $\zeta = \mu/\lambda$, i.e.

$$\max_{\zeta,\lambda} \Pi(\zeta, \lambda, y) = m(\lambda) f^1 + \phi(\lambda\zeta, \lambda) \Delta f - \zeta\lambda W_2 - (1-\zeta)\lambda W_1, \tag{12}$$

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

$$m'(\lambda) f^1 + \frac{\partial \phi(\zeta\lambda, \lambda)}{\partial \lambda} \Delta f = W_1 + \zeta(W_2 - W_1), \tag{13}$$

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 (13) 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.

When an optimal ζ for firm y is interior $(0 < \zeta < 1)$, it must satisfy the FOC

$$\phi_\mu(\zeta\lambda, \lambda)\Delta f = W_2 - W_1. \tag{14}$$

This condition 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 (14) 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.[22] Of course, the optimal ζ might be at a corner, i.e. $\zeta = 0$ or 1, in which case the appropriate complementary slackness condition must hold.

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 cares primarily about matching probability while match quality is of secondary importance, which implies a unique optimal queue for each firm type (see the proof of Proposition 1 for details). Further, as $x_2 \to x_1$, the queue faced by firms of type y converges to a limit $(\zeta(y), \lambda(y))$, which is determined by the FOCs (13) and (14) 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 $\zeta(y) = \zeta \in (0, 1)$, 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 (13) 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.

$$\frac{d \log f(x, y)}{d \log y} = -\frac{d \log m'(\lambda(y))}{d \log \lambda(y)} \frac{d \log \lambda(y)}{d \log y}, \tag{15}$$

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

At the same time, it follows from (14) that for constant ζ to be optimal in the

[22] 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 $\lambda\Delta\zeta$.

limit, the elasticity of $f_x(x, y)$ with respect to y must equal

$$\frac{d \log f_x(x, y)}{d \log y} = -\frac{\partial \log \phi_\mu(\zeta\lambda(y), \lambda(y))}{\partial \log \lambda(y)} \frac{d \log \lambda(y)}{d \log y}. \tag{16}$$

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 (15) and (16), which yields

$$\rho(x, y) = \frac{\partial \log \phi_\mu(\zeta\lambda(y), \lambda(y))}{\partial \log m'(\lambda(y))}, \tag{17}$$

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 (17) 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,

$$a^c(\zeta, \lambda) \equiv \frac{\partial \log \phi_\mu(\zeta\lambda, \lambda)}{\partial \log m'(\lambda)} > 0. \tag{18}$$

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 4, $a^c(\zeta, \lambda)$ 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

$$a^m(\zeta, \lambda) \equiv \left. \frac{d \log \phi_\mu(\zeta\lambda, \lambda)}{d \log m'(\lambda)} \right|_{h(\zeta,\lambda)=\overline{h}} = a^c(\zeta, \lambda) \left(1 - \frac{\partial \phi_\mu / \partial \zeta}{\partial \phi_\mu / \partial \lambda} \frac{\partial h / \partial \lambda}{\partial h / \partial \zeta} \right) > 0, \tag{19}$$

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 (19) represents the relative effect of adjusting ζ so that $h(\zeta, \lambda)$ stays constant; as we will prove in Lemma 4, 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

$$\underline{\rho} \equiv \inf_{x,y} \rho(x, y) \geq \sup_{\zeta,\lambda} a^i(\zeta, \lambda) \equiv \overline{a}^i. \tag{20}$$

In fact, (20) 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 (20) is also necessary for PAC/PAM to always occur in the limit ($x_2 \to x_1$). The following proposition formalizes this idea.

Proposition 1. *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 (20) 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

$$\overline{\rho} \equiv \sup_{x,y} \rho(x,y) \leq \inf_{\zeta,\lambda} a^i(\zeta,\lambda) \equiv \underline{a}^i. \tag{21}$$

Proof. See Appendix A.2. □

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 4. *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)*

$$\overline{a}^c = \overline{a}^m = \frac{1+\sigma}{2} \quad \text{and} \quad \underline{a}^c = \underline{a}^m = \frac{1-\sigma}{2}. \tag{22}$$

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.3. □

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 $\zeta = 1$.[23] 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 (32) in Appendix A.3. It satisfies $a^m(\zeta, \lambda) = 1/2$ when $\sigma = 0$ while $a^m(\zeta, \lambda) = \zeta$ when $\sigma = 1$.

Together with Proposition 1, Lemma 4 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 (20) and (21) 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 (20) 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 positive sorting always holds when $x_2 \to x_1$, then it also holds for any x_1 and x_2. The following proposition formally establishes this result.

Proposition 2. *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*

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

is unique if $\underline{\rho} \geq (1+\sigma)/2$ *or* $\overline{\rho} \leq (1-\sigma)/2$.

Proof. See Appendix B.4. □

Given Definition 1, we can alternatively state Proposition 2 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 (12) is nonconcave. It is therefore natural to consider the methods of monotone comparative statics to derive the sufficient condition for sorting (see e.g. Milgrom and Shannon, 1994).[24] However, this approach does not suffice for our model because it does not take into account that whenever multiple solutions arise, one solution must be $\zeta = 0$. In the proof of Proposition 2, we therefore separately consider firm types for which multiple optimal queues exist and firm types for which a unique optimal queue exists.

More precisely, we show in Appendix B.4 that the two reasons that PAC fails in Figure 1b never arise under the sufficient condition for PAC/PAM: First, if firms of type y_m have two submarkets (for example, $y_m = 0.6$ in both Figure 1a and 1b), then $\zeta(y)$ will jump up around type y_m. Second, if firms of each type have a unique optimal queue within some interval of firm types (this is the case in both Figure 1a and 1b when $y < 0.6$ or $y > 0.6$), then both $\zeta(y)$ and $h(\zeta(y), \lambda(y))$ are increasing within the interval. Therefore, if all firm types have a unique optimal queue, then the second case above already implies that PAC/PAM holds; else the planner's solution must look like Figure 1a, and then PAC/PAM holds.

4.5 Effect of Screening

We can now consider how screening affects sorting. It follows from Proposition 2 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

[24] For PAC, this approach requires that one uses (13) to obtain the optimal λ as a function of ζ and y, denoted by $\lambda^o(\zeta, y)$. Plugging it into (12) gives firms' expected profit as a function of ζ and y only: $\widetilde{\Pi}(\zeta, y) = \Pi(\zeta, \lambda^o(\zeta, y), y)$. PAC then holds if $\widetilde{\Pi}(\zeta, y)$ is supermodular in (ζ, y).

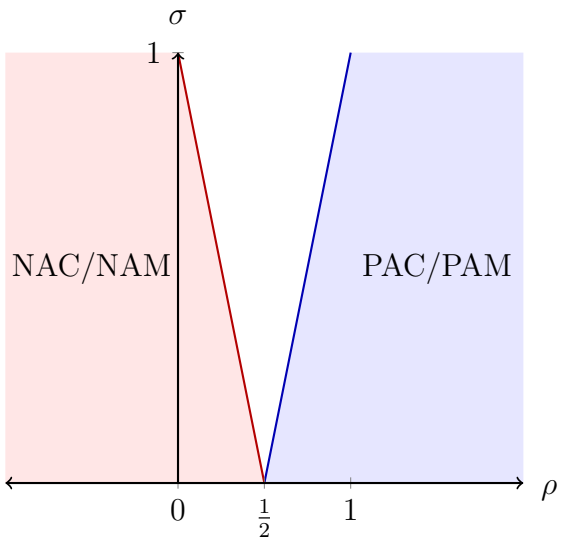


Figure 2: Combinations of ρ and σ that give rise to PAC/PAM (blue) or NAC/NAM (red) for any distribution of agents' types, assuming a CES production function.

requires square-root supermodularity, in line with Eeckhout and Kircher (2010). 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 again the contribution to surplus by a high-type applicant. 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.

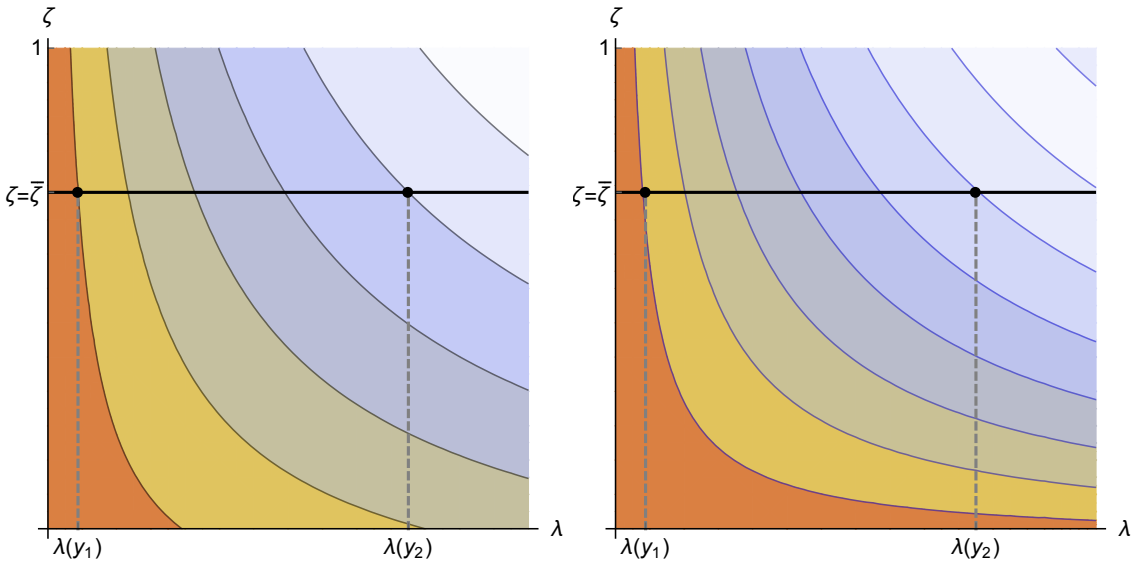
In contrast, when the queue mainly consists of high types (ζ is high), multiple interviews with high-type applicants are a key concern. A longer queue makes this outcome more likely and this force is *amplified* by an increase in firms' screening ability σ, since it increases every applicant's interviewing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

Figure 3 illustrates this intuition for PAC/NAC by showing the level curves of ϕ_μ as a function of ζ and λ for two different values of σ. An increase in σ brings these level curves further apart for low values of ζ but closer together for high values of ζ, in line with the discussion above. For PAM/NAM, the same is true as can readily been seen from the fact that $a^m(\zeta, \lambda)$ is strictly decreasing in σ when $\zeta < 1/2$ and strictly increasing in σ when $\zeta > 1/2$.

Bilateral Limit. Eeckhout and Kircher (2010) analyze a model where firms can interview only a single applicant ($\sigma = 0$ in our benchmark technology) with probability $m(\lambda)$. In that case, $\phi(\mu, \lambda) = m(\lambda)\mu/\lambda$ and $a^c(\zeta, \lambda) = a^m(\zeta, \lambda) = m'(\lambda)(\lambda m'(\lambda) - m(\lambda))/(\lambda m(\lambda) m''(\lambda))$, which is independent of ζ and which is precisely the object (the elasticity of substitution of the total number of matches) that Eeckhout and Kircher (2010) show to be important in their study of sorting patterns for bilateral technologies.

Sorting with Few High-Type Workers. Since Proposition 2 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 3. 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 patterns.)



(a) PAC/NAC: low σ (b) PAC/NAC: high σ

Figure 3: Level curves of ϕ_μ for two values of σ, where ϕ_μ is decreasing from the bottom left to the top right.

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]

Proposition 3. *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 B.5. □

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

[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).

$a^m(\zeta, \lambda)$, 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 Market Equilibrium

In this section, we establish that the market equilibrium where firms post wage menus 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).

5.1 Beliefs, Strategies, and Equilibrium Definition

Beliefs. A firm of type y posting a wage menu $\boldsymbol{w}$ has to form beliefs about its queues $(\mu(\boldsymbol{w}, y), \lambda(\boldsymbol{w}, y))$. Following the standard approach in the literature, we restrict these beliefs in the spirit of subgame perfection through what is known as the *market utility condition*. To state this condition, consider a worker of type x_i. Define $V_i(\boldsymbol{w}, \mu, \lambda, y)$ as his expected payoff in a submarket $(\boldsymbol{w}, y)$ with queues (μ, λ), and his *market utility* U_i as the maximum expected payoff that he can obtain in equilibrium, either by visiting one of the submarkets or by remaining inactive. Firms' beliefs $(\mu(\boldsymbol{w}, y), \lambda(\boldsymbol{w}, y))$ must then satisfy

$$\begin{cases} V_1(\boldsymbol{w}, \mu, \lambda, y) \leq U_1, \text{ with equality if } \lambda - \mu > 0, \\ V_2(\boldsymbol{w}, \mu, \lambda, y) \leq U_2, \text{ with equality if } \mu > 0. \end{cases} \tag{23}$$

For common meeting technologies, including our benchmark as we will show in Lemma 5 below, (23) admits a unique solution (μ, λ), which is then the firm's belief. For other technologies, there can be multiple solutions to (23). The standard assumption is then that firms are optimistic and expect the solution that maxi-

[26] By Lemma 4, $a^m(\zeta, \lambda)$ 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 $\lambda\zeta < \sqrt{2(1-\zeta)} - 1$).

mizes their expected payoff $\pi(\boldsymbol{w}, \mu, \lambda, y)$. Explicit expressions for π and V_i will be provided in Section 5.2.

Strategies. Let $G(\boldsymbol{w} \mid y)$ denote the (conditional) probability that a firm of type y offers a wage menu $\widetilde{\boldsymbol{w}} \leq \boldsymbol{w}$, where $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \widetilde{w}_2)$, $\boldsymbol{w} = (w_1, w_2)$, $\widetilde{w}_1 \leq w_1$ and $\widetilde{w}_2 \leq w_2$. Given market utilities (U_1, U_2), firm optimality means that every $\boldsymbol{w}$ in the support of $G(\boldsymbol{w} \mid y)$ must maximize $\pi(\boldsymbol{w}, \mu, \lambda, y)$ subject to (23).

Similarly, let $H_i(\boldsymbol{w}, y)$ denote the probability that workers of type x_i apply to a firm with $\widetilde{\boldsymbol{w}} \leq \boldsymbol{w}$ and $\widetilde{y} \leq y$. The following accounting identities then link workers' strategies $H_1(\boldsymbol{w}, y)$ and $H_2(\boldsymbol{w}, y)$ to the queues in different submarkets.

$$H_1(\boldsymbol{w}, y) = \frac{1}{L(1-z)} \int_{\widetilde{y} \leq y} \int_{\widetilde{\boldsymbol{w}} \leq \boldsymbol{w}} \left[\lambda(\widetilde{\boldsymbol{w}}, \widetilde{y}) - \mu(\widetilde{\boldsymbol{w}}, \widetilde{y})\right] dG(\widetilde{\boldsymbol{w}} \mid \widetilde{y})\, dJ(\widetilde{y}). \tag{24}$$

$$H_2(\boldsymbol{w}, y) = \frac{1}{Lz} \int_{\widetilde{y} \leq y} \int_{\widetilde{\boldsymbol{w}} \leq \boldsymbol{w}} \mu(\widetilde{\boldsymbol{w}}, \widetilde{y})\, dG(\widetilde{\boldsymbol{w}} \mid \widetilde{y})\, dJ(\widetilde{y}), \tag{25}$$

Optimality requires that workers must obtain exactly U_i at any firm to which they apply with positive probability, and weakly less at other firms i.e. (23) must hold. Further, note that no firm will post a wage menu $\boldsymbol{w} \geq \overline{\boldsymbol{w}} \equiv (f(x_1, \overline{y}), f(x_2, \overline{y}))$. Thus, $H_i(\overline{\boldsymbol{w}}, \overline{y})$ is the probability that workers of type x_i apply, which must equal 1 if $U_i > 0$, as the payoff from not sending an application is zero. This condition can be interpreted as "market clearing": in equilibrium, demand for each type of applicant must equal supply, which determines the "market prices" U_1 and U_2.

