Efficiency of Simultaneous Search

Philipp Kircher

University of Pennsylvania and Institute for the Study of Labor

This paper presents an equilibrium labor search model in which workers can simultaneously apply to multiple firms to increase their search intensity. They observe firms' wage postings before choosing where to apply. Owing to coordination frictions, a firm may not receive any applications; otherwise it is able to hire unless all its applicants have better offers. It is shown that the equilibrium converges to the efficient Walrasian outcome as application costs vanish. Even for nonnegligible application costs, the entry of firms, the search intensity, and the number of filled vacancies are constrained efficient. Wage dispersion is essential for constrained efficiency.

I. Introduction

In many markets, the decentralized nature of the interaction prevents efficient (Walrasian) exchange. For example, in the labor market, an unemployed worker who applies for a job might not be the only applicant, in which case one of the other applicants might be hired and he remains unemployed. Nevertheless, when the intensity of search is sufficiently high, one might expect the market allocation to be approximately Walrasian. This is indeed the case in sequential search models in which workers apply to one firm at a time, as long as the period length between applications becomes sufficiently small (e.g., Gale 1987;

This paper circulated under the title "Efficiency of Simultaneous Directed Search with Recall.'' The idea for this paper grew out of earlier fruitful collaboration with Manolis Galenianos. I am also grateful to Ken Burdett, Jan Eeckhout, Stephan Lauermann, Wolfram Merzyn, Georg Nöldeke, Andrew Postlewaite, and Randy Wright for many insightful discussions and to the editor Robert Shimer and two anonymous referees for very helpful comments. This paper greatly benefited from the input of seminar participants at numerous seminars and conferences. Financial support by the National Science Foundation, grant SES-0752076, is gratefully acknowledged.

[*Journal of Political Economy*, 2009, vol. 117, no. 5]

Mortensen and Wright 2002; Lauermann 2006; Satterthwaite and Shneyerov 2007).

In many search environments the period length may not be small, though. The hiring process, for example, may induce a substantial lag between the time when a worker applies for a job and the point at which the final hiring decision is reached. Van Ours and Ridder (1992) find in Dutch data a delay of 1–2 months to select an employee from an initial pool of applicants.[1] Delays are even larger in occupations in which entry positions are filled only in fixed (e.g., annual) intervals. In such environments the efficient Walrasian outcome is unattainable if workers can apply to only one firm in each selection period. However, the threat of not obtaining a job and having to wait another period is exactly what provides incentives for workers to send multiple applications simultaneously to several firms. The idea to model search as a simultaneous application process goes back to the foundations of labor search in Stigler (1962), and the ability of this channel to generate efficient equilibrium outcomes has recently been explored by Acemoğlu and Shimer (2000), Gautier and Moraga-Gonzáles (2005), Albrecht, Gautier, and Vroman (2006), and Galenianos and Kircher (2008). None of these find convergence to the Walrasian limit when the application costs vanish.[2]

This paper considers the efficiency of simultaneous search in a directed search environment in which each firm has one job and posts a wage to attract workers. Subsequently, workers decide to which firms they want to apply. Search frictions imply that workers might not obtain the job they apply for, which provides incentives to apply to several firms despite the fact that applications are costly. The frictions arise because workers cannot coordinate their applications and sometimes multiple workers apply for the same job. In contrast to the earlier literature that has restricted firms to communicate with only one of their applicants and to leave the job vacant if this applicant turns down the offer (even if many other workers applied for the job), here firms are allowed to communicate with all their applicants. It is shown that the market al-

[1] A concern for delays in hiring is warranted mainly in skill-intensive occupations. Van Ours and Ridder (1992) find little delay for low-education jobs, but delays become substantial as the educational requirement of the job increases.

[2] Acemoğlu and Shimer (2000) and Albrecht et al. (2006) explicitly consider limit results, which still feature a non-Walrasian component. Gautier and Moraga-Gonzáles (2005) and Galenianos and Kircher (2008) do not explicitly consider limit results, but by the close resemblance of their trading environment with that of Albrecht et al., they inherit the inefficiencies of that model. Classical equilibrium search papers such as Butters (1977) and Burdett and Judd (1983) also consider simultaneous search but focus on the distribution of wages. Except for the costs of information acquisition, these classical papers lack a margin of efficiency, since each firm can service the entire market. Efficiency questions are also absent in the original contribution by Stigler (1962), since he models the workers in a partial equilibrium setting in which the benefit from matching for the firms is not considered.

location approaches the Walrasian outcome when search costs become small. This result highlights how Stigler's notion of simultaneous search provides a foundation for Walrasian equilibrium when search costs are small and wages are determined by optimizing firms. Furthermore, it is shown that the equilibrium is constrained efficient even when search costs are nonnegligible.

The theory presented in this paper relies on two assumptions regarding the ability of firms to interact with workers: firms can compete for workers through binding wage announcements, and firms are able to communicate with all their applicants in the hiring process. To incorporate the latter, a new theory of matching is proposed. It is based on the idea that communication leads to a stable matching among the firms and their applicants. Stability here means that a firm does not leave the job vacant while one of its applicants starts to work at a lower wage or remains unemployed.[3] In such a case, both the firm and the worker would be better off by deviating and forming a match between themselves. While stability is imposed by assumption, it arises naturally if firms offer their job sequentially to applicants and workers are free to reconsider their options. During such a process a high-wage firm can entice a worker away from lower-wage competitors, and a job remains vacant only if all the firm's offers get rejected in favor of better alternatives.[4] In this setup, a firm can communicate only with those workers who applied to the firm, and therefore the model naturally differs from the existing stability analysis that presumes that all agents in the economy communicate with one another.