Equilibrium Definition. We can now define an equilibrium as follows.

Definition 4. *A (directed search) equilibrium is a triple* $(G, \{H_1, H_2\}, \{U_1, U_2\})$ *satisfying ...*

(i) Firm Optimality. *Given* (U_1, U_2)*, every wage menu* $\boldsymbol{w}$ *in the support of* $G(\cdot \mid y)$ *maximizes* $\pi(\boldsymbol{w}, \mu(\boldsymbol{w}, y), \lambda(\boldsymbol{w}, y), y)$ *for each firm type* y*, where the queue lengths* $(\mu(\boldsymbol{w}, y), \lambda(\boldsymbol{w}, y))$ *are determined by* (23)*.*

(ii) Worker Optimality. *Given* (U_1, U_2)*, the application strategy of high-type and low-type workers satisfies* (25) *and* (24)*, respectively, where the queue lengths* $(\mu(\boldsymbol{w}, y), \lambda(\boldsymbol{w}, y))$ *are determined by* (23)*. Further,* $H_i(\overline{\boldsymbol{w}}, \overline{y}) = 1$ *if* $U_i > 0$*.*

5.2 Payoffs and Efficiency

Ranking. How firms rank workers depends on which worker types are most profitable, which in turn depends on the wage contracts that firms post. To simplify exposition, assume for now that firms post wage menus (w_1, w_2) satisfying

$$f(x_2, y) - w_2 > f(x_1, y) - w_1, \tag{26}$$

i.e. more productive workers are more profitable and are therefore preferred by firms. Later, in Lemma 6, we will show that (26) must indeed hold when firms act optimally, making our assumption without loss of generality.

Payoffs. The expected payoff of a firm facing a queue (μ, λ) equals

$$\pi(\boldsymbol{w}, \mu, \lambda, y) = \phi(\mu, \lambda)\left[f(x_2, y) - w_2\right] + \left[m(\lambda) - \phi(\mu, \lambda)\right]\left[f(x_1, y) - w_1\right]. \tag{27}$$

The firm 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 with probability $m(\lambda) - \phi(\mu, \lambda)$.

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

$$\psi_1(\mu, \lambda) = \frac{m(\lambda) - \phi(\mu, \lambda)}{\lambda - \mu} \qquad \text{or} \qquad \psi_2(\mu, \lambda) = \frac{\phi(\mu, \lambda)}{\mu}. \tag{28}$$

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 $\psi_2(0, \lambda) = \phi_\mu(0, \lambda)$.

Uniqueness of Queues. In a submarket $(\boldsymbol{w}, y)$, the queues (μ, λ) are determined by (23). Since this is a system of non-linear equations, it is not immediate that there is a unique solution. Lemma 5 guarantees uniqueness.

Lemma 5. *Suppose that ϕ is given by (2). Given market utilities U_1 and U_2, there exists exactly one solution (μ, λ) to the market utility condition (23) for any wage menu $\boldsymbol{w}$.*

Proof. See Appendix B.6. □

Competitive Market for Queues. As standard in the literature, we can use the market utility condition (23) to substitute the wages w_1 and w_2 out of (27) and rewrite the firm's problem with queue lengths as choice variables. This yields

$$\max_{0\leq\mu\leq\lambda} m(\lambda) f(x_1, y) + \phi(\mu, \lambda) [f(x_2, y) - f(x_1, y)] - \lambda U_1 - \mu (U_2 - U_1). \quad (29)$$

Equation (29) 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.[27] This payoff is similar to equation (12) except that the costs that the firm faces are now workers' market utilities instead of their marginal contribution to surplus.

Productivity versus Profitability. We now show that it is without loss of generality to only consider wage menus satisfying (26). To do so, Lemma 6 establishes two results.[28] First, the maximum profit in (29) can always be obtained with a wage menu (w_1^*, w_2^*) that satisfies (26). Second, a wage menu that violates (26) always yields a strictly lower profit. To understand the latter result, suppose that a firm posts a wage menu where low-type workers yield a higher profit ex post, i.e. $f(x_2, y) - w_2 < f(x_1, y) - w_1$, and attracts a queue (μ, λ) with $0 < \mu < \lambda$. Workers must obtain their market utility, so the expected transfer from the firm to the workers equals $\mu U_2 + (\lambda - \mu) U_1$. However, giving priority to low- rather than high-type workers reduces surplus relative to $S(\mu, \lambda, y)$ in (3). Hence, the firm's expected profit is strictly smaller than the maximum profit in (29).

Lemma 6. *A solution (μ^*, λ^*) (interior or corner) to the firm's problem (29) can be implemented with the wage menu $(w_1^*, w_2^*) = (U_1/\psi_1(\mu^*, \lambda^*), U_2/\psi_2(\mu^*, \lambda^*))$, which satisfies (26). Further, any wage menu violating (26) yields a strictly lower payoff than (w_1^*, w_2^*).*

Proof. See Appendix B.7. □

[27] 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 (29) also holds if $\mu = 0$ or $\mu = \lambda$.

[28] A similar result appears in Shimer (2005) for urn-ball meetings. Lemma 6 generalizes his result to arbitrary meeting technologies.

If the solution is interior ($0 < \mu^* < \lambda^*$), then the wage menu that firms need to post to attract the optimal queue is unique. In a corner solution ($\mu^* = 0$ or $\mu^* = \lambda^*$), the wage menu is not unique, but Lemma 6 describes the maximum wages satisfying (26).[29]

Observability of Firm Productivity. By Lemma 6, all firms will post wage contracts such that high-type workers are more profitable. Given the wage contract, the market utility condition then determines the queue length and composition. Since workers only care about their hiring probability and the wage, this then means that all our results carry through if they do not observe firm types.

Efficiency. In sum, we have demonstrated that the market equilibrium with wage menus 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 4. *The market equilibrium is constrained efficient, i.e., the equilibrium outcome solves the planner's problem given by* (7).

6 Extensions

6.1 Signals

In our benchmark model, firms have no information about applicants' types when selecting interviewees. In practice, it is often easy to obtain some information from applicants' resumes. To capture this idea, suppose that firms can costlessly observe a signal for every applicant. For high-type applicants, the signal is positive with certainty. In contrast, a low-type applicant generates a negative signal with probability $\tau \in [0, 1]$ and a positive signal with probability $1 - \tau$. Hence, the signal is perfect if $\tau = 1$, but pure noise if $\tau = 0$. Using this information, firms first interview applicants with positive signals and only interview applicants with negative signals if interview capacity remains. As before, an interview reveals the applicant's true type.

The following proposition establishes that this modified environment is isomorphic to our baseline model, as long as we transform the parameter σ.

[29] For example, if $\mu^* = \lambda^*$, then the optimal w_2 is uniquely given by $U_2/\psi_2(\lambda^*, \lambda^*)$, but the optimal w_1 can take any value between zero and $w_1^* = U_1/\psi_1(\lambda^*, \lambda^*)$.

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 B.8. □

As a corollary, all our earlier results carry over to the environment with signals, 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.

6.2 Endogenous Screening

In our baseline model, the screening intensity σ is exogenous. However, firms can generally influence the number of applicants that they interview. In Appendix B.9, we therefore analyze an extension in which firms can choose (and post) their recruiting intensity $\sigma \in [0,1]$ at a linear cost $c\sigma$ where $c \geq 0$. Since $\phi(\mu, \lambda)$ given by equation (2) is convex in σ, the surplus function S in (3) and hence firms' expected profit is also convex in σ, which implies that firms will either choose $\sigma = 0$ or 1. This simplifies the analysis considerably. Firms either rely on ex ante screening through wage menus and try to hire a particular type of worker, or combine ex ante screening with the maximal amount of ex post screening and encourage both types of workers to apply. The main result is that it is more difficult to obtain PAC/PAM with endogeneous σ. First, there does not exist a sufficient condition that gurentees PAC/PAM for any distribution of agents' types and any screening cost c. Second, for a given distribution of agent types, the sufficient condition (30) below is more stringent than log-supermodularity. The reason is that the additional gain from ex-post screening relative to only ex-ante screening can be highly non-monotonic in firm types. The following proposition formalizes this.[30]

Proposition 6. *In our environment with endogenous screening, the following holds:*

(i) Equilibrium exhibits NAC/NAM for any distribution of agents' types and any cost c *if (resp. only if)* $f(x,y)$ *is strictly (resp. weakly) submodular.*

[30] The result that log-supermodularity is not sufficient for PAC/PAM may seem puzzling. We provide detailed intuition in Appendix B.9.

(ii) Given any log-supermodular function f, we can find a distribution of agents' types and a screening cost c such that PAC/PAM fails in equilibrium. However, given a distribution of agents' types, PAC/PAM holds in equilibrium (for any screening cost c) if

$$\underline{\rho} \geq \Omega\left(\kappa(\underline{y})\right), \tag{30}$$

where $\kappa(\cdot)$ is defined by (4), $\underline{y}$ is the lowest firm type, and $\Omega(\kappa) \equiv 1/2 + \ln(\sqrt{\kappa}+\sqrt{1+\kappa})/\ln(1+\kappa)$, which is strictly decreasing with $\lim_{\kappa\to 0}\Omega(\kappa) = \infty$ and $\lim_{\kappa\to\infty}\Omega(\kappa) = 1$.

Proof. See Appendix B.9. □

7 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 types of 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 ex-post screening is easier (firms can screen more applicants), sorting may be harder in the sense that stronger complementarities in the production technology 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).

There are several promising avenues for future research. On the theoretical side, in markets with a long hiring cycle, like the academic job market, workers may have strong incentives to send multiple applications simultaneously. This reduces 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. In order to identify complementarities in production, we may 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.

References

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

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.

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. (2020). Job applications and labor market flows. mimeo.

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.

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.

Chade, H., Eeckhout, J., and Smith, L. (2017). Sorting through search and matching models in economics. *Journal of Economic Literature*, 55(2):493–544.

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.

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.

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.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1):157–180.

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.

Appendix A Proofs

A.1 Proof of Lemma 2

The Hessian $\mathcal{H}(\mu, \lambda, y)$ of $S(\mu, \lambda, y)$ equals

$$\mathcal{H}(\mu, \lambda, y) = \begin{pmatrix} \phi_{\mu\mu}\Delta f & \phi_{\mu\lambda}\Delta f \\ \phi_{\mu\lambda}\Delta f & m''f^1 + \phi_{\lambda\lambda}\Delta f \end{pmatrix}.$$

When $\sigma > 0$, we have $\phi_{\mu\mu} < 0$. So, the Hessian is negative definite if and only if its determinant is positive, i.e. $\Delta f \left[m''\phi_{\mu\mu}f^1 + \left(\phi_{\mu\mu}\phi_{\lambda\lambda} - \phi_{\mu\lambda}^2\right)\Delta f\right] > 0$. Using $\Delta f > 0$ and the definition of $\kappa(y)$, we obtain condition (5). □

A.2 Proof of Proposition 1

Consider first the degenerate case $x = x_1 = x_2$. Surplus in a submarket equals $m(\lambda)f(x, y)$, so the marginal contribution of a worker is $m'(\lambda)f(x, y)$, which must be the same across different submarkets. That is, the optimal queue length $\lambda(y)$ satisfies $m'(\lambda(y))f(x, y) = W$, where W is a constant such that $\int_{\underline{y}}^{\overline{y}} \lambda(y) = L$.

When x_2 is sufficiently close to $x_1 = x$, then the marginal contributions of low- and high-type workers will be close to W, which implies that to solve the planner's problem, it is without loss of generality to limit the queue length of each firm to $\overline{\lambda} \equiv 2\lambda(\overline{y})$, where $\lambda(\overline{y})$ is the optimal queue length of the firm with the highest type in the degenerate case. That is, to solve the planner's problem in (7), we can restrict $(\mu(y), \lambda(y))$ to be in the convex set $\Delta \equiv \{(\mu, \lambda) \,|\, 0 \leq \mu \leq \lambda \leq \overline{\lambda}\}$.

In this set Δ, the right-hand side of the firm's SOC (5) is bounded due to continuity. Hence, (5) will hold for all (μ, λ) in Δ when $\kappa(y)$, or equivalently $x_2 - x_1$, is sufficiently small. That is, for each firm type y, the surplus function $S(\mu, \lambda, y)$ is strictly concave on the set Δ, which implies that in the planner's problem in (7), we can replace $\widetilde{S}(\mu, \lambda, y)$ with $S(\mu, \lambda, y)$ because two submarkets in Δ are strictly suboptimal. Thus the planner solves a standard (strictly) concave maximization problem; the optimal solution $(\mu(y), \lambda(y))$ is unique and continuous. Furthermore, when $\mu(y)$ and $\lambda(y)$ satisfy $0 < \mu(y) < \lambda(y)$ for some firm type y, they are jointly determined by the FOCs (13) and (14).

As $x_2 \to x_1 = x$, the FOCs (13) and (14) converge to $m'(\lambda(y))f(x, y)$ (which is constant across firms with $\lambda(y) > 0$) and $\phi_\mu(\lambda(y)\zeta(y), \lambda(y))f_x(x, y)$ (which is

constant across firms with $\zeta(y) \in (0,1)$). Differentiating the two FOCs with respect to y yields

$$0 = m''(\lambda(y))\lambda'(y)f(x,y) + m'(\lambda(y))f_y(x,y)$$
$$0 = \left(\frac{\partial \phi_\mu}{\partial \zeta(y)}\zeta'(y) + \frac{\partial \phi_\mu}{\partial \lambda(y)}\lambda'(y)\right) f_x(x,y) + \phi_\mu f_{xy}(x,y),$$

where we suppress the arguments of $\phi_\mu(\lambda(y)\zeta(y), \lambda(y))$. Combining these two equations yields $\zeta'(y)$ and $\lambda'(y)$, which then implies that $\zeta'(y) \geq 0$ if and only if $\rho(x,y) \geq a^c(\zeta(y), \lambda(y))$ and $\frac{d}{dy}h(\zeta(y), \lambda(y)) \geq 0$ if and only if $\rho(x,y) \geq a^m(\zeta(y), \lambda(y))$.