The central results in this paper concern the efficiency of the equilibrium. The efficient Walrasian outcome obtains in the limit as search costs vanish. Even when search costs are nontrivial, the equilibrium interaction is constrained efficient. The miscoordination between workers in their application decisions renders perfect matching impossible, leading to the standard result of the search literature that the labor market does not perfectly clear and therefore some workers face prolonged unemployment. Nevertheless, given the miscoordination and the

[3] Wages are treated as fixed in the matching stage. The implications and the applicability of this assumption are discussed in Sec. VI. This notion of stability has been successfully applied in many areas. In a recent paper, Bulow and Levin (2006) also deploy matching at fixed wages after a noncooperative wage-setting stage but assume that all workers apply to all firms.

[4] Such a sequential process resembles the well-known deferred acceptance process. For finite economies, it converges in finite time. For the continuum economy, stability is imposed to capture the spirit of a process by which firms contact all their applicants. In contrast to the standard matching literature that pioneered these concepts (see Gale and Shapley 1962), this paper deals with much less heterogeneity (essentially coming only through the wage-setting process, though some additional heterogeneity is discussed in Sec. VI) but adds the novelty that stability is defined only for agents that got to know each other in the application process.

notion of stability, the equilibrium is constrained efficient along the following operative margins even when application costs are nonnegligible. *Entry efficiency*: the constrained optimal number of firms enter. *Efficiency of search intensity*: the number of applications that workers send is constrained efficient. *Search efficiency*: workers send their applications in a way that induces the optimal number of matches. Thus, even when workers choose their search intensity strategically by deciding on a finite set of firms to which they apply at nonnegligible costs, market wages are such that workers make the socially optimal application decisions and firms provide the socially optimal number of jobs.

As a by-product the analysis reveals that wage dispersion is crucial to achieve constrained efficiency when workers apply for more than one job, even though workers and firms are homogeneous and risk neutral. Low-wage jobs endogenously create a safety net for workers who are unsuccessful with their other applications. In contrast, with a single wage, workers would apply randomly to firms, which too often generates several offers for some workers and none to others. The idea that matching frictions induce wage dispersion even in settings in which workers and firms are homogeneous has been pursued in a large body of work,[5] yet the insight that wage dispersion might arise in such a setting as the optimal response of the market to deal with the matching frictions is new to the literature. In this model, wage dispersion divides the market into separate wage segments that attract one application from each worker. These segments resemble standard one-application markets, and the insights from the simpler one-application models carry over. In particular, the equilibrium wage in each segment is characterized by a modified version of Hosios' (1990) condition for optimal entry by firms.

Earlier work on simultaneous search did not find constrained efficient outcomes, nor convergence in the limit to the Walrasian allocation. This has two main reasons. In Acemoğlu and Shimer (2000) and Gautier and Moraga-Gonzáles (2005), workers do not observe all wage offers when they make their search decision, which induces equilibrium wages that do not reflect the workers' marginal product. In directed search models, workers can observe all wage offers, and a central insight of this literature is that equilibrium wages do reflect the workers' marginal product (see, e.g., Moen 1997; Shi 2001; Shimer 2005) when workers send one application. Albrecht et al. (2006) and Galenianos and Kircher (2008) show that this insight might not be valid when workers apply to multiple firms. The main reason is a contractual incompleteness: A firm's wage offer rewards a worker only for applying to it, but this reward

[5] See, e.g., Butters (1977), Burdett and Judd (1983), Burdett and Mortensen (1998), Acemoğlu and Shimer (2000), Mortensen (2003), Delacroix and Shi (2004), Albrecht et al. (2006), Gaumont, Schindler, and Wright (2006), and Galenianos and Kircher (2008).

is not conditional on the other firms to which the worker applied (even though job offers from these other firms potentially affect the first firm negatively). In these earlier models this problem was exacerbated by the assumption that each firm can talk to only one of its applicants and its job remains vacant if this applicant declines its offer in favor of a different firm. This has the counterintuitive effect that the number of matches may go down when workers apply more often since some workers get multiple offers that prevent others from obtaining jobs. This feature prevents convergence to the Walrasian limit. In this model, firms can communicate with all workers in a way that leads to a stable matching. While stability is no guarantee for efficiency, it does have the feature that a worker who rejects a job does not block other workers from getting employed there, which is crucial for the limit result. Additionally, in equilibrium the contractual incompleteness is solved because of the interaction among workers themselves: if some applicants apply elsewhere, the job becomes more attractive and other workers will apply there more strongly. Through this channel the noncontracted externality is internalized. This point is somewhat subtle and is elaborated on in the discussion section.

To my knowledge, this is the first attempt to integrate the two-sided strategic considerations of a search environment with stability concepts used in matching markets.[6] The paper draws on three strands of literature. Insights from the directed search literature (e.g., Peters 1991; Burdett, Shi, and Wright 2001) informed the modeling of the frictions and information flows in the market. With multiple applications, workers face a simultaneous portfolio choice. For this type of problem Chade and Smith (2006) consider an individual agent's choice and derive a simple characterization of the optimal decision rule; Galenianos and Kircher (2008) derive implications in an equilibrium search framework. For the final matching, the stability concept pioneered by Gale and Shapley (1962) is applied to the network formed in the search process.

Since notation and exposition are much more tractable when at most two applications per worker are considered, Sections II–IV are restricted to this case. Section II presents the model. Section III characterizes the constrained optimal allocation. Section IV shows that the decentralized economy implements the constrained efficient allocation. Section V lifts the restriction of two applications per worker and, additionally, discusses convergence when application costs vanish. Section VI discusses in more detail why efficiency prevails in this environment but failed in related models and which of the assumptions are crucial for this result. It also

[6] Gautier and Moraga-González (2005) present a three-player example with a similar concept for random search.

outlines how heterogeneity can be introduced into this setting. Section VII presents conclusions. Omitted proofs are gathered in the Appendix.

II. The Model

This section presents the physical environment, as well as some restrictions on the equilibrium set that are intended to capture the search frictions of a large market.

Players and preferences.—The economy consists of a continuum of size 1 of unemployed workers and a large continuum of firms, with a measure v active in equilibrium.[7] Each firm has a single vacant job, which, if filled, produces one unit of output. All agents are risk neutral. Each firm maximizes its expected output minus wage and entry costs. Each worker maximizes his expected wage payments minus the cost of applying for jobs.

The labor market interaction.—The market interaction proceeds in three stages. In the first stage, firms can become active by posting a wage that they commit to pay when they hire a worker. A firm that becomes active incurs an entry cost $K < 1$ that reflects the costs of setting up the job and advertising it to workers. In the second stage, workers observe all active firms and their wage postings. Each worker decides on the number i of applications he wants to send at cost $c(i)$, which subsumes monetary as well as time and effort costs. A worker who does not send applications does not incur any costs. The marginal costs $c_i = c(i) - c(i-1)$ are everywhere weakly increasing, and they are strictly positive at some finite $i \in \mathbb{N}$. The worker also selects the i active firms to which he applies. For expositional simplicity, the restriction $i \leq 2$ is initially imposed; a generalization to an arbitrary number of applications is presented in Section V. After the applications are sent out, they form the links in a network between workers and firms: each worker is linked to the firms to which he applied and each firm is linked to its applicants. In the final stage, workers and firms form matches, the announced wage is paid, and matched pairs start production.

Matching.—The matching in the final stage is assumed to be stable on the network: matches form such that a firm's job remains vacant only if either it has no applicants or all its applicants are matched at weakly higher wages. This concept of pairwise stability is motivated by the fact that otherwise the vacant firm could offer its job to the applicant who is matched at a lower wage, and both would be jointly better off by forming a match between themselves. For economies with a finite number of workers and firms, a stable matching is reached by an ex-

[7] Since constant returns to scale in matching will be assumed, only the ratio of workers to active firms matters.

tensive form game in which firms successively make offers to workers, workers hold on to the best offer they have so far received and decline others, and rejected firms make additional offers until every firm either has a worker or has no applicants left to whom it could offer the job (see Gale and Shapley 1962). The approach in this paper can be viewed as a limit of such an interaction when the number of agents gets large, but following the literature on matching with a continuum number of agents, the paper imposes stability directly rather than considering the convergence of the extensive form.[8] The difference between this paper and the earlier literature is that here stability refers only to firms and workers that got to know each other through the (endogenous) application process. The main implication of pairwise stability is that firms with higher wages hire "first," and firms at lower wages can hire only those applicants who have not obtained a higher-paying job. This notion of stability presumes that firms cannot react to other firms' offers and are committed to their wage announcement—an assumption that is discussed in detail in Section VI, where the forces leading to efficiency are highlighted.

Anonymity assumptions.—There exist many subgame-perfect strategies by which workers might apply for jobs after observing the wages, including the frictionless case in which each worker applies to a different firm (if enough firms entered). Here the attention is on a market that is large and anonymous. To capture this, in the spirit of the directed search literature, consider equilibrium strategies that are symmetric and anonymous. Anonymity means that agents that are observationally identical except for their name are treated identically. Anonymity is assumed both at the application and the hiring stages. In particular, the following assumptions are made:

[8] The extensive form process terminates in finite time, i.e., after a finite number of rounds of offers, for any economy with a finite number of workers and firms as shown in Gale and Shapley (1962). Yet as the size of the economy extends to infinity, the time to termination might tend to infinity as well. In this regard, note the following: If workers who currently hold an offer inform all firms with lower offers that they are no longer available, even in this continuum economy the process terminates in finite time in the optimal assignment and in the decentralized equilibrium. In particular, the number of rounds of offers equals the highest number of applications sent by an individual worker. This cannot be ensured for suboptimal assignments and nonequilibrium interactions, though. Even for such interactions, Kircher (2006) shows that the measure of agents that are not stably matched (in the sense that they could form blocking pairs among the agents they are linked to) converges to zero as the number of rounds goes to infinity. Nevertheless, full stability might not be reached in finite time.