Next, we show the necessity of (20) and (21). We only consider the case of PAC; the other cases (PAM, NAC and NAM) follow the same logic. Suppose that (20) does not hold for $i = c$, so that there exist x_0, y_0, ζ_0, and λ_0 such that $\rho(x_0, y_0) < a^c(\zeta_0, \lambda_0)$. We can then construct a counterexample in which worker/firm heterogeneity is small and NAC holds at the planner's solution. In particular, by continuity, we can assume that $0 < \zeta_0 < 1$ (note the strict inequality), and that there exists a small ϵ_0 such that the above inequality holds for all $x \in [x_0, x_0 + \epsilon_0]$, $y \in [y_0 - \epsilon_0, y_0 + \epsilon_0]$, $\zeta \in [\zeta_0 - \epsilon_0, \zeta_0 + \epsilon_0]$, and $\lambda \in [(1-\epsilon_0)\lambda_0, (1+\epsilon_0)\lambda_0]$. Fix ϵ_0 from now on and set $x_1 = x_0$, $Lz = \lambda_0\zeta_0$, $L(1-z) = \lambda_0(1-\zeta_0)$, $\underline{y} = y_0 - \epsilon_1$, and $\overline{y} = y_0 + \epsilon_1$ for some $\epsilon_1 \leq \epsilon_0$. Next, we reduce firm heterogeneity by letting $\epsilon_1 \to 0$. When ϵ_1 is sufficiently small, $\lambda(y) \in [(1-\epsilon_0)\lambda_0, (1+\epsilon_0)\lambda_0]$ and $\zeta(y) \in [\zeta_0 - \epsilon_0, \zeta_0 + \epsilon_0]$ for all y. Thus, NAC holds at the planner's solution. □

A.3 Proof of Lemma 4

We first consider $a^c(\zeta, \lambda)$. Since $\phi(\mu, \lambda)$ is given by equation (2) and $a^c(\zeta, \lambda)$ is defined by equation (18), direct calculation yields

$$a^c(\zeta, \lambda) = \frac{1+\lambda}{2\lambda}\left(1 + \frac{1}{1+(1-\sigma)\lambda} - \frac{2}{1+\sigma\zeta\lambda+(1-\sigma)\lambda}\right). \tag{31}$$

Note that $a^c(\zeta, \lambda)$ is strictly increasing in ζ. Thus, we have $\max_\zeta a^c(\zeta, \lambda) = a^c(1, \lambda)$ and $\min_\zeta a^c(\zeta, \lambda) = a^c(0, \lambda)$. Moreover, (31) reveals that $a^c(0,\lambda) + a^c(1,\lambda) = 1$ and $\frac{da^c(1,\lambda)}{\lambda} = -\frac{\sigma(1-\sigma)}{2(1+(1-\sigma)\lambda)^2} < 0$. Therefore, $a^c(1,\lambda)$ approaches its supremum when $\lambda \to 0$ and $a^c(0,\lambda)$ approaches its infimum when $\lambda \to 0$. Hence, we have $\sup_{\zeta,\lambda} a^c(\zeta,\lambda) = \lim_{\lambda\to 0} a^c(1,\lambda) = (1+\sigma)/2$ and $\inf_{\zeta,\lambda} a^c(\zeta,\lambda) = 1 - \sup_{\zeta,\lambda} a^c(\zeta,\lambda) = (1-\sigma)/2$, where

neither the infimum nor the supremum can be reached because we require $\lambda > 0$. Furthermore,

$$\frac{\partial a^c(\zeta,\lambda)}{\partial \sigma} = \frac{1+\lambda}{2}\left(\frac{1}{(1+\lambda(1-\sigma))^2} - \frac{2(1-\zeta)}{(1+\lambda(1-\sigma)+\lambda\sigma\zeta)^2}\right).$$

Hence, $a^c(\zeta,\lambda)$ is strictly increasing in σ if and only if $\frac{\lambda\zeta\sigma}{1+\lambda(1-\sigma)} > \sqrt{2(1-\zeta)} - 1$.

Next, we consider $a^m(\mu,\lambda)$. Analogous to above, direct computation yields

$$a^m(\zeta,\lambda) = \frac{1}{2}\left(1 + \frac{\sigma(2\zeta-1)}{1+(1-\sigma)\lambda}\right). \tag{32}$$

Note that $a^m(\zeta,\lambda)$ is strictly increasing in ζ. For a given λ, $a^m(\zeta,\lambda)$ therefore reaches its minimum at $\zeta = 0$ and its maximum at $\zeta = 1$. Because $a^m(0,\lambda) = a^c(0,\lambda)$ and $a^m(1,\lambda) = a^m(1,\lambda)$, we have $\overline{a}^m = \overline{a}^c$ and $\underline{a}^m = \underline{a}^c$. The above equation implies that $a^m(\zeta,\lambda)$ is strictly increasing in σ if and only if $\zeta > 1/2$. When $\zeta = 1/2$, $a^m(1/2,\lambda) = 1/2$, independent of λ.

Finally, note that

$$a^c(\zeta,\lambda) - a^m(\zeta,\lambda) = \frac{\zeta(1-\zeta)\sigma^2\lambda}{(1+(1-\sigma)\lambda)(1+\sigma\zeta\lambda+(1-\sigma)\lambda)} \geq 0.$$

Thus, when $\sigma > 0$, $a^c(\zeta,\lambda) = a^m(\zeta,\lambda)$ if and only if $\zeta = 0$ or $\zeta = 1$. □

Appendix B Online Appendix

B.1 Proof of Lemma 1

First, consider the application stage. Given queue length λ, a firm's number of applicants n_A in our benchmark model follows a geometric distribution with support $\mathbb{N}_0$ and mean λ, i.e. $\mathbb{P}[n_A = n \,|\, \lambda] = \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^n$ for $n = 0, 1, 2, \ldots$. If $\sigma = 1$ (firms can interview all candidates), then we have

$$\phi(\mu, \lambda) = 1 - \sum_{n=0}^{\infty} \mathbb{P}[n_A = n \,|\, \lambda] \left(1 - \frac{\mu}{\lambda}\right)^n = \frac{\mu}{1+\mu},$$

where the first equality uses the fact that the probability that an applicant is high-type is μ/λ and is independent across applicants.

Next, consider the screening stage. A firm's *potential* number of interviews, n_C, follows a geometric distribution with support $\mathbb{N}_1$ and mean $(1-\sigma)^{-1}$. That is, $\mathbb{P}[n_C \geq n \,|\, \sigma] = \sigma^{n-1}$ for $n = 1, 2, \ldots$. Since interviewing might be constrained by the number of applications, the firm's *actual* number of interviews is $n_I = \min\{n_A, n_C\} \in \mathbb{N}_0$, distributed according to $\mathbb{P}[n_I \geq n \,|\, \lambda, \sigma] = \mathbb{P}[n_A \geq n \,|\, \lambda]\, \sigma^{n-1} = \left(\frac{\lambda}{1+\lambda}\right)^n \sigma^{n-1}$. An interview reveals a high-type worker with probability μ/λ, independently across applicants. The firm therefore interviews at least one high-type worker with probability

$$\phi(\mu, \lambda) = 1 - \sum_{n=0}^{\infty} \mathbb{P}[n_I = n | \lambda, \sigma] \left(1 - \frac{\mu}{\lambda}\right)^n = \sum_{n=1}^{\infty} \mathbb{P}[n_I \geq n | \lambda, \sigma] \frac{\mu}{\lambda} \left(1 - \frac{\mu}{\lambda}\right)^{n-1},$$

where the second equality follows from summation by parts, analogous to (**??**). Substituting $\mathbb{P}[n_I \geq n \,|\, \lambda, \sigma] = \left(\frac{\lambda}{1+\lambda}\right)^n \sigma^{n-1}$ yields equation (2). □

B.2 Marginal Contributions

Adding more low-type workers to a submarket only increases λ, while adding more high-type workers increases both μ and λ. Thus, the marginal contribution of low-type and high-type workers at a firm of type y with queues (μ, λ) are $S_\lambda(\mu, \lambda, y)$ and $S_\mu(\mu, \lambda, y) + S_\lambda(\mu, \lambda, y)$, respectively. Because of constant returns to scale, the firm's marginal contribution is the difference between total surplus and the sum of the

marginal contributions of its applicants, i.e. $S(\mu,\lambda,y)-\mu S_\mu(\mu,\lambda,y)-\lambda S_\lambda(\mu,\lambda,y)$.[31] Using $S(\mu,\lambda,y)$ from (3), $f^1 \equiv f(x_1,y)$ and $\Delta f = f(x_2,y)-f(x_1,y)$, we get

$$T_1(\mu,\lambda,y) = m'(\lambda)f^1 + \phi_\lambda(\mu,\lambda)\Delta f, \tag{33}$$

$$T_2(\mu,\lambda,y) = m'(\lambda)f^1 + (\phi_\mu(\mu,\lambda)+\phi_\lambda(\mu,\lambda))\,\Delta f, \tag{34}$$

$$R(\mu,\lambda,y) = (m(\lambda)-\lambda m'(\lambda))\,f^1 + (\phi(\mu,\lambda)-\mu\phi_\mu(\mu,\lambda)-\lambda\phi_\lambda(\mu,\lambda))\,\Delta f, \tag{35}$$

where T_1, T_2 and R are the marginal contribution to surplus of low-type workers, high-type workers, and firms, respectively.

B.3 Proof of Lemma 3

We prove this result and discuss it extensively in Cai et al. (2022). Here, we state the single-crossing condition and briefly argue why it leads to Lemma 3. To do so, we define $H(\mu,\lambda)$ as the right-hand side of (5), i.e.

$$H(\mu,\lambda) \equiv \frac{\phi_{\lambda\lambda}-\phi_{\mu\lambda}^2/\phi_{\mu\mu}}{-m''}. \tag{36}$$

Cai et al. (2022) then show that Lemma 3 holds whenever a meeting technology satisfies Property A0, A1, A2 and the following A3.

A3. (single-crossing condition) At any point (ζ,λ) where $H(\lambda\zeta,\lambda)>0$, we have $\partial H(\lambda\zeta,\lambda)/\partial\lambda > 0$ and

$$-\frac{\partial\phi_\mu(\lambda\zeta,\lambda)/\partial\zeta}{\partial\phi_\mu(\lambda\zeta,\lambda)/\partial\lambda} < -\frac{\partial H(\lambda\zeta,\lambda)/\partial\zeta}{\partial H(\lambda\zeta,\lambda)/\partial\lambda}. \tag{37}$$

Note that Property A0 states that $\partial\phi_\mu(\lambda\zeta,\lambda)/\partial\zeta<0$, while Property A2 states that $\partial\phi_\mu(\lambda\zeta,\lambda)/\partial\lambda<0$, making the left-hand side of (37) strictly negative. When $\phi(\mu,\lambda)$ is given by (2), direct computation reveals that both $H(\lambda\zeta,\lambda)$ and the right-hand side of (37) are strictly positive. Thus, Property A3 is trivially satisfied in this case.

Let $R(\mu,\lambda,y)$ and $T_2(\mu,\lambda,y)$ denote the marginal contributions to surplus of firms and high-type workers, respectively, as derived in Appendix B.2. The idea of the proof of Cai et al. (2022) is then as follows. Suppose that the marginal

[31] Alternatively, increase the number of firms by a factor $1+\Delta s$. The additional surplus is then $(1+\Delta s)S(\mu/(1+\Delta s),\lambda/(1+\Delta s),y)-S(\mu,\lambda,y)$, which yields the same result when $\Delta s\to 0$.

contribution to surplus of firms equals R^*. Property A3 then implies that, in the λ-ζ plane, the level curve $R(\lambda\zeta, \lambda, y) = R^*$ crosses the level curve $H(\lambda\zeta, \lambda) = 1/\kappa(y)$ at most once and from the left, as illustrated in Figure 1 of Cai et al. (2022). If the intersection exists, denote it by (λ^*, ζ^*). Along the level curve $R(\lambda\zeta, \lambda, y) = R^*$, the SOC (5) is then satisfied for $\zeta > \zeta^*$ and violated for $\zeta < \zeta^*$. The only feasible submarket when $\zeta < \zeta^*$ is therefore the corner solution $\zeta = 0$. Furthermore, along the level curve $R(\lambda\zeta, \lambda, y) = R^*$, the marginal contribution to surplus by high-type workers, $T_2(\lambda\zeta, \lambda, y)$ is monotonically decreasing in ζ for $\zeta \geq \zeta^*$. Since the marginal contribution of high-type workers must be the same among all submarkets containing such workers, there can exist only one submarket with $\zeta \geq \zeta^*$. Hence, there exist at most two submarkets: one with $\zeta = 0$ and the other with $\zeta \geq \zeta^*$. Cai et al. (2022) then show that there exists only one pair of (γ, Δ) which satisfies the FOC for the maximization problem in (6). Hence, the planner's solution is unique. □

B.4 Proof of Proposition 2

To prove PAM/PAC, we establish two results (in Secion B.4.3 and B.4.4, respectively). First, we show that if there exists a firm type y_m that is present in two submarkets, then $\zeta(y)$ must jump up around type y_m under the assumption $\underline{\rho} \geq (1+\sigma)/2$ (note that $\underline{\rho} > 1/2$ is actually sufficient; see Lemma 9).

Second, we show that if firm types have a unique optimal queue within some interval, then both $\zeta(y)$ and $h(\zeta(y), \lambda(y))$ are increasing in y within this interval when $\underline{\rho} \geq (1+\sigma)/2$.

These two results jointly imply that PAC/PAM holds at the planner's solution. Note that if there exist no firm types with two submarkets, then the second result above implies that PAC/PAM holds. Suppose that there exists a single firm type y_m which has two submarkets where $\zeta(y_m)$ is 0 and $\zeta_1 > 0$ (in Figures 1a and 1b, $y_m = 0.6$ and $\zeta_1 = 0.2$). Then when $y < y_m$ or $y > y_m$, firms of type y have a unique optimal queue. The second result above implies that both $\zeta(y)$ and $h(\zeta(y), \lambda(y))$ are increasing when $y < y_m$ and when $y > y_m$. Recall that $\mathcal{Q}(y)$ is the set of queues that firms of type y face at the planner's solution. Since $\mathcal{Q}(y)$ solves the maximization problem in (10), it is an upper hemi-continuous correspondence by the Theorem of the Maximum. The first result above then implies $\lim_{y \uparrow y_m} \zeta(y) = 0$ and $\lim_{y \downarrow y_m} \zeta(y) = \zeta_1$. Therefore, the resulting optimal queues must look like the

one in Figure 1a. Hence, PAC/PAM holds.

Finally, note that there exists at most one firm type that is present in two submarkets when $\underline{\rho} \geq (1+\sigma)/2$. As before, suppose that firms of type y_m have two submarkets. Then, $\zeta(y_m)$ is 0 and $\zeta_1 > 0$. Then the first result above implies that firms with type y slightly above y_m have a unique submarket whose $\zeta(y)$ is close to ζ_1 and firms with types slightly below y_m have a unique submarket whose $\zeta(y)$ is close to 0. Therefore, firm types that have two submarkets are isolated from each other so that we can list them as $y_m^1 < \cdots < y_m^K$. Assume that $K \geq 2$, and that $\zeta(y_m^i)$ is either 0 or ζ_1^i for $i = 1, \ldots, K$. Then firms of type $y \in (y_m^i, y_m^{i+1})$ have a unique optimal queue, and by the first result above, $\lim_{y \downarrow y_m^i} \zeta(y) = \zeta_1^i$ and $\lim_{y \uparrow y_m^{i+1}} \zeta(y) = 0$, which contradicts with the second result above. Hence there exists at most one firm type that is present in two submarkets.