For stability in continuum models, see, e.g., Kaneko and Wooders (1986) or Gretzky, Ostroy, and Zame (1999).

1. A firm only observes whether a worker applied to it or not. It cannot obtain a report or direct evidence from a worker about the other firms to which he has applied. Anonymity then implies that a firm hires at random one of those applicants who lack better offers.
2. All workers use the same application strategies and do not condition on the firms' identities. They treat firms at different wages differently, but firms with the same wage are treated identically.

The first assumption deals with ties in the hiring process that arise because workers are identical. These ties are broken randomly, which implies that the process selects neither the best nor the worst stable matching. To see this, consider a firm that has two applicants A and B and breaks the tie between them by randomly hiring one of them. If applicant A applied nowhere else whereas applicant B also applied for a lower-wage job for which he is the only applicant, then it would be strictly more efficient to deterministically break the tie in favor of A and strictly less efficient to break it in favor of B. Note, though, that firms do not directly benefit from improved matches at other (lower-wage) firms. Therefore, any small amount of noise about match-specific quality of the applicants would lead firms to pick the most suitable candidate in a way that looks random to an outsider rather than choose the candidate who improves the match at another firm.

The second assumption creates the frictions that make tie-breaking necessary: the use of identical and anonymous strategies by the workers implies that they randomize equally over firms with the same wage and sometimes miscoordinate in the sense that several of them apply for the same job that only one of them can get.[9] The next section discusses in detail how these assumptions translate into the hiring probabilities in the economy.

III. Social Planner's Problem

Before turning to the equilibrium interaction between workers and firms, it is instructive to first consider the related problem of a social planner. Following Pissarides (2000) and others, consider a planner who has the same choice margins and restrictions that the agents in the decentralized equilibrium face. In terms of choice margins, the planner determines the number of firms that enter the economy, the number of applications that workers send, and where these applications are sent. In terms of restrictions, the planner is subject to the two anonymity

[9] Miscoordination arises even when there are fewer workers than firms, and even when all firms offer different wages. In the latter case, symmetric, i.e., identical, strategies still imply that workers all apply to any given set of firms with the same probability. For a careful argument in one-application environments, see Peters (1997).

restrictions and to the final matching stage that was imposed on the decentralized equilibrium. Placing these restrictions on the planner gives a relevant benchmark for comparing the efficiency of the decentralized wage competition. Without the anonymity restrictions the planner could trivially do better by assigning only one worker to each firm and avoiding the miscoordination in the application process, or by coordinating the tie-breaking in the matching stage across firms. Since the efficiency concept takes these features of the environment as given and does not assess which other matching concepts can improve the outcome further, the resulting optimum is called constrained efficient.

A. *Planner's Choices, Resulting Output, and Efficiency Criterion*

Planner's choice set.—The social planner can choose the measure v of firms in the economy. He can also choose the probability γ_i that a worker sends i applications. By the usual law of large number convention, this coincides with the fraction of workers who send i applications. Since we are considering the case in which workers have at most two applications, the planner chooses a search intensity $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ in the three-dimensional unit simplex Δ_3.

In the decentralized economy, firms are distinguished only by a one-dimensional attribute, the wage. This one-dimensional attribute allows workers to discriminate between firms. Symmetry and anonymity preclude any further differentiation. While the planner does not care about wages per se because they are just transfers, he might want to distinguish between firms to allow workers to apply in a better way. To incorporate this, the planner is allowed to distinguish firms along a single dimension by assigning them to different locations. Stability in the final matching stage translates into the requirement that a firm remains vacant only if all its applicants are matched at a weakly higher location.

The planner can choose a finite set of locations $L \subset \mathbb{R}$ and can then assign firms and applications to any location $l \in L$. To formalize the assignment, let Φ_L^1 and Φ_L^2 be the set of cumulative distribution functions over L and $L \times L$. Then the planner chooses a distribution $F \in \Phi_L^1$ for the firms, where $F(l)$ is the fraction of firms at locations weakly lower than l. For workers who send one application, let $G_1 \in \Phi_L^1$ denote the distribution of their choice of location, where again $G_1(l)$ is the fraction of the one-application workers who send their application to a location weakly lower than l. Workers who send two applications send them according to $G_2 \in \Phi_L^2$, where $G_2(l_1, l_2)$ is the fraction of these workers who send their first application to firms at a location weakly below l_1 and their second application to firms at a location weakly below l_2. These distributions are fully characterized by their respective mass points f, g_1, and g_2, where $vf(l)$ is the measure of firms at location l, $\gamma_1 g_1(l)$ is

the measure of workers who apply to l and send one application, and $\gamma_2 g_2(l_1, l_2)$ is the measure of workers who send two applications to l_1 and l_2. Without loss of generality let the first argument refer to the lower location; that is, $g_2(l_1, l_2) > 0$ only if $l_1 \leq l_2$.

Output and hiring probabilities.—The output in the economy coincides with the number of matches in the economy. How these are determined by the application behavior and the assumptions on matching is laid out in the following. Let η_l denote the probability that a firm at location l hires a worker, and let p_l be the probability that a worker who applies to location l would get hired there if he wants the job (i.e., if he has no offer from a higher location). As in other directed search models, the probabilities η_l and p_l depend on the number of applications relative to the number of firms at a given location. This ratio is called the gross queue length because it describes the average queue of applications per job at a location l and is defined as

$$\lambda_l = \frac{\gamma_1 g_1(l) + \gamma_2 \sum_{l'=1}^{L} [g_2(l', l) + g_2(l, l')]}{vf(l)} \tag{1}$$

when $vf(l) > 0$. When $vf(l) = 0$, the absence of firms trivially implies that there are no matches at this location.

The important novelty in this setup is that a firm might not be able to hire even if it obtains an application, because the applicant might have applied to another firm at a higher location and obtained a job there. Applications are called "effective" if the applicant does not obtain a higher job. Denote by ψ_l the fraction of applications that are not effective in the sense that the applicant is unavailable for hiring because of higher-ranked alternatives. The ratio of effective applications to firms at location l is called the *effective* queue length at location l and is defined as

$$\mu_l = (1 - \psi_l)\lambda_l. \tag{2}$$

Owing to the anonymity of the workers' strategy, the effective applications that are sent to firms at location l are uniformly distributed over these firms. Therefore, the number of effective applications at any individual firm at this location is random. It has been shown that the distribution of effective applications at any individual firm is given by a Poisson distribution with parameter μ_l, which implies that the probability that a firm receives no effective applications is $e^{-\mu_l}$.[10] If the firm

[10] This is obtained by considering a finite number of effective applications that are randomly distributed to a finite number of firms and by taking the limit as the number of applications and firms converges to infinity while retaining a ratio μ_l. A simple derivation is provided, e.g., in Burdett et al. (2001). They also provide careful and intuitive derivations for the following expressions for the hiring probabilities $\eta(\cdot)$ and $p(\cdot)$ in eqq. (3) and (4).

receives at least one effective application, it is able to fill its vacancy, because once a firm has a worker who is not employed at a better location, it will not remain unmatched by our stability assumption. Therefore, the hiring probability for a firm at location l is

$$\eta_l = 1 - e^{-\mu_l}. \tag{3}$$

Next, consider the probability of getting a job at location l for a worker who wants to obtain a job at that location, that is, who is effective in the sense that he did not get a job offer from a higher location. The worker cares about only those rival applicants who do not have a higher offer either, because those workers who take higher jobs are not competing for this job. Given that there are $1 - e^{-\mu_l}$ matches per firm and μ_l effective applications per firm, the probability of an effective application yielding a match is given by

$$p_l = \frac{1 - e^{-\mu_l}}{\mu_l}, \tag{4}$$

with the convention that $p_l = 1$ if $\mu_l = 0$.

The probability ψ_l that an applicant is not available for hiring because he accepts some (weakly) higher offer is trivially zero if workers send only one application and irrelevant if no worker applies to location l. Otherwise, consider some application sent to location l, and let $\hat{g}(l'|l)$ denote the probability that the sender mailed a second application and sent it to location l'.[11] The applicant takes a job from a strictly higher location if he applied to a strictly higher location and receives an offer. He takes an equally high offer if he applied to another firm at the same location, this other firm also wants to hire him, and the applicant chooses the other firm over the current firm. One can think about workers who apply twice to the same location as randomizing in advance about which offer they would prefer to accept in case they get equivalent offers. Since workers do not condition their choices on the identities of firms, both firms have equal chances of attracting the worker, and therefore an application is not effective with probability $\hat{g}(l|l)p_l/2$ because of offers from other equally high firms. Then ψ_l is given by

$$\psi_l = \sum_{l'>l} p_{l'}\hat{g}(l'|l) + \frac{p_l}{2}\hat{g}(l|l). \tag{5}$$

The system defined by (2), (4), and (5) is recursive: At the highest

[11] The likelihood that an application at l was sent from someone who sent another application to l' is

$$\hat{g}(l'|l) = \frac{\gamma_2[g_2(l', l) + g_2(l, l')]}{\gamma_1 g_1(l) + \gamma_2 \sum_{l'' \in L}[g_2(l'', l) + g_2(l, l'')]}.$$

location the probabilities p_L, η_L, and ψ_L can be uniquely determined without knowledge of hiring probabilities at other locations. This is due to the assumption that ties are broken irrelevant of the matching at lower locations. These values are then used to evaluate the corresponding terms at lower locations. Since the specification allows firms to hire any worker who is not matched at a higher or equally high location, the final matching is stable in the sense that no firm is vacant while one of its applicants is unmatched or matched at a strictly lower location.

Planner's objective.—The firms' hiring probability η_l is uniquely determined by the planner's choice of entry v, number of applications $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$, number of locations L, and distributions F, G_1, and G_2 for firms and workers over these locations. Let $M(v, \boldsymbol{\gamma}, L, F, G_1, G_2) = v\sum_l [\eta_l f(l)]$ denote the resulting output in the economy. The economy is called constrained efficient if its surplus is maximized, that is, if output minus costs achieves the maximum of

$$\max_{v,\boldsymbol{\gamma},L,F,G_1,G_2} M(v, \boldsymbol{\gamma}, L, F, G_1, G_2) - vK - \gamma_1 c(1) - \gamma_2 c(2). \tag{6}$$

The optimal solution can be characterized in three successive steps. First, search efficiency is analyzed by considering the optimal application behavior when the number of firms v and the search intensity $\boldsymbol{\gamma}$ are taken as given. Next, entry efficiency is analyzed by additionally considering the optimal entry of firms. Finally, the efficiency of the search intensity is analyzed by additionally characterizing the optimal number of applications. The last step coincides with the notion of constrained efficiency.