After presenting two helpful lemmas in Section B.4.1 and B.4.2, we prove the two main results in Section B.4.3 and B.4.4. Finally, we show that the planner's solution is unique in Section B.4.5.

B.4.1 The Elasticity of Complementarity Revisited.

Note that $\rho(x,y)$ is the ratio of the percentage change in $f_y(x,y)$ (the marginal output by firms) and the percentage change in $f(x,y)$ caused by increasing the worker type to $x + \Delta x$. That is, for sufficiently small $\Delta x > 0$, we have

$$\frac{f_y(x+\Delta x, y)}{f_y(x,y)} \approx 1 + \rho(x,y)\frac{f_x(x,y)}{f(x,y)}\Delta x \approx \left(\frac{f(x+\Delta x, y)}{f(x,y)}\right)^{\rho(x,y)}.$$

In general, when x is discrete and $\rho(x,y)$ is not necessarily constant, the elasticity of f_y with respect to f is bounded by $\underline{\rho}$ and $\overline{\rho}$, as summarized by the following lemma.

Lemma 7. *For given y, $f_y(x,y)/f(x,y)^{\underline{\rho}}$ is increasing in x, and $f_y(x,y)/f(x,y)^{\overline{\rho}}$ is decreasing in x. That is,*

$$\left(\frac{f(x_2,y)}{f(x_1,y)}\right)^{\underline{\rho}} \leq \frac{f_y(x_2,y)}{f_y(x_1,y)} \leq \left(\frac{f(x_2,y)}{f(x_1,y)}\right)^{\overline{\rho}}, \tag{38}$$

where the first (resp. second) inequality holds as equality if and only if $\underline{\rho}$ (resp. $\overline{\rho}$) is equal to $\rho(x,y)$ for all $x \in [x_1, x_2]$.

Proof. Given ρ_0, the derivative of $\log f_y(x,y) - \rho_0 \log f(x,y)$ with respect to x equals

$$\frac{\partial}{\partial x}\left(\log f_y - \rho_0 \log f\right) = \frac{f_{xy}}{f_y} - \rho_0 \frac{f_x}{f} = \frac{f_{xy} f - \rho_0 f_x f_y}{f f_y},$$

where we suppress the arguments of $f(x,y)$ and its partial derivatives for simplicity. The right-hand side is weakly positive (resp. negative) if $\rho_0 = \underline{\rho}$ (resp. $\rho_0 = \overline{\rho}$), which means that $\log f_y(x_2,y) - \underline{\rho} \log f(x_2,y) \geq \log f_y(x_1,y) - \underline{\rho} \log f(x_1,y)$, and $\log f_y(x_2,y) - \overline{\rho} \log f(x_2,y) \geq \log f_y(x_1,y) - \overline{\rho} \log f(x_1,y)$, which jointly imply (38). □

B.4.2 A Technical Lemma

The first two parts of the following lemma are trivial, whereas the third part is non-trivial and critical for our results.

Lemma 8. *(i) If $\rho > 1$, then $\frac{1}{\kappa}((1+\kappa)^\rho - 1)$ is strictly increasing for $\kappa > 0$; (ii) if $\rho \in (0,1)$, then $\frac{1}{\kappa}((1+\kappa)^\rho - 1)$ is strictly decreasing for $\kappa > 0$; and (iii) if $\rho \in (0,1)$, then $\left(\frac{1}{\kappa} + \frac{1-\rho}{2}\right)((1+\kappa)^\rho - 1)$ is strictly increasing for $\kappa > 0$.*

Proof. For (i) and (ii), define $g(\kappa) = (1+\kappa)^\rho$, which is strictly concave if $\rho \in (0,1)$ and strictly convex if $\rho > 1$. Observe that $((1+\kappa)^\rho - 1)/\kappa = (g(\kappa) - g(0))/(\kappa - 0)$, which is strictly increasing in κ if $g(\kappa)$ is strictly convex, and strictly decreasing in κ if $g(\kappa)$ is strictly concave.

For (iii), direct computation gives

$$\frac{d}{d\kappa}\left[\left(\frac{1}{\kappa} + \frac{1-\rho}{2}\right)((1+\kappa)^\rho - 1)\right] = \frac{2(1+\kappa)^{1-\rho} - 2 - \kappa(1-\rho)(2-\kappa\rho)}{2\kappa^2(1+\kappa)^{1-\rho}}.$$

The numerator on the right-hand side equals zero for $\kappa = 0$. Moreover, its derivative is $\frac{d}{d\kappa}[2(1+\kappa)^{1-\rho} - 2 - \kappa(1-\rho)(2-\kappa\rho)] = 2(1-\rho)[(1+\kappa)^{-\rho} - (1-\kappa\rho)] > 0$, because convexity of $(1+\kappa)^{-\rho}$ implies $(1+\kappa)^{-\rho} - (1-\kappa\rho) > 0$. Hence, the numerator on the right-hand side is strictly positive for $\kappa > 0$, which proves (iii). □

B.4.3 Local Analysis: Around a Firm Type with Two Submarkets

We now present a lemma which guarantees that the planner's choice is well behaved around a multiplicity point y_m. Note that the sufficient condition for PAC/PAM ($\underline{\rho} \geq (1+\sigma)/2$) is more than we need here ($\underline{\rho} \geq 1/2$) for the first case.

Lemma 9. *Suppose that at the planner's solution, firms of type y_m have two submarkets with queues $(0, \lambda_0)$ and $(\lambda_1\zeta_1, \lambda_1)$ and $\zeta_1 > 0$. If $\underline{\rho} > 1/2$, then there exists a small interval of firm types containing y_m such that within this interval, if $y > y_m$ then firms of type y have a single submarket whose queue is close to $(\lambda_1\zeta_1, \lambda_1)$, and if $y < y_m$ then firms of type y have a single submarket whose queue is close to $(0, \lambda_0)$.*

When $\overline{\rho} \leq (1-\sigma)/2$, then the conclusion is reversed: Within the interval, if $y > y_m$ then firms of type y have a single submarket whose queue is close to $(0, \lambda_0)$, and if $y < y_m$ then firms of type y have a single submarket whose queue is close to $(\lambda_1\zeta_1, \lambda_1)$.

Proof. Suppose that the queues in the two submarkets for firms of type y_m are (ζ_0, λ_0) and $(\lambda_1\zeta_1, \lambda_1)$, where $0 = \zeta_0 < \zeta_1$. Since the marginal contribution to surplus by firms of type y_m must be the same for the two submarkets, by (35) we have

$$m(\lambda_0) - \lambda_0 m'(\lambda_0) = m(\lambda_1) - \lambda_1 m'(\lambda_1) + \left(\phi(\zeta_1\lambda_1, \lambda_1) - \lambda_1 \frac{d\phi(\zeta_1\lambda_1, \lambda_1)}{d\lambda}\right) \frac{\Delta f}{f^1}, \tag{39}$$

where $\Delta f = f(x_2, y_m) - f(x_1, y_m)$ and $f^1 = f(x_1, y_m)$. The left-hand side is the firm's marginal contribution to surplus with a queue $(0, \lambda_0)$, divided by $f(x_1, y_m)$, and the right-hand side is the corresponding value with a queue $(\lambda_1\zeta_1, \lambda_1)$.

If $\zeta_1 \in (0, 1)$, then low-type workers are present in both queues and their marginal contribution to surplus must be the same. Equation (33) then yields

$$m'(\lambda_0) = m'(\lambda_1) + \phi_\lambda(\zeta_1\lambda_1, \lambda_1)\frac{\Delta f}{f^1} \quad \text{if } \zeta_1 \in (0, 1). \tag{40}$$

Low-type workers are not present in the shorter queue if $\zeta_1 = 1$. In this special case, optimality requires that the left-hand side of (40) is larger than the right-hand side.

Recall that $\mathcal{Q}(y)$ is the set of queues that firms of type y face at the planner's solution. By the Theorem of the Maximum, $\mathcal{Q}(y)$ is an upper hemi-continuous correspondence. That is, for firm types y close to y_m, the element(s) in $\mathcal{Q}(y)$ must be close to either $(0, \lambda_0)$ or $(\lambda_1\zeta_1, \lambda_1)$.

By the envelope theorem, if a firm with type y close to y_m is constrained to choose only (μ, λ) close to $(\lambda_1\zeta_1, \lambda_1)$, then its return is approximately (first-order) $\Pi(\zeta_1, \lambda_1, y_m) + \Pi_y(\zeta_1, \lambda_1, y_m)\Delta y$ where $\Delta y = y - y_m$. Similarly, if the firm is constrained to choose $\zeta = 0$, then its maximum expected profit is approximately $\Pi(0, \lambda_0, y_m) + \Pi_y(0, \lambda_0, y_m)\Delta y$. Recall that $\Pi(\zeta_1, \lambda_1, y_m) = \Pi(0, \lambda_0, y_m)$. When $\Pi_y(\zeta_1, \lambda_1, y_m) > \Pi_y(0, \lambda_0, y_m)$, then a firm type $y > y_m$ strictly prefers to choose ζ around ζ_1 instead of around zero, and a firm type $y < y_m$ strictly prefers to choose ζ around zero instead of around ζ_1. As mentioned before, by continuity, it is without loss of generality to constrain the firm to choose between zero and all ζ close to ζ_1.

Note that by the envelope theorem, the condition $\Pi_y(0, \lambda_0, y_m) < \Pi_y(\zeta_1, \lambda_1, y_m)$ can be written as

$$m(\lambda_0) < m(\lambda_1) + \phi(\zeta_1\lambda_1, \lambda_1)\frac{\Delta f_y}{f_y^1}, \tag{41}$$

where $\Delta f_y = f(x_2, y_m) - f(x_1, y_m)$ and $f_y^1 = f_y(x_1, y_m)$. Similarly, $\Pi_y(0, \lambda_0, y_m) > \Pi_y(\zeta_1, \lambda_1, y_m)$ when the reverse inequality holds in (41).

First consider the case in which $\zeta_1 < 1$, such that (40) holds with equality. From (39) and (40), we can solve for $\kappa(y_m)$ and λ_0 in terms of ζ_1 and λ_1. This yields

$$\kappa(y_m) = \frac{4\sigma(1 + \lambda_1 - \lambda_1\sigma(1 - \zeta_1))^2}{(1 + \lambda_1)(\lambda_1 - \sigma - \lambda_1\sigma(1 - \zeta_1) + 1)^2}, \tag{42}$$

$$\lambda_0 = \frac{\lambda_1(\lambda_1 + \sigma(-\lambda_1 + (\lambda_1 + 2)\zeta_1 - 1) + 1)}{1 - \sigma - \lambda_1(1 - \sigma - \sigma\zeta_1)}. \tag{43}$$

Assume $\underline{\rho} > 1/2$. Rewrite (41) as

$$1 + \frac{m(\lambda_0) - m(\lambda_1)}{\phi(\zeta_1\lambda_1, \lambda_1)} < \frac{f_y(x_2, y_m)}{f_y(x_1, y_m)}. \tag{44}$$

Since $\underline{\rho} > 1/2$, $f_y(x_2, y_m)/f_y(x_1, y_m) > (1 + \kappa(y_m))^{1/2}$ by (38). Note that

$$(1 + \kappa(y_m)) - \left(1 + \frac{m(\lambda_0) - m(\lambda_1)}{\phi(\zeta_1\lambda_1, \lambda_1)}\right)^2 = \frac{4\lambda_1\sigma^3(1 - \zeta_1)(1 + \lambda_1(1 - \sigma(1 - \zeta_1)))}{(1 + \lambda_1)^2(1 - \sigma + \lambda_1(1 - \sigma(1 - \zeta_1)))^2} > 0,$$

hence (44) holds.

On the other hand, if $\overline{\rho} \leq (1-\sigma)/2$, then we have

$$\frac{\Delta f_y}{f_y^1} - \frac{m(\lambda_0)-m(\lambda_1)}{\phi(\zeta_1\lambda_1,\lambda_1)} < (1+\kappa(y_m))^{\overline{\rho}} - 1 - \frac{m(\lambda_0)-m(\lambda_1)}{\phi(\zeta_1\lambda_1,\lambda_1)}$$
$$< \frac{1-\sigma}{2}\kappa(y_m) - \frac{m(\lambda_0)-m(\lambda_1)}{\phi(\zeta_1\lambda_1,\lambda_1)} = -\frac{2\sigma^2\lambda_1(1-\sigma(1-\zeta_1))(1+\lambda_1(1-\sigma(1-\zeta_1)))}{(1+\lambda_1)\left(1-\sigma+\lambda_1(1-\sigma(1-\zeta_1))\right)^2} \leq 0,$$

where the first inequality follows from $f_y(x_2,y_m)/f_y(x_1,y_m) < (1+\kappa(y_m))^{\overline{\rho}}$ (see (38)), the second inequality follows from $(1+\kappa)^{\overline{\rho}} < 1+\overline{\rho}\kappa \leq 1+\frac{1-\sigma}{2}\kappa$, and the equality follows from equations (42) and (43). Hence, (41) holds with $>$.

Next, consider the case $\zeta_1 = 1$, where (39) holds with equality and (40) holds with $>$. From (39) we can solve

$$\frac{f(x_2,y_m)}{f(x_1,y_m)} = \kappa(y_m) + 1 = \frac{(\lambda_0/(1+\lambda_0))^2}{(\lambda_1/(1+\lambda_1))^2} \tag{45}$$

The sorting condition (41) becomes $\frac{\lambda_0/(1+\lambda_0)}{\lambda_1/(1+\lambda_1)} < \frac{f_y(x_2,y_m)}{f_y(x_1,y_m)}$, which, by (45), is equivalent to $\sqrt{\frac{f(x_2,y_m)}{f(x_1,y_m)}} < \frac{f_y(x_2,y_m)}{f_y(x_1,y_m)}$. If $\underline{\rho} > 1/2$, then the above inequality holds by Lemma 7; if $\underline{\rho} < 1/2$, then similarly, the above inequality holds with $>$. □

B.4.4 Local Analysis: An Interval of Firm Types That Have Unique Queues and Both Types of Workers

We now consider an interval of firm types that have unique queues ($\mathcal{Q}(y)$ contains a single element) and attract both types of workers ($\zeta(y) \in (0,1)$). The FOCs (13) and (14) jointly determine $\lambda(y)$ and $\zeta(y)$. Differentiating (14) with respect to y yields

$$-\frac{1}{\phi_\mu}\left(\frac{\partial\phi_\mu}{\partial\zeta}\zeta'(y) + \frac{\partial\phi_\mu}{\partial\lambda}\lambda'(y)\right) = \frac{\Delta f_y}{\Delta f}, \tag{46}$$

which states that the percentage decrease in ϕ_μ must equal the percentage increase in Δf.