B. Search Efficiency, Given Entry and Search Intensity

Take entry v and search intensity $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ as given and consider the solution to the maximization problem (6) over the restricted parameter set $\{L, F, G_1, G_2\}$. The solution to this restricted problem is called "search efficient." Clearly, if either $v = 0$ or $\gamma_0 = 1$, there is nothing to analyze because there will be no matches in the economy. The case in which workers send only one application ($\gamma_2 = 0$) is well understood: Search efficiency is achieved with only one location at which workers who send an application randomly apply for jobs (see, e.g., Shimer 2005).

Here, the focus is on the novel case in which at least some workers send two applications (i.e., $\gamma_2 > 0$). Given the possibly large number of locations and the possibility to correlate applications across locations, this is a nontrivial problem. The first result significantly simplifies the problem by showing that search efficiency does not require more locations than applications sent by an individual worker, which means that

two locations $L = \{1, 2\}$ suffice. Moreover, the optimal search strategy is simple: every worker with a single application sends it to location 1 and workers with two applications send one to location 1 and the other to location 2.

LEMMA 1. For given entry $v > 0$ and search intensity $\boldsymbol{\gamma}$ with $\gamma_2 > 0$, output $M(v, \boldsymbol{\gamma}, L, F, G_1, G_2)$ is maximized with $L = \{1, 2\}$ locations, $g_1(1) = 1$, $g_2(1, 2) = 1$, and appropriate F.

The intuitive argument behind the formal proof, which is presented in the Appendix, is the following. Any location with effective queue length μ can be subdivided into two locations that both have queue length μ. This does not change the number of matches since firms hire with exactly the same probability as before. Moreover, it is possible to assign low and single applications to one of the locations and high applications to the other while retaining the ratio μ of effective applications to firms at each location, because the number of firms at each of the locations can be varied appropriately. That leaves us with a set of locations that receive low (or single) applications and a set of locations that receive high applications. Because the firms' matching probability $1 - e^{-\mu}$ is concave, it is optimal that all locations in each set have the same queue length, and so the locations in each set can be merged without loss of efficiency. There is no loss to optimality by having both single and low applications at the same location, because a worker who sends two applications but is unsuccessful with his high application is in the same situation as a worker who sends only one application. So conditional on failing at the high location, their problem looks identical. The trade-off for high and low applications is different, though, and it will not be optimal to have equal expected queue lengths in both locations, as we will see in the following.

Given lemma 1, we have to characterize only the distribution F, that is, the fraction of firms at each of the two locations. Let $f(1) = 1 - \rho$ and $f(2) = \rho$. The social planner chooses the optimal fraction of firms at each of the locations to maximize the output according to

$$\max_{\rho \in [0,1]} \tilde{M}(\rho) = (1 - \rho)v(1 - e^{-\mu_1}) + \rho v(1 - e^{-\mu_2}), \tag{7}$$

where the first term reflects the measure of firms at the low location times their matching probability. Similarly, the second term accounts for the firms at the high location. By equations (1)–(4) we have $p_i = (1 - e^{-\mu_i})/\mu_i$ with $\mu_2 = \lambda_2 = \gamma_2/(v\rho)$, $\mu_1 = [1 - \gamma_2 p_2/(\gamma_1 + \gamma_2)]\lambda_1$, and $\lambda_1 = (\gamma_1 + \gamma_2)/[v(1 - \rho)]$. The first-order condition to problem (7) is

given by[12]

$$\frac{d\tilde{M}}{d\rho}\frac{1}{v} = -(1 - e^{-\mu_1} - \mu_1 e^{-\mu_1}) + (1 - e^{-\mu_2} - \mu_2 e^{-\mu_2})(1 - e^{-\mu_1}) = 0. \tag{8}$$

Since the first two terms in parentheses are strictly increasing in μ_i, equation (8) immediately implies that $\mu_2 > \mu_1$. Therefore, it is optimal that jobs at the lower locations are easier to get. The intuition for this result comes from the idea of a safety net: Those workers who end up getting jobs at the low location will be unemployed if they do not get the job, and so the planner assigns many firms to this location to make it easy for these workers to get matched. In contrast, those workers who get jobs at the high location also applied for low-location jobs and might get hired there even in the absence of the high-location job. Therefore, workers who end up at the high location are less at risk, and it is optimal to make it harder for them and easier for the others.