Similarly, differentiating (13) with respect to y yields

$$\zeta'(y)(W_2 - W_1) = m' f_y^1 + m''\lambda'(y) f^1 + (\zeta(y)\phi_\mu + \phi_\lambda)\Delta f_y \\ + \left[\zeta'(y)\phi_\mu + \zeta\frac{\partial\phi_\mu}{\partial\zeta}\zeta'(y) + \zeta\frac{\partial\phi_\mu}{\partial\lambda}\lambda'(y) + \frac{\partial\phi_\lambda}{\partial\zeta}\zeta'(y) + \frac{\partial\phi_\lambda}{\partial\lambda}\lambda'(y)\right]\Delta f,$$

where we have suppressed the arguments $\mu(y)$ and $\lambda(y)$ from the functions m and ϕ. By (14), we can substitute $\phi_\mu \Delta f$ for $W_2 - W_1$ on the left-hand side. The resulting equation and equation (46) are two linear equations in $\zeta'(y)$ and $\lambda'(y)$. A simple but tedious calculation then yields the percentage change of $m'(\lambda)$ across firm types, i.e.

$$-\frac{m''(\lambda(y))}{m'(\lambda(y))}\lambda'(y) = \frac{f_y^1}{f^1}\frac{1 - \frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_\lambda\right)\frac{\Delta f_y}{f_y^1}}{1 - \frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}} - \phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}}. \tag{47}$$

When the meeting technology exhibits no congestion externalities (i.e. $\sigma = 1$), the second factor on the right-hand side reduces to 1. That is, when we move towards more productive jobs, the percentage decrease in $m'(\lambda)$ (as a result of a longer queue) is independent of ζ and simply equals the percentage increase in $f(x_1, y)$. When there are congestion externalities between heterogeneous workers, however, the optimal queue involves a trade-off between quantity and quality, and more of one affects the marginal contribution of the other. The second factor on the right-hand side of (47) represents this complex interplay between quality and quantity.

Dividing both sides of (46) by the corresponding side of (47) then gives the relative change in ϕ_μ and $m'(\lambda)$ across firm types,

$$\frac{\frac{1}{\phi_\mu}\left(\frac{\partial\phi_\mu}{\partial\zeta}\zeta'(y) + \frac{\partial\phi_\mu}{\partial\lambda}\lambda'(y)\right)}{\frac{m''}{m'}\lambda'(y)} = \frac{f^1\Delta f_y}{f_y^1\Delta f}\frac{1 - \frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}} - \phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}}{1 - \frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_\lambda\right)\frac{\Delta f_y}{f_y^1}}. \tag{48}$$

The left-hand side reflects the relative change in ϕ_μ and $m'(\lambda)$ across firm types. Recall that $a^c(\zeta, \lambda)$, as defined by equation (18), measures the relative change in ϕ_μ and $m'(\lambda)$, while fixing ζ. Thus if the right-hand side of (48) is larger than $a^c(\zeta(y), \lambda(y))$, then it must be the case that $\zeta'(y) \geq 0$. Similarly, if the right-hand

side of (48) is larger than $a^m(\zeta(y), \lambda(y))$, as defined by equation (19), then it must be the case that $\frac{d}{dy}h(\zeta(y), \lambda(y)) \geq 0$. We can summarize this in the following Lemma.

Lemma 10. *Assume that at the planner's solution, there exists an interval of firm types that have unique queues and attract both types of workers ($\zeta(y) \in (0,1)$). If type y is in this interval, then $\zeta'(y) \geq 0$ (resp. $\frac{d}{dy}h(\zeta(y), \lambda(y)) \geq 0$) if and only if*

$$\frac{f^1 \Delta f_y}{f^1_y \Delta f} \geq a^i \frac{1 - \frac{1}{m'}\left(\phi_\mu \frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_\lambda\right)\frac{\Delta f_y}{f^1_y}}{1 - \frac{1}{m''}\left(\frac{\phi^2_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}}, \tag{49}$$

where $i = c$ (resp. $i = m$), and we suppress the arguments of $\phi(\zeta(y)\lambda(y), \lambda(y))$, $m(\lambda(y))$ and $a^i(\zeta(y), \lambda(y))$.

Proof. Rearranging equation (48) gives

$$-\frac{1}{\phi_\mu}\frac{\partial \phi_\mu}{\partial \zeta}\zeta'(y) = \frac{f^1_y}{f^1}\left(\frac{f^1 \Delta f_y}{f^1_y \Delta f} - \frac{\frac{1}{\phi_\mu}\frac{\partial \phi_\mu}{\partial \lambda}}{\frac{m''}{m'}} \frac{1 - \frac{1}{m'}\left(\phi_\mu \frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_\lambda\right)\frac{\Delta f_y}{f^1_y}}{1 - \frac{1}{m''}\left(\frac{\phi^2_{\mu\lambda}}{\phi_{\mu\mu}} - \phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}}\right) \tag{50}$$

where we used equation (47) to substitute out $\lambda'(y)$. Since $\phi(\mu, \lambda)$ is strictly concave, $\frac{\partial \phi_\mu}{\partial \zeta} = \lambda \phi_{\mu\mu} < 0$, which implies that $\zeta'(y) \geq 0$ if and only if the term in the parenthesis on the right-hand side is positive, i.e. (49) holds with $i = c$.

By definition, PAM is equivalent to $\frac{\partial h}{\partial \zeta}\zeta'(y) + \frac{\partial h}{\partial \lambda}\lambda'(y) \geq 0$. Combining (47) and (50) then shows that PAM is obtained if and only if (49) holds with $i = m$. □

We now show that the necessary condition (20) implies that PAC/PAM holds locally at all interior points, so it is also sufficient. The same conclusion also applies to the case of NAC/NAM.

Recall $\kappa(y) \equiv \Delta f / f^1$. Throughout we will then use the following inequalities which result from rewriting (38):

$$\frac{(1 + \kappa(y))^{\underline{\rho}} - 1}{\kappa(y)} \leq \frac{f^1 \Delta f_y}{f^1_y \Delta f} \leq \frac{(1 + \kappa(y))^{\overline{\rho}} - 1}{\kappa(y)}. \tag{51}$$

First, consider PAC/PAM. Assume the necessary condition (20) holds, i.e. $\underline{\rho} \geq \overline{a}^i$. Since $\overline{a}^i \geq 0$, this implies that $\Delta f_y \geq 0$ (i.e. f is supermodular) such that the

left-hand side of (49) is positive. We now prove a stronger version of (49), i.e.

$$\frac{f^1\Delta f_y}{f_y^1\Delta f} \geq \overline{a}^i \frac{1-\frac{1}{m'}\left(\phi_\mu \frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)\frac{\Delta f_y}{f_y^1}}{1-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}},$$

where $a^i(\zeta,\lambda)$ is replaced by its supremum $\overline{a}^i$. This is justified because if the second factor on the right-hand side is negative then we have nothing to prove; if it is positive, then we have a stronger version of the original inequality. Firms' SOC implies that the denominator of this factor is positive. Rearranging terms therefore gives

$$\frac{f^1\Delta f_y}{f_y^1\Delta f}+\frac{\Delta f_y}{f_y^1}\left[\overline{a}^i\frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)\right]\geq \overline{a}^i. \tag{52}$$

Since $\phi(\mu,\lambda)$ is given by (2) and $\overline{a}^i=(1+\sigma)/2$ by Lemma 4, the above condition can be rewritten as

$$\frac{f^1\Delta f_y}{f_y^1\Delta f}+\frac{\Delta f_y}{f_y^1}\frac{(1-\sigma)(2+(1-\sigma)\lambda)(1+\lambda)^2}{4(1+\lambda(1-\sigma))(1+\lambda(1-\sigma)+\lambda\sigma\zeta)}\geq\frac{1+\sigma}{2}.$$

Consider now two subcases, determined by the value of $\underline{\rho}$. If $\underline{\rho}\geq 1$, then the first term on the left-hand side is greater than 1 by (51); hence the above condition holds. Next, consider the case $\underline{\rho}\in(0,1)$. Note that

$$\frac{(1-\sigma)(2+(1-\sigma)\lambda)(1+\lambda)^2}{4(1+\lambda(1-\sigma))(1+\lambda(1-\sigma)+\lambda\sigma\zeta)}\geq\frac{(1-\sigma)(2+(1-\sigma)\lambda)(1+\lambda)}{4(1+\lambda(1-\sigma))}\geq\frac{1-\sigma}{2}$$

where the first inequality is because the denominator reaches its maximum at $\zeta=1$, and the second one is because $1+\lambda\geq 1+(1-\sigma)\lambda$. Thus a sufficient condition for (52) is

$$\frac{f^1\Delta f_y}{f_y^1\Delta f}+\frac{\Delta f_y}{f_y^1}\frac{1-\sigma}{2}\geq\frac{1+\sigma}{2}.$$

Note that

$$\frac{f^1\Delta f_y}{f_y^1\Delta f}+\frac{\Delta f_y}{f_y^1}\frac{1-\sigma}{2}\geq\frac{(1+\kappa(y))^{\underline{\rho}}-1}{\kappa(y)}+\left((1+\kappa(y))^{\underline{\rho}}-1\right)(1-\underline{\rho})\geq\underline{\rho}\geq\frac{1+\sigma}{2},$$

where the first inequality holds by (51) and the assumption $\underline{\rho}\geq(1+\sigma)/2$, the second inequality holds because the second term reaches its minimum value $\underline{\rho}$ at $\kappa(y)=0$, by part (iii) of Lemma 8. Therefore, (52) holds when $\underline{\rho}\geq\overline{a}^i$.

Next, consider NAC/NAM. If $\Delta f_y\leq 0$, then the left-hand side of (49) is negative. The denominator on the right-hand side is positive because of the SOC, and the numerator is positive because

$$\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda=\frac{1-\sigma}{2\sigma(1+\lambda(1-\sigma)+\lambda\sigma\zeta)}\geq 0.$$

Thus, it follows immediately that (49) holds with $\leq$.

In contrast, if $\Delta f_y\geq 0$, then we have

$$\begin{aligned}\frac{f^1\Delta f_y}{f_y^1\Delta f}\frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)&\leq\frac{1-\sigma}{2}\frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)\\&=-\frac{\lambda(1+\lambda)^2(1-\sigma)^2}{4(1+(1-\sigma)\lambda)(1+\sigma\mu+(1-\sigma)\lambda)}\leq 0.\end{aligned}$$

which then implies

$$1\leq\frac{1-\frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)\frac{\Delta f_y}{f_y^1}}{1-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}}.$$

Therefore, we have

$$\frac{f^1\Delta f_y}{f_y^1\Delta f}\leq\overline{\rho}\leq\underline{a}^i\leq a^i\leq a^i\frac{1-\frac{1}{m'}\left(\phi_\mu\frac{\phi_{\mu\lambda}}{\phi_{\mu\mu}}-\phi_\lambda\right)\frac{\Delta f_y}{f_y^1}}{1-\frac{1}{m''}\left(\frac{\phi_{\mu\lambda}^2}{\phi_{\mu\mu}}-\phi_{\lambda\lambda}\right)\frac{\Delta f}{f^1}},$$

where the three inequalities follow from (51), part ii) of Lemma 8, and our assumption $\overline{\rho}\leq\underline{a}^i$, respectively, and the last inequality follows from the result above. Hence, we have proved the case of NAC/NAM.

B.4.5 Uniqueness of the Planner's Solution

Suppose that the solution to the planner's problem is not unique: there exist two allocations $(\overline{\mu}(y), \overline{\lambda}(y))$ and $(\widetilde{\mu}(y), \widetilde{\lambda}(y))$ that solve (7). Consider a new allocation which has queue schedule $(\gamma\overline{\mu}(y) + (1-\gamma)\widetilde{\mu}(y), \gamma\overline{\lambda}(y) + (1-\gamma)\widetilde{\lambda}(y))$ for some $\gamma \in (0,1)$, which must yield the same maximum surplus as the original two allocations. Hence for each firm type y, we have $\gamma\widehat{S}(\overline{\mu}(y), \overline{\lambda}(y), y) + (1-\gamma)\widehat{S}(\widetilde{\mu}(y), \widetilde{\lambda}(y), y) = \widehat{S}(\gamma\overline{\mu}(y) + (1-\gamma)\widetilde{\mu}(y), \gamma\overline{\lambda}(y) + (1-\gamma)\widetilde{\lambda}(y), y)$.

Since the two allocations $(\overline{\mu}(y), \overline{\lambda}(y))$ and $(\widetilde{\mu}(y), \widetilde{\lambda}(y))$ are different, there exist at least two firm types y_1 and y_2 such that $(\overline{\mu}(y), \overline{\lambda}(y)) \neq (\widetilde{\mu}(y), \widetilde{\lambda}(y))$. Consider firms of type y_1. Recall that $\widehat{S}(\mu, \lambda, y_1)$ is linear in (μ, λ) on the line segment between $(\overline{\mu}(y_1), \overline{\lambda}(y_1))$ and $(\widetilde{\mu}(y_1), \widetilde{\lambda}(y_1))$. Given the average queue lengths $(\overline{\mu}(y), \overline{\lambda}(y))$ and $(\widetilde{\mu}(y), \widetilde{\lambda}(y))$, the planner must create two submarkets $(0, \lambda_a(y_1))$ and $(\mu_b(y_1), \lambda_b(y_1))$ in either case. The same is true for firms of type y_2. Therefore, in each of the two allocations $(\overline{\mu}(y), \overline{\lambda}(y))$ and $(\widetilde{\mu}(y), \widetilde{\lambda}(y))$, there are two firm types each of which has two submarkets, which contradicts with PAC/PAM. We have thus proved that the planner's solution must be unique.

B.5 Proof of Proposition 3

When $\sigma = 1$, $\phi(\mu, \lambda)$ is independent of λ: $\phi_\lambda(\mu, \lambda) = 0$; hence $\phi(\mu, \lambda) = m(\mu)$. Therefore, $S(\mu, \lambda, y)$ in (3) reduces to $m(\lambda) f(x_1, y) + m(\mu)[f(x_2, y) - f(x_1, y)]$, which is strictly concave in (μ, λ). Thus, $\widehat{S}(\mu, \lambda, y) = S(\mu, \lambda, y)$, and the planner's problem in (7) is strictly concave, which implies a unique optimal solution $(\mu(y), \lambda(y))$ that is continuous in y, and is determined by the FOCs (13) and (14) and the complementary slackness conditions. Given that the surplus function is separable in μ and λ (see (53) and (54)), below we derive the FOCs with respect to μ and λ, which are equivalent but simpler than the corresponding version with $\zeta = \mu/\lambda$ and λ given by (13) and (14).

Our proof below consists of three steps: 1) we assume that no firms attract x_2 workers only and show that this assumption is valid if and only if the fraction of x_2 workers z is smaller than some threshold $\widehat{z}$. Furthermore, we derive some characterizations of the planner's solution under this assumption. 2) We show that for PAC/PAM to occur, this assumption is necessary when $\rho \in (0,1)$. 3) We derive the conditions for PAC/PAM, and by utilizing the characterizations derived in step

1, show that they hold if and only if z is sufficiently small.

Step 1: Assume that at the planner's solution, there exist no firms that attract x_2 workers only: if $\lambda(y) > 0$, then $\mu(y) < \lambda(y)$). Then the FOC with respect to λ is given by,

$$m'(\lambda(y))f(x_1, y) = W_1, \tag{53}$$

where W_1 is determined by the budget constraint: $\int_{\underline{y}}^{\overline{y}} \lambda(y) = L$. As long as the above assumption holds, then $\lambda(y)$ and W_1 are independent of z, since the FOC (53) and the corresponding budget constraint do not depend on z.

If $\mu(y) > 0$, then the FOC with respect to μ is

$$m'(\mu(y))\left[f(x_2, y) - f(x_1, y)\right] = W_2 - W_1. \tag{54}$$

where $W_2 - W_1$ and hence W_2 are determined by the budget constraint: $\int_{\underline{y}}^{\overline{y}} \mu(y) = Lz$. Therefore, for a given y if $f(x_2, y) - f(x_1, y) > W_2 - W_1$, then $\mu(y) > 0$ and is strictly decreasing in $W_2 - W_1$. Thus, $W_2 - W_1$ is strictly decreasing in z for a given λ.

Given $\lambda(y)$ and W_1, as long as $W_2 - W_1 > \max_{y \in [\underline{y}, \overline{y}]} m'(\lambda(y))[f(x_2, y) - f(x_1, y)]$, where the right-hand side is (54) evaluated at $\mu(y) = \lambda(y)$ (the knife-edge case), then no firms will attract only x_2 workers. Since $W_2 - W_1$ is strictly decreasing in z, there exists a threshold $\widehat{z}$ such that the above assumption holds if and only if $z < \widehat{z}$.

Step 2: Suppose that the above assumption fails and there exists some firm type y_1 with $0 < \mu(y_1) = \lambda(y_1)$. The FOCs for firms of type y_1 are: $m'(\mu(y_1))f(x_2, y_1) = W_2$ and $m'(\mu(y_1))f(x_1, y_1) \leq W_1$, which implies that

$$\frac{f(x_2, y_1) - f(x_1, y_1)}{f(x_1, y_1)} \geq \frac{W_2 - W_1}{W_1}$$

But, since $(\mu(y), \lambda(y))$ is continuous in y, there must exist some firm type y_2 with $0 < \mu(y_2) < \lambda(y_2)$. For firms of type y_2, both (53) and (54) must hold, which

implies that

$$\frac{W_2 - W_1}{W_1} = \frac{m'(\mu(y_2))}{m'(\lambda(y_2))} \frac{f(x_2, y_2) - f(x_1, y_2)}{f(x_1, y_2)} > \frac{f(x_2, y_2) - f(x_1, y_2)}{f(x_1, y_2)}$$

Combining the above two equations implies that $f(x_2, y_1)/f(x_1, y_1) > f(x_2, y_2)/f(x_1, y_2)$. Since we assume $\rho < 1$, $f(x, y)$ is strictly log-submodular: $f(x_2, y)/f(x_1, y)$ is strictly decreasing in y. Thus $y_1 < y_2$ and PAC/PAM fails at the planner's solution.

Step 3: Assume $z < \widehat{z}$ or equivalently that there exist no firms that attract x_2 workers only. By differentiating (53) and (54) with respect to y (or equivalently equation (49) in Lemma 10 in Appendix B.4), PAC/PAM holds at the planner's solution if and only if for each y,

$$\frac{(1 + \kappa(y))^\rho - 1}{\kappa(y)} \geq a^i(\zeta(y), \lambda(y)) \tag{55}$$

where, as before, $i = c$ for the case of PAC and $i = m$ for the case of PAM. Note that when $\sigma = 1$, $a^m(\zeta, \lambda) = \zeta$ and $a^c(\zeta, \lambda) = \zeta(1 + \lambda)/(1 + \zeta\lambda) > \zeta$. Since $\rho < 1$, the left-hand side above is strictly decreasing in $\kappa(y)$ and at $\kappa(y) = 0$, it equals ρ. Thus, (55) implies that when PAC/PAM holds, $\rho > \zeta(y)$ for all y.

Recall that when $z < \widehat{z}$, both $\lambda(y)$ and W_1 are independent of z. For $i = c$ and m, define $\overline{\zeta}^i(y)$ as the value of $\zeta(y)$ such that (55) holds with equality. Since both $a^c(\zeta, \lambda)$ and $a^m(\zeta, \lambda)$ are decreasing in ζ, PAC/PAM holds if and only if for each y, $\zeta(y) \leq \overline{\zeta}^i(y)$. As before, as long as $W_2 - W_1 \geq \max_{y \in [\underline{y}, \overline{y}]} m'(\lambda(y)\overline{\zeta}^i(y))[f(x_2, y) - f(x_1, y)]$ (the knife-edge case), then $\zeta(y) \leq \overline{\zeta}^i(y)$ for all y and PAC/PAM holds. Thus following the same logic as before, there exists a threshold $\overline{z}^i$ such that $\zeta(y) \leq \overline{\zeta}^i(y)$ for all y if and only if $z \leq \overline{z}^i$. Since $a^c(\zeta, \lambda) > a^m(\zeta, \lambda)$, $\overline{\zeta}^m(y) > \overline{\zeta}^c(y)$ and thus $\overline{z}^m > \overline{z}^c$. □

B.6 Proof of Lemma 5

Given U_1/w_1 and U_2/w_2, consider then the level curves $\psi_2(\lambda\zeta, \lambda) = U_2/w_2$ and $\psi_1(\lambda\zeta, \lambda) = U_1/w_1$ in the λ-ζ space. Note that

$$\psi_1(\lambda\zeta, \lambda) = \frac{1 + (1 - \sigma)\lambda}{(1 + \lambda)(1 + (1 - \sigma + \sigma\zeta)\lambda)} \text{ and } \psi_2(\lambda\zeta, \lambda) = \frac{1}{1 + (1 - \sigma + \sigma\zeta)\lambda},$$

both of which are strictly decreasing in ζ. We now show that the two curves intersect at most once so that there exists exactly one solution (μ, λ). At any intersection point, the difference between the slopes of the two level curves is

$$-\frac{\partial\psi_1(\lambda\zeta,\lambda)/\partial\lambda}{\partial\psi_1(\lambda\zeta,\lambda)/\partial\zeta}+\frac{\partial\psi_2(\lambda\zeta,\lambda)/\partial\lambda}{\partial\psi_2(\lambda\zeta,\lambda)/\partial\zeta}=\frac{1+(1-\sigma+\sigma\zeta)\lambda}{\lambda(\lambda+1)(1+(1-\sigma)\lambda)}>0.$$

Hence, by a standard single-crossing argument, the two level curves cross each other at most once. Note that we can also derive the solution (μ, λ) explicitly. However, with this approach we need to discuss the conditions under which we have a corner solution ($\mu = 0$ or $\mu = \lambda$) or an interior solution ($0 < \mu < \lambda$). □

B.7 Proof of Lemma 6

We first show that given a solution (μ^*, λ^*) (interior or corner) to the firm's problem (29), the corresponding wage menu $(w_1^*, w_2^*) = (U_1/\psi_1(\mu^*, \lambda^*), U_2/\psi_2(\mu^*, \lambda^*))$ satisfies (26). This proof is based on Shimer (2005), but extends his result to arbitrary $\phi(\mu, \lambda)$. Because $\phi(\mu, \lambda)$ is concave in μ, we have

$$\psi_1\left(\mu^*,\lambda^*\right)\leq\phi_\mu(\mu^*,\lambda^*)\leq\psi_2\left(\mu^*,\lambda^*\right), \tag{56}$$

where ψ_1 and ψ_2 are defined by equation (28). Consequently, the wages must satisfy

$$w_1^*=\frac{U_1}{\psi_1\left(\mu^*,\lambda^*\right)}\geq\frac{U_1}{\phi_\mu(\mu^*,\lambda^*)}\quad\text{and}\quad w_2^*=\frac{U_2}{\psi_2\left(\mu^*,\lambda^*\right)}\leq\frac{U_2}{\phi_\mu(\mu^*,\lambda^*)}. \tag{57}$$

Moreover, the FOC of (29) with respect to μ implies $\phi_\mu(\mu^*, \lambda^*)(f(x_2, y)-f(x_1, y)) = U_2 - U_1$. Combining this FOC with (57) implies $w_2^* - w_1^* \leq \frac{U_2-U_1}{\phi_\mu(\mu^*,\lambda^*)} = f(x_2, y) - f(x_1, y)$. The strict inequality in $f(x_2, y) - w_2^* > f(x_1, y) - w_1^*$ then follows because the two inequalities in (56) cannot hold simultaneously; that would imply that $\phi(\mu, \lambda^*)$ is linear for $\mu \in [0, \lambda^*]$, in which case the firm's problem never has an interior solution.

Next, we show that posting a wage menu that violates (26) is always strictly suboptimal. Suppose that low-type workers are strictly preferred. The firms' expected profit in this case is

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

where $\phi(\lambda-\mu,\lambda)$ is the probability that firms interview at least one low-type worker. The matching probabilities in (28) become $\psi_1(\mu,\lambda) = \frac{\phi(\lambda-\mu,\lambda)}{\lambda-\mu}$ and $\psi_2(\mu,\lambda) = (m(\lambda) - \phi(\lambda-\mu,\lambda))/\mu$. The firms' expected profit can then be rewritten as $m(\lambda) f^1 + (m(\lambda) - \phi(\lambda-\mu,\lambda))\Delta f - \lambda U_1 - \mu(U_2 - U_1)$. Note that the expected costs are the same as the case where high-type workers are preferred; both equal $\lambda U_1 + \mu(U_2 - U_1)$. However, surplus is strictly smaller than that in (29). The case where firms randomize between low-type and high-type workers follows the same logic. □

B.8 Proof of Proposition 5

First, we consider the unconditional probability that an applicant generates a positive signal $\widetilde{x}_2$. The probability of this event equals $\mathbb{P}(\widetilde{x}_2) = \frac{\mu}{\lambda} + \frac{\lambda-\mu}{\lambda}(1-\tau)$, and the queue length of such applicants is $\widetilde{\lambda} = \lambda\mathbb{P}(\widetilde{x}_2) = \mu + (\lambda-\mu)(1-\tau)$. Given a positive signal $(\widetilde{x}_2)$, the probability that an applicant is of high type (x_2) is $\mathbb{P}(x_2 \,|\, \widetilde{x}_2) = \mathbb{P}(x_2)\mathbb{P}(\widetilde{x}_2 \,|\, x_2)/\mathbb{P}(\widetilde{x}_2) = \mu/\widetilde{\lambda}$, where the first equality is simply Bayes' rule.

Next, we consider the probability that the firm interviews at least one high-type worker, $\phi(\mu,\lambda)$. For this, we can ignore the existence of applicants with negative signals; they are low-type workers for sure and do not affect the meeting process between firms and workers with positive signals. By equation (2), the probability that a firm interviews someone from the queue μ of high-type applicants, given a queue $\widetilde{\lambda}$ of applicants with positive signals, is $\phi(\mu,\lambda) = \mu/(1+\sigma\mu+(1-\sigma)\widetilde{\lambda})$, which yields the desired result after substitution of $\widetilde{\lambda}$. □

B.9 Endogenous Screening

B.9.1 Individual Firm's Problem

Consider an environment which is like our benchmark model, except that firms additionally choose (and post) their recruiting intensity $\sigma \in [0,1]$ at a linear cost

$c\sigma$, where $c \geq 0$.[32] That is, they solve

$$\max_{\sigma,\mu,\lambda} \quad \frac{\lambda}{1+\lambda} f^1 + \frac{\mu}{1+\sigma\mu+(1-\sigma)\lambda}\Delta f - \lambda U_1 - \mu \Delta U - c\sigma. \tag{58}$$

Since the second term above is convex in σ and $c\sigma$ is linear, the above profit function is *convex* in σ. The maximum is therefore reached at a corner, i.e. when $\sigma = 0$ or 1. To determine firms' choice, we compare the profits from the two options.

Profits with No Screening. Consider a firm of type y choosing $\sigma = 0$. This firm's optimal queue then consists of either low-type workers or high-type workers, but not both. Suppose the firm attracts workers of type x_i. Equation (58) then reduces to $\max_{\lambda_i} m(\lambda_i) f(x_i, y) - \lambda_i U_i$. Because $m(\lambda)$ is strictly concave, the FOC of this problem is both necessary and sufficient. Assuming that $f(x_i, y) > U_i$, the optimal queue length is $\lambda_i = \sqrt{f(x_i, y)/U_i} - 1$, which yields an expected payoff of

$$\pi_i(y) = \left(\sqrt{f(x_i, y)} - \sqrt{U_i}\right)^2. \tag{59}$$

Naturally, the firm chooses the type of workers it wishes to attract based on whether $\pi_1(y)$ or $\pi_2(y)$ is higher, which requires comparing $\sqrt{f(x_2, y)} - \sqrt{f(x_1, y)}$ with $\sqrt{U_2} - \sqrt{U_1}$. If the former is strictly increasing in y, i.e. f is strictly square-root supermodular, then there exists a unique y^{EK} such that $\pi_2(y) > \pi_1(y)$ if $y > y^{EK}$ and vice versa. This result is a special case of Eeckhout and Kircher (2010).

Profits with Perfect Screening. When the firm chooses $\sigma = 1$, (58) reduces to

$$\overline{\pi}(y) \equiv \max_{0 \leq \mu \leq \lambda} \frac{\lambda}{1+\lambda} f^1 + \frac{\mu}{1+\mu}\Delta f - \lambda U_1 - \mu \Delta U. \tag{60}$$

This problem is strictly concave in (μ, λ), so that the FOCs are both necessary and sufficient. The only complexity lies in the constraint $0 \leq \mu \leq \lambda$, which, as we illustrate in Figure 4, implies that there are four possibilities with respect to the

[32] Posting contracts that include σ in addition to wages is necessary for constrained efficiency in this environment. More restrictive contract spaces and more general cost functions are left for future research. Wolthoff (2018) endogenizes σ in a similar way as us, but with a cost function that is sufficiently convex (in an otherwise quite different model). In the random search model of Birinci et al. (2020), firms have the option to learn all their applicants' types after paying a fixed cost.