Equation (8) becomes even more intuitive when we interpret it from the firms' point of view. The terms $1 - e^{-\mu_i} - \mu_i e^{\mu_i}$ reflect the additional probability of generating a match at location i when adding one more firm to that location. It entails the probability $1 - e^{-\mu_i}$ that the additional firm can hire minus the "business stealing" effect on other firms at the same location. The latter effect coincides with the probability $\mu_i e^{\mu_i}$ that the other firms had exactly one applicant who is stolen from them. At location 1, business stealing is restricted to this location only, because no worker with a job offer from location 2 will be "stolen" by an additional firm at location 1. This is different at location 2. Every additional match created at location 2 takes away an effective applicant from location 1. Therefore, the additional output is not one unit but is decreased by the probability $e^{-\mu_1}$ that the firm at the low location had only this one worker and can no longer produce.[13] Equation (8) therefore equates the marginal benefit of a firm at the first location to the marginal benefit at the second location. Since both μ_i's are functions only of the fraction ρ of firms at the high location (given $\boldsymbol{\gamma}$ and v), equation (8)

[12] The first derivative of (7) is

$$\frac{d\tilde{M}}{d\rho} = \left\{-(1 - e^{-\mu_1}) + (1 - e^{-\lambda_2}) + \left[-\gamma_2\lambda_1\frac{dp_2}{d\rho} + (\gamma_1 + \gamma_2 - \gamma_2 p_2)\frac{d\lambda_1}{d\rho}\right]\frac{e^{-\mu_1}}{v\lambda_1} + \rho e^{-\lambda_2}\frac{d\lambda_2}{d\rho}\right\}v.$$

The last summand in the braces equals $-\lambda_2 e^{-\lambda_2}$, the third summand equals $e^{-\mu_1}[-(1 - e^{-\lambda_2} - \lambda_2 e^{-\lambda_2}) + \mu_1]$, and $\mu_2 = \lambda_2$.

[13] The adjustment $1 - e^{-\mu_1}$ also coincides with the probability that a low-location firm hires a worker, but this is coincidence in the special case in which all firms have equal productivity, as can be seen in n. 36, which considers differences in firms' productivities.

suffices to determine the optimal fraction of firms uniquely.[14] Optimality is ensured by global concavity.[15]

Since high-location firms have an additional business stealing effect, the optimal number of firms relative to the number of effective applications is lower. Therefore, despite homogeneity of labor, it is optimal to implement different hiring probabilities at different firms, which necessitates more than one location.

LEMMA 2. For given entry v and search intensity $\boldsymbol{\gamma}$ with $\gamma_2 > 0$, the output with $|L| = 1$ locations is always strictly smaller than the optimal output with $|L| = 2$ locations.

In the following, let $M^P(v, \boldsymbol{\gamma})$ be the planner's solution for the optimal number of matches according to optimality condition (8), let $p_i^P(v, \boldsymbol{\gamma})$ be the optimal probability of getting hired at each location, and let $\rho^P(v, \boldsymbol{\gamma})$ be the optimal fraction of firms at the high location.

C. *Optimal Entry, Given Search Intensity*

We now additionally derive the condition for optimal entry, for given search intensity $\boldsymbol{\gamma}$ with $\gamma_0 < 0$. The objective is

$$\max_{v \in \mathbb{R}} M^P(v, \boldsymbol{\gamma}) - vK. \tag{9}$$

In the Appendix the following first-order condition with respect to entry v is derived, which uniquely determines the level of entry as a function of the fixed entry cost:

$$K = 1 - e^{-\mu_1} - \mu_1 e^{-\mu_1}. \tag{10}$$

Together with condition (8) for efficient search this implies

$$K = (1 - e^{-\mu_2} - \mu_2 e^{-\mu_2})(1 - e^{-\mu_1}) \tag{11}$$

if $\gamma_2 > 0$. Similarly to the previous discussion of equation (8), the right-hand side of these equations captures the marginal benefit of adding one more firm to the first and second location, respectively. Optimality requires these to equal the entry cost. Let $v^P(\boldsymbol{\gamma})$ be the optimal level of entry given the application behavior, and let μ_1^P and μ_2^P denote the unique effective queue lengths that solve (10) and (11).

[14] Some tedious algebra establishes uniqueness. It can be found in Kircher (2006, lemma 1, pt. 1).

[15] We have

$$\frac{d^2\tilde{M}}{d\rho^2} = -v\left[\frac{\lambda_2^2 e^{-\lambda_2}(1 - e^{-\mu_1})}{\rho} + \frac{e^{-\mu_1}(1 - e^{-\lambda_2} - \lambda_2 e^{-\lambda_2} - \mu_1)^2}{1 - \rho}\right] < 0.$$

D. *Optimal Search Intensity*

Finally, the optimal number of applications $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ solves

$$\max_{\boldsymbol{\gamma} \in \Delta_3} M^P(v^P(\boldsymbol{\gamma}), \boldsymbol{\gamma}) - v^P(\boldsymbol{\gamma})K - \gamma_1 c(1) - \gamma_2 c(2). \tag{12}$$

This objective is equivalent to the notion of constrained efficiency in program (6). The optimal solution for this program is derived in the Appendix. Note that equations (10) and (11) uniquely determine the optimal effective queue lengths μ_1^P and μ_2^P at each location independent of the exact search intensity $\boldsymbol{\gamma}$. The reason is that under optimal entry, the number of firms adjusts to keep the hiring probabilities constant. The optimal search intensity is determined by two cutoffs, $t_1^P = e^{-\mu_1^P}$ and $t_2^P = e^{-\mu_2^P}(1 - e^{-\mu_1^P})$, in the following way:

$$\begin{aligned} \gamma_2 &= 1 \quad \text{if } c_2 < t_2^P, \\ \gamma_1 &= 1 \quad \text{if } c_1 < t_1^P \text{ but } c_2 > t_2^P, \\ \gamma_0 &= 1 \quad \text{if } c_1 > t_1^P. \end{aligned} \tag{13}$$