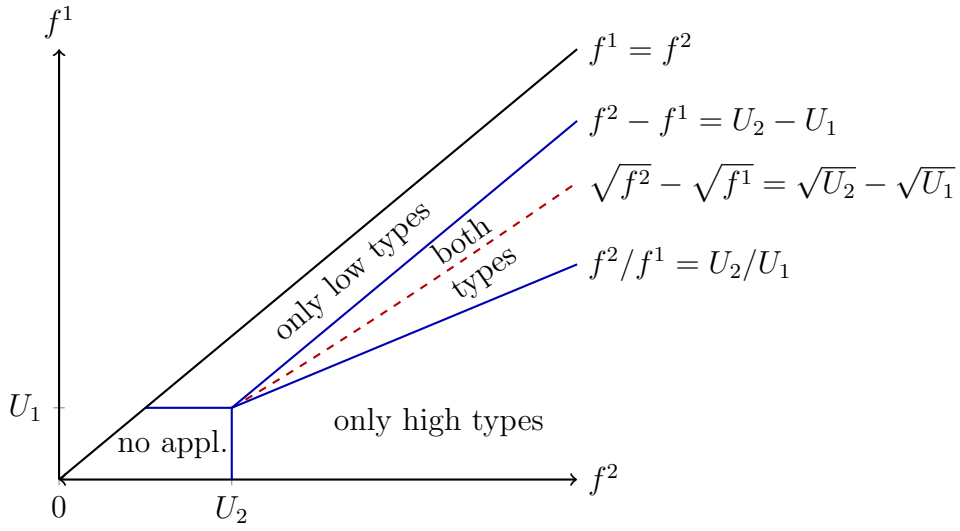


Figure 4: Optimal applicant pool for a firm, conditional on $\sigma = 1$.

optimal applicant pool:

(i) *No applicants.* If $f(x_1, y) \leq U_1$ and $f(x_2, y) \leq U_2$, then the firm will not attract any applicants, such that $\overline{\pi}(y) = 0$.

(ii) *Only low-type applicants.* If $f(x_1, y) > U_1$ and $f(x_2, y) - f(x_1, y) \leq U_2 - U_1$, the firm will attract low-type workers, but not high-type workers as their marginal product is less than their marginal cost; in this case, $\overline{\pi}(y) = \pi_1(y)$.

(iii) *Only high-type applicants.* If $f(x_2, y) > U_2$ and $f(x_2, y)/f(x_1, y) \geq U_2/U_1$, the firm will attract only high-type workers since their relative productivity is higher than their relative cost; in this case, $\overline{\pi}(y) = \pi_2(y)$.

(iv) *Both types of applicants.* If $f(x_2, y) - f(x_1, y) > U_2 - U_1$ and $f(x_2, y)/f(x_1, y) < U_2/U_1$, then the firm strictly prefers a mix of both types of workers in their application pool. By the FOCs, the optimal queue is given by $\mu = \sqrt{\Delta f/\Delta U} - 1$ and $\lambda = \sqrt{f^1/U_1} - 1$. In this case, $\overline{\pi}(y)$ is given by

$$\overline{\pi}(y) = \left(\sqrt{f^1} - \sqrt{U_1}\right)^2 + \left(\sqrt{\Delta f} - \sqrt{\Delta U}\right)^2. \tag{61}$$

Clearly, a necessary condition for $\sigma = 1$ to yield higher profits than $\sigma = 0$ is that the firm attracts both types of applicants. In what follows, we will therefore

focus on this case, which occurs when

$$\Delta f > \Delta U \quad \text{and} \quad \frac{f(x_2, y)}{f(x_1, y)} < \frac{U_2}{U_1}. \tag{62}$$

As the red dashed line in Figure 4 shows, the region described by (62) is divided into two parts by the curve $\pi_1(y) = \pi_2(y)$, or equivalently

$$\sqrt{f^2} - \sqrt{f^1} = \sqrt{U_2} - \sqrt{U_1}. \tag{63}$$

We therefore have to distinguish between two cases when calculating the difference in profits between $\sigma = 0$ and $\sigma = 1$ in this region, i.e. $\Delta\pi(y) \equiv \overline{\pi}(y) - \max\{\pi_1(y), \pi_2(y)\}$. The following lemma formalizes this.

Lemma 11. *If a firm is indifferent between attracting low- and high-type workers conditional on $\sigma = 0$, i.e. $\pi_1(y) = \pi_2(y)$ or equivalently (63) holds, then this firm attracts both types of workers conditional on $\sigma = 1$, i.e. (62) also holds. In the region characterized by (62), the difference in profits between $\sigma = 1$ and $\sigma = 0$ equals*

$$\Delta\pi(y) = \begin{cases} \left(\sqrt{\Delta f} - \sqrt{\Delta U}\right)^2 & \text{if } \pi_1(y) \geq \pi_2(y), \quad (64a) \\ 2\left(\sqrt{f^2 U_2} - \sqrt{f^1 U_1} - \sqrt{\Delta f \Delta U}\right) & \text{if } \pi_1(y) \leq \pi_2(y). \quad (64b) \end{cases}$$

Proof. Equation (63) can be rewritten as $\sqrt{f^2/f^1} - 1 = \sqrt{U_1/f^1}(\sqrt{U_2/U_1} - 1)$. Since $U_1/f^1 < 1$, it follows that $\sqrt{U_2/U_1} - 1 > \sqrt{f^2/f^1} - 1$, and thus $U_2/U_1 > f^2/f^1$. Similarly, (63) can also be rewritten as $(f^2 - f^1)/(\sqrt{f^2} + \sqrt{f^1}) = (U_2 - U_1)/(\sqrt{U_2} + \sqrt{U_1})$. Because $f^1 > U_1$ and $f^2 > U_2$, we have $\Delta f > \Delta U$. Hence, (62) holds. Equation (64) then follows from substituting the relevant version of (59) into $\Delta\pi(y) = \overline{\pi}(y) - \max\{\pi_1(y), \pi_2(y)\}$. □

Choice of Screening Intensity. The characterization of $\Delta\pi(y)$ completes the analysis of the firm's choice problem given by (58): the firm's optimal σ is 1 if $\Delta\pi(y) > c$, 0 if $\Delta\pi(y) < c$, and indeterminate in the knife-edge case $\Delta\pi(y) = c$. If the optimal σ is 1, then the optimal (μ, λ) must be interior, and given by $\mu = \sqrt{\Delta f/\Delta U} - 1$ and $\lambda = \sqrt{f^1/U_1} - 1$. When the optimal σ is 0, then the firm

will attract either only low-type or only high-type workers, depending on whether $\sqrt{f^2} - \sqrt{f^1}$ is larger than $\sqrt{U_2} - \sqrt{U_1}$, as discussed after (59).

B.9.2 Discussion of Proposition 6

Before proving Proposition 6, we first offer some discussion of the results. Consider first the special case $c = 0$, where all firms choose $\sigma = 1$: Given that $c = 0$, the necessary and sufficient condition for PAC/PAM (resp. NAC/NAM) is that $f(x, y)$ needs to be log-supermodular (resp. submodular). Of course, the question remains whether the above conditions are sufficient for any screening cost c. For NAC/NAM, the answer is (almost) true: we find that *strict* submodularity is sufficient for NAC/NAM for any distribution of agents' types and any screening cost c.

However, a sufficient condition for PAC/PAM for any distribution of agents' types and any screening cost c does not exist: For any log-supermodular $f(x, y)$, we can find counterexamples where PAC/PAM fails in equilibrium. The sufficient condition (30) in Proposition 6 for PAC/PAM is for a given distribution of agent types so that $\kappa(\underline{y})$, the lower bound of the output dispersion parameter, is fixed.[33] It requires that either production complementarity measured by $\underline{\rho}$, the lower bound of the production complementarities, or output dispersion measured by $\kappa(\underline{y})$ is sufficiently large. Note that condition (30) is quite sharp: in the proof of Proposition 6, we show that with CES production we can construct counterexamples where PAC/PAM fails in equilibrium whenever $\rho < \Omega(\kappa(\underline{y}))$.

At first, our result regarding PAC/PAM may seem puzzling. One may have expected that, with strong complementarities, firms' incentives to invest in (ex post) screening are increasing in their productivity and that since the least-productive firms can only afford to attract low-type workers, PAC/PAM arises. This intuition turns out to be wrong. When x_1 and x_2 are sufficiently close, the most-productive firms find it optimal to attract high-type applicants only. They are not willing to provide low-type workers with their market utility, because compensating them for the low matching probability that results from the presence of many high-type workers requires a very high wage. Therefore, firms in the middle of the productivity distribution have the strongest incentives to screen ex post. Although those firms also prefer to hire high-type workers, they can only afford to offer modest wages

[33] Since f is assumed to be log-supermodular, $\kappa(y)$ is smallest at $y = \underline{y}$.

to them and therefore they attract relatively few of them. As a consequence, they can attract low-type workers for a relatively low wage (since they offer them a high hiring probability). However, some firm types below those screening firms are not productive enough to be willing to pay the screening cost as an insurance device (since the opportunity costs of remaining unmatched is lower for those firms), but conditional on not screening ex-post, they are productive enough to target high-type applicants. When this happens, PAC/PAM fails in the middle. In the limit where $x_1 \to x_2$, we can always find a distribution of agent types and a screening cost c such that PAC/PAM fails in equilibrium, even for log-supermodular production functions.

This scenario does not arise when either the degree of complementarity $\underline{\rho}$ or output dispersion $\kappa(\underline{y})$ is large, i.e. (30) holds. In that case, the incentive to attract low-type workers as insurance against failing to hire is decreasing in firms' type. Then, the most-productive firms attract only high-type workers, firms in the middle attract both types and screen ex post, and the least-productive firms attract only low-type workers. More precisely, the gains from ex-post screening are first increasing in y, reach their maximum at y^{EK} (defined by $\pi_1(y^{EK}) = \pi_2(y^{EK})$) and from then onwards are decreasing in y.

B.9.3 Proof of Proposition 6

The Analysis of NAC/NAM. As mentioned in the main text, necessity of submodularity of $f(x, y)$ for NAC/NAM follows from the special case $c = 0$ (see Proposition 2). Next, we show that strict submodularity of $f(x, y)$ is sufficient for NAC/NAM. From the discussion after equation (59), it follows that when $f(x, y)$ is strictly submodular, and thus strictly square-root submodular, there exists a unique y^{EK} which solves (63). Furthermore, $\pi_2(y) > \pi_1(y)$ for firms with $y < y^{EK}$, and vice versa.

Since f is strictly submodular, both $f^2 - f^1$ and f^2/f^1 are strictly decreasing in y. The first part of Lemma 11 states that y^{EK} must belong to the region characterized by (62). There exists at most one $y' < y^{EK}$ such that $f(x_2, y')/f(x_1, y') = U_2/U_1$ (otherwise set $y' = \underline{y}$), and at most one $y'' > y^{EK}$ such that $f(x_2, y'') - f(x_1, y'') = U_2 - U_1$ (otherwise set $y'' = \overline{y}$). The region characterized by (62) is thus $y \in (y', y'')$. The following Lemma establishes that $\Delta\pi(y)$ is single-peaked at $y = y^{EK}$.

Lemma 12. *Suppose that $f(x,y)$ is strictly submodular. In the region characterized by (62), $\Delta\pi(y)$ is strictly increasing in y for $y \leq y^{EK}$ and strictly decreasing in y for $y \geq y^{EK}$.*

Proof. For submodular f, $\pi_2(y) > \pi_1(y)$ if $y < y^{EK}$, and vice versa. As we remarked before, the region characterized by (62) is (y', y''), which contains y^{EK}. Hence,

$$\Delta\pi'(y) = \begin{cases} \left(1 - \dfrac{\sqrt{\Delta U}}{\sqrt{\Delta f}}\right) \Delta f_y & \text{if } y > y^{EK}, \\ -\left(\sqrt{\dfrac{\Delta U}{\Delta f}} - \sqrt{\dfrac{U_2}{f^2}}\right) f_y^2 + \left(\sqrt{\dfrac{\Delta U}{\Delta f}} - \sqrt{\dfrac{U_1}{f^1}}\right) f_y^1 & \text{if } y < y^{EK}. \end{cases} \tag{65a}$$

(65b)

To establish the sign of (65a), note that $\Delta f_y = f_y^2 - f_y^1 < 0$ when f is strictly submodular; hence, $\Delta\pi'(y) < 0$ for $y > y^{EK}$. To establish the sign of (65b), note that $f^2/f^1 < U_2/U_1$ is equivalent to $\Delta U/\Delta f > U_1/f^1$ or $\Delta U/\Delta f > U_2/f^2$. The coefficient of f_y^2 in (65b) is therefore negative. Since f is submodular, $f_y^2 \leq f_y^1$, and we have

$$\Delta\pi'(y) \geq -f_y^1 \left(\sqrt{\frac{\Delta U}{\Delta f}} - \sqrt{\frac{U_2}{f^2}}\right) + f_y^1 \left(\sqrt{\frac{\Delta U}{\Delta f}} - \sqrt{\frac{U_1}{f^1}}\right) = f_y^1 (\sqrt{U_2/f^2} - \sqrt{U_1/f^1}),$$

where the right-hand side is strictly positive because $U_2/U_1 > f^2/f^1$. Hence, $\Delta\pi'(y) > 0$ for $y < y^{EK}$, i.e. $\Delta\pi(y)$ is strictly increasing in y for $y \leq y^{EK}$. □

This result implies that firms with type y^{EK} have the strongest incentive to screen. If all firms choose $\sigma = 1$ in equilibrium, then sufficiency follows from Proposition 2; if all firms choose $\sigma = 0$ in equilibrium, then sufficiency follows from Proposition 2 or Eeckhout and Kircher (2010). In the remaining case, where the equilibrium features both firms choosing $\sigma = 1$ and firms choosing $\sigma = 0$, we must have $\Delta\pi(y^{EK}) > c$ (otherwise all firms will choose $\sigma = 0$). There exist then two firm types $\underline{y}^s$ and $\overline{y}^s$ with $y' \leq \underline{y}^s < y^{EK} < \overline{y}^s \leq y''$, where firms of type $\underline{y}^s$ and $\overline{y}^s$ are indifferent between choosing $\sigma = 0$ and 1, i.e. $\Delta\pi(\underline{y}^s) = \Delta\pi(\overline{y}^s) = c$.

Firms with $y < \underline{y}^s$ will choose $\sigma = 0$ and attract only high-type workers; firms with $y \in (\underline{y}^s, \overline{y}^s)$ will choose $\sigma = 1$ and attract both types of workers; finally, firms with $y > \overline{y}^s$ will choose $\sigma = 0$ and attract only low-type workers. Since all firm types y between $\underline{y}^s$ and $\overline{y}^s$ choose $\sigma = 1$, submodularity implies that NAC/NAM holds within this interval. Combining the above results implies that NAC/NAM holds globally.

Note that we can not weaken the requirement of strict submodularity to mere submodularity for the sufficient condition. To see this, set $f(x, y) = x + y$ and initially set c large enough so that all firms choose $\sigma = 0$. Then for $y \geq y^{EK}$, $\Delta\pi(y)$ is a constant by equation (64a). If we set $c = \Delta\pi(y^{EK})$, all firms with $y \geq y^{EK}$ are indifferent between choosing $\sigma = 0$ with low-type applicants and $\sigma = 1$ with both types of applicants. This indeterminacy violates NAC/NAM.

The Analysis of PAC/PAM. First, with a slight abuse of notation, given x_1 and x_2, we define $\rho(x_1, x_2, y)$ as the solution to

$$\frac{f_y(x_2, y)}{f_y(x_1, y)} = \left(\frac{f(x_2, y)}{f(x_1, y)}\right)^{\rho(x_1, x_2, y)}. \tag{66}$$

By Lemma 7, $\rho(x_1, x_2, y) \in [\underline{\rho}, \overline{\rho}]$. Note that $\rho(x_1, x_2, y)$ is the discrete version of $\rho(x, y)$ defined in (1). We have $\rho(x_1, x_2, y) \to \rho(x, y)$ when $x_1, x_2 \to x$.

Second, to simplify exposition, we introduce a transformation $\Omega(\cdot)$ of $\kappa(y)$, the output dispersion parameter defined by equation (4). Define

$$\Omega(\kappa) \equiv \frac{1}{2} + \frac{\ln(\sqrt{\kappa} + \sqrt{1+\kappa})}{\ln(1+\kappa)}.$$

Lemma 13. *$\Omega(\kappa)$ is strictly decreasing with $\lim_{\kappa\to 0}\Omega(\kappa) = \infty$ and $\lim_{\kappa\to\infty}\Omega(\kappa) = 1$.*

Proof. By L'Hospital's Rule, $\lim_{\kappa\to 0}\Omega(\kappa) = \lim_{\kappa\to 0}\frac{1}{2} + \frac{1}{\sqrt{\kappa}+\sqrt{1+\kappa}}\left(\frac{1}{2\sqrt{\kappa}} + \frac{1}{2\sqrt{1+\kappa}}\right)(1+\kappa) = \infty$. In contrast, when $\kappa \to \infty$, we have $\kappa \approx 1 + \kappa$ and $\lim_{\kappa\to\infty}\Omega(\kappa) = \lim_{\kappa\to\infty}\frac{1}{2} + \frac{\ln(\sqrt{\kappa}+\sqrt{\kappa})}{\ln(\kappa)} = 1$.

Next, we prove that $\Omega(\kappa)$ is strictly decreasing. By direct computation,

$$\Omega'(\kappa) = \frac{\ln(1+\kappa) - 2\sqrt{\frac{\kappa}{1+\kappa}}\ln(\sqrt{\kappa} + \sqrt{1+\kappa})}{4\sqrt{\kappa(1+\kappa)}\ln(1+\kappa)}.$$

The derivative of the numerator above is $-\ln(\sqrt{\kappa}+\sqrt{1+\kappa})\sqrt{\frac{1+\kappa}{\kappa}}(1+\kappa)^{-2} < 0$. At $\kappa = 0$, the numerator is zero, which implies that it is strictly negative and hence $\Omega'(\kappa) < 0$ when $\kappa > 0$. □

We now provide a claim which is stronger than the statements in Proposition 6.

Claim. *Consider a log-supermodular function f. Given a distribution of agents' types, PAC/PAM holds in equilibrium as long as, for each y,*

$$\rho(x_1, x_2, y) \geq \Omega(\kappa(y)). \tag{67}$$

In contrast, given x_1, x_2 and $J(y)$, if for some $y^* \in (\underline{y}, \overline{y})$, we have

$$\rho(x_1, x_2, y^*) < \Omega(\kappa(y^*)), \tag{68}$$

then we can find (L, z) and c such that PAC/PAM fails in equilibrium.

Since $\Omega(\cdot)$ is strictly decreasing and with log-supermodular f, $\kappa(y)$ is increasing in y), the right-hand side of (67) reaches its maximum at $y = \underline{y}$. Also since $\rho(x_1, x_2, y) \geq \underline{\rho}$, the sufficient condition (30) in Proposition 6 then implies (67). On the other hand, given any log-supermodular function, whenever $x_1, x_2 \to x$, then $\kappa(y) \to 0$ and $\Omega(\kappa(y)) \to \infty$, and (68) holds for all $y^* \in [\underline{y}, \overline{y}]$, which, by the above claim, implies that we can find (L, z) and c such that PAC/PAM fails in equilibrium.

Note that for a CES production function, (67) reduces to $\rho \geq \Omega(\kappa(\underline{y}))$ and and (68) reduces to $\rho < \Omega(\kappa(\underline{y}))$. Thus, although the sufficient condition (30) is slightly weaker than (67), it is still sharp in the special case of CES production functions.

Similar to the analysis of NAC/NAM, since $f(x, y)$ is log-supermodular, and therefore strictly square-root supermodular, there exists a unique y^{EK} which solves (63). The first part of Lemma 11 states that y^{EK} must belong to the region characterized by (62). Furthermore, $f^2 - f^1$ is strictly increasing so that there exists at most one $y' < y^{EK}$ such that $f^2 - f^1 = U_2 - U_1$ (otherwise set $y' = \underline{y}$). Since we only assume weak log-supermodularity, f^2/f^1 is weakly increasing. Set $y'' = \min\{y \,|f^2/f^1 \geq U_2/U_1\}$ (if this set is empty, then set $y'' = \overline{y}$). The region characterized by (62) is then $y \in (y', y'')$. The following Lemma establishes

that under the sufficient condition (67), $\Delta\pi(y)$ is single-peaked at $y = y^{EK}$, so PAC/PAM follows from the same logic that was used for the case of NAC/NAM.

Lemma 14. *Suppose that $f(x, y)$ is log-supermodular. In the region characterized by (62), $\Delta\pi(y)$ is strictly increasing in y for $y \leq y^{EK}$, and if condition (67) holds for each $y \in (\underline{y}, \overline{y})$, then it is strictly decreasing in y for $y \geq y^{EK}$.*

Proof. If $y \in (y', y^{EK}]$, then $\Delta\pi(y)$ is given by (64a) and its derivative is given by (65a), so it is strictly increasing in y since $\Delta f_y > 0$. If $y \in [y^{EK}, y'')$, then $\Delta\pi(y)$ is given by (64b) and its derivative is now given by (65b) and can be rewritten as

$$\Delta\pi'(y) = f_y^1 \sqrt{\frac{\Delta U}{\kappa(y) f^1}} \left[-(1+\kappa(y))^{\rho(y)} \left(1 - \sqrt{\frac{\kappa(y)}{1+\kappa(y)}} \sqrt{\frac{U_2}{\Delta U}} \right) + 1 - \sqrt{\frac{\kappa(y)}{\Delta U / U_1}} \right],$$

where, to simplify notation, we shorten $\rho(x_1, x_2, y)$ as $\rho(y)$, and we used the identities $f^2/f^1 = 1 + \kappa(y)$ and $f_y^2/f_y^1 = (1+\kappa(y))^{\rho(y)}$.

Furthermore, define

$$\delta(y) \equiv \sqrt{\frac{\kappa(y)}{\Delta U / U_1}}, \tag{69}$$

which implies $\sqrt{U_2/\Delta U} = \sqrt{(\kappa(y) + \delta(y)^2)/\kappa(y)}$, and $\Delta\pi'(y)$ can be rewritten as

$$\begin{aligned}
\Delta\pi'(y) &= f_y^1 \sqrt{\frac{\Delta U}{\kappa(y) f^1}} \left[(1+\kappa(y))^{\rho(y)} \left(\sqrt{\frac{\kappa(y) + \delta(y)^2}{1+\kappa(y)}} - 1 \right) + 1 - \delta(y) \right] \\
&= f_y^1 \sqrt{\frac{\Delta U}{\kappa(y) f^1}} \left[(1+\kappa(y))^{\rho(y) - \frac{1}{2}} \sqrt{\kappa(y) + \delta(y)^2} - \left((1+\kappa(y))^{\rho(y)} - 1 + \delta(y) \right) \right] \\
&= f_y^1 \sqrt{\frac{\Delta U}{\kappa(y) f^1}} \frac{(1+\kappa(y))^{2\rho(y)-1} \left(\kappa(y) + \delta(y)^2\right) - \left((1+\kappa(y))^{\rho(y)} - 1 + \delta(y) \right)^2}{(1+\kappa(y))^{\rho(y) - \frac{1}{2}} \sqrt{\kappa(y) + \delta(y)^2} + \left((1+\kappa(y))^{\rho(y)} - 1 + \delta(y) \right)}.
\end{aligned}$$

Thus, $\Delta\pi'(y)$ has the same sign as the numerator of the last factor in the last line. Single out the numerator and define

$$\mathcal{S}(\delta, \kappa, \rho) = (1+\kappa)^{2\rho-1} \left(\kappa + \delta^2\right) - \left((1+\kappa)^{\rho} - 1 + \delta \right)^2, \tag{70}$$

which is a quadratic function of δ with a strictly positive second-order coeffi-

cient since we assume $\rho \geq 1$ (log-supermodularity). Note that $\mathcal{S}(1,\kappa,\rho) = 0$ and $\frac{\partial \mathcal{S}(\delta,\kappa,\rho)}{\partial \delta}\big|_{\delta=1} = 2(1+\kappa)^{\rho}((1+\kappa)^{\rho-1}-1) \geq 0$. Therefore, if $\mathcal{S}(0,\kappa,\rho) \leq 0$, then $\mathcal{S}(\delta,\kappa,\rho) < 0$ for all $\delta \in (0,1)$. Note that $\mathcal{S}(0,\kappa,\rho) = \kappa(1+\kappa)^{2\rho-1} - ((1+\kappa)^{\rho}-1)^2$, Thus $\mathcal{S}(0,\kappa,\rho) \leq 0$ if and only if $\sqrt{\frac{\kappa}{1+\kappa}}(1+\kappa)^{\rho} \leq (1+\kappa)^{\rho}-1$, or equivalently $\rho \geq \Omega(\kappa)$.

If for each $y \in (\underline{y},\overline{y})$, we have $\rho(y) \geq \Omega(\kappa(y))$, then by the above argument, $\mathcal{S}(\delta(y),\kappa(y),\rho(y)) < 0$ and hence $\Delta\pi'(y) < 0$ for $y \in [y^{EK}, y'')$. □

Similar to the case of NAC/NAM, we only need to consider the case where the equilibrium features both firms choosing $\sigma = 1$ and firms choosing $\sigma = 0$. Then there exist two firm types $\underline{y}^s$ and $\overline{y}^s$ that are indifferent between choosing $\sigma = 0$ and 1, where $y' \leq \underline{y}^s < y^{EK} < \overline{y}^s \leq y''$. Firms with $y < \underline{y}^s$ will choose $\sigma = 0$ and attract only low-type workers; firms with $y \in (\underline{y}^s, \overline{y}^s)$ will choose $\sigma = 1$ and attract both types of workers; finally, firms with $y > \overline{y}^s$ will choose $\sigma = 0$ and attract only high-type workers. Since all firms of y between $\underline{y}^s$ and $\overline{y}^s$ choose $\sigma = 1$, log-supermodularity implies that PAC/PAM holds within this interval. Combining the above results then implies that PAC/PAM holds globally.

Now consider the second part of the claim. Before we move to the detailed proof, we first give a brief sketch. If (68) holds, then we can find (L,z) and a large c such that all firms choose $\sigma = 0$ in equilibrium, and $\Delta\pi(y)$ reaches its maximum at some point $\widetilde{y} > y^{EK}$ (note that the maximum is between 0 and c here). Now decrease c gradually till firms near $\widetilde{y}$ find it optimal to choose $\sigma = 1$ and screen ex-post while firms with types slightly above y^{EK} will continue choosing $\sigma = 0$ and accordingly attract high-type applicants only. PAC/PAM then fails in this case. Below, we prove this claim formally.

We first prove the following. Given a log-supermodular function $f(x,y)$ and a distribution of agents' types, a necessary condition for PAC/PAM to hold for all c is that $\Delta\pi'_+(y^{EK}) \leq 0$ when c is sufficiently large (for example, $c \geq f(x_2,\overline{y})$) so that all firms choose $\sigma = 0$, where $\Delta\pi'_+(y^{EK})$ is the right derivative of $\Delta\pi(y)$ at point y^{EK}.

Suppose otherwise that $\Delta\pi'_+(y^{EK})$ is strictly positive; the maximum value of $\Delta\pi(y)$ must then be reached at some point $\widetilde{y} > y^{EK}$, since $\Delta\pi(y)$ is always strictly increasing when $y \in (y', y^{EK})$ (see Lemma 14). Now define $\widetilde{c} = \Delta\pi(\widetilde{y})$ and gradually decrease it from $f(x_2,\overline{y})$ to values around $\widetilde{c}$. What is the impact of this

change on the sorting pattern? As long as $c \geq \widetilde{c}$, no firm is willing to invest in screening, so the equilibrium allocation remains the same. When c is slightly below $\widetilde{c}$, then firms with types sufficiently close to $\widetilde{y}$ will choose $\sigma = 1$. Note that the equilibrium market utilities U_1 and U_2 will change slightly, so that y^{EK} also changes only slightly. As before, firms with types slightly above y^{EK} will therefore choose $\sigma = 0$ and hire high-type workers only, while firms with types sufficiently close to $\widetilde{y}$ will attract both types of workers. Hence, PAC/PAM fails to hold when c is slightly below $\widetilde{c}$.

Below, we complete the proof by showing that for any log-supermodular function $f(x, y)$ and $(x_1,\ x_2,\ J(y))$, if (68) holds for some $y^* \in (\underline{y}, \overline{y})$, then we can choose (L, z) such that $\Delta\pi'_+(y^{EK}) > 0$ when c is sufficiently large that all firms choose $\sigma = 0$.

Step 1: Since $\rho(y^*) < \Omega(\kappa(y^*))$, we have $\mathcal{S}\left(0, \kappa(y^*), \rho(y^*)\right) > 0$, where $\mathcal{S}$ is defined in equation (70). Thus, by continuity, we can find a δ^* small enough such that $\mathcal{S}(\delta^*, \kappa(y^*), \rho(y^*)) > 0$. Next, we construct (U_1^*, U_2^*) from the following two equations,

$$\sqrt{f(x_2, y^*)} - \sqrt{f(x_1, y^*)} = \sqrt{U_2^*} - \sqrt{U_1^*}$$
$$\delta^* = \sqrt{\frac{(f(x_2, y^*) - f(x_1, y^*))/f(x_1, y^*)}{(U_2^* - U_1^*)/U_1^*}}.$$

These equations are reminiscent of (63) and (69), respectively. The main difference is that there we considered the market utilities as known and solved for y^{EK} and $\delta(y)$; here we treat y^* and δ^* as known and solve for market utilities instead.

Step 2: Given (U_1^*, U_2^*), y^* is then the firm type that corresponds to y^{EK} defined before. Since f is log-supermodular and hence strictly square-root supermodular, firms with types $y > y^*$ will attract only high-type applicants, and firms with types $y < y^*$ will attract only low-type applicants. The firms' problem is $\max_\lambda\ m(\lambda)f(x_1, y) - \lambda U_1^*$ for $y \leq y^*$, and $\max_\lambda\ m(\lambda)f(x_2, y) - \lambda U_2^*$ for $y \geq y^*$. Denote the solution by $\lambda(y)$ for all y.

Step 3: Set $L(1 - z) = \int_{\underline{y}}^{y^*} \lambda(y) dJ(y)$ and $Lz = \int_{y^*}^{\overline{y}} \lambda(y) dJ(y)$. Then, by construction, (U_1^*, U_2^*) are indeed the market utilities, $y^* = y^{EK}$ for the equilibrium where all firms choose $\sigma = 0$, and $\Delta\pi'_+(y^{EK}) > 0$ because $\mathcal{S}(\delta^*, \kappa(y^*), \rho(y^*)) > 0$ and $y^* = y^{EK}$. □