These cutoffs have a straightforward interpretation: t_1^P is the probability of sending the first application to a firm that does not yet have an applicant. The benefits from sending any applications are positive only if the costs are lower than this threshold. The term t_2^P corresponds to the probability $e^{-\mu_2^P}$ that the second application goes to a firm that does not yet have an applicant times the probability $1 - e^{-\mu_1^P}$ that the first application goes to a firm that has an applicant, in which case the second application leads to a new match. Again, a second application is worthwhile only if this benefit outweighs the costs. When $c_1 = t_1^P$, the surplus from having workers and firms interact is exactly zero as long as no worker sends more than one application, so any γ_1 is efficient as long as $\gamma_2 = 0$. Similarly, for $c_2 = t_2^P$, the surplus from sending a second application is zero, and any γ_2 is efficient as long as $\gamma_0 = 0$.

In summary, this section has laid out a set of necessary and sufficient conditions for optimality in (10), (11), and (13).

IV. The Decentralized Economy

This section focuses on the interaction in the decentralized economy, and it is shown that it implements the constrained efficient solution.

A. *Strategies, Expected Payoffs, and Equilibrium Definition*

Strategies.—In the decentralized economy, a pure strategy for a firm that enters is a wage offer $w \in [0, 1]$. In order to highlight the con-

nection between the decentralized problem and the social planner's problem, with slight abuse of notation, the same names will be used for distributions over wages that were used in the planner's problem for distributions over locations. Then F denotes the wage distribution, with support denoted by $\mathcal{F}$. That is, $F(w)$ gives the proportion of firms that offer wages below w. A worker observes the distribution of posted wages and then chooses the number of applications and the firms to which he sends them. Given the focus on anonymous strategies, a worker applies with equal probability to all firms with the same wage.[16] Therefore, his choice of firm can be summarized by its wage. A mixed strategy for a worker is then given by the tuples $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$ and $\boldsymbol{G} = (G_1, G_2)$. The term γ_i is the probability of sending i applications; $G_1(w)$ gives the probability that the worker sends his application to a firm with a wage below w when he sends just one application; and $G_2(w_1, w_2)$ gives the probability that the worker applies for a job with a wage below w_1 with the first and below w_2 with the second application when he sends two applications, where we assume $w_2 \geq w_1$ throughout. Since we assume that workers use symmetric strategies, these probabilities also reflect the fraction of the population that undertake a given action.

Expected payoffs.—Expected payoffs for firms and workers obtain as follows. Let $\eta(w)$ denote the probability that a firm posting wage w hires a worker. Let $p(w)$ be the probability that an effective applicant at wage w gets hired. These probabilities depend on the strategies, as described below. The expected profit of a firm posting wage w, omitting entry costs, is

$$\pi(w) = \eta(w)(1 - w). \tag{14}$$

The profits equal the hiring probability multiplied by the profit margin if a worker is hired.

The utility of a worker who sends no application is $U_0 = 0$. A worker who sends one application to a firm with wage w obtains utility

$$U_1(w) = p(w)w - c(1), \tag{15}$$

that is, the expected wage minus the cost of the application. A worker who applies for a job with wages (w_1, w_2), $w_2 \geq w_1$, obtains utility

$$U_2(w_1, w_2) = p(w_2)w_2 + [1 - p(w_2)]p(w_1)w_1 - c(2). \tag{16}$$

The worker's utility is given by the wage w_2 if he is hired at that wage, which happens with probability $p(w_2)$. With the complementary probability $1 - p(w_2)$, he does not receive an offer at the high wage, and his

[16] This is optimal if all other workers also follow anonymous strategies. At the expense of substantial additional notation, one can show that symmetry rather than anonymity of the workers' strategy is sufficient to yield identical hiring probabilities to firms with the same wage.

utility is w_1 if he gets an offer for his low-wage application, which happens with probability $p(w_1)$. He always incurs the cost for the two applications.

Hiring probabilities.—The hiring probabilities are analogous to those in the planner's problem. Let $\eta(w)$ be the effective queue length, $\lambda(w)$ the gross queue length, and $\psi(w)$ the probability that an applicant is not available for hiring at wage w. In direct analogy to equations (2)–(4) we have

$$\mu(w) = [1 - \psi(w)]\lambda(w), \tag{2'}$$

$$\eta(w) = 1 - e^{-\mu(w)}, \tag{3'}$$

and

$$p(w) = \frac{1 - e^{-\mu(w)}}{\mu(w)}, \tag{4'}$$

with the convention that $p(w) = 1$ if $\mu(w) = 0$.

To fix the probability $\psi(w)$ that an applicant is not effective, consider some application sent to a firm with a wage w in the support of the wage offer distribution, and let $\hat{G}(w'|w)$ denote the probability that the sender mailed a second application and sent it to a firm with a wage weakly lower then w'. Let $\hat{g}(w|w)$ be the probability that the sender mailed another application to a firm with the same wage.[17] Similar to (5), $\psi(w)$ is given by

$$\psi(w) = \int_{w'>w} p(w')d\hat{G}(w'|w) + \frac{p(w)}{2}\hat{g}(w|w). \tag{5'}$$

Finally, the gross queue length is still the ratio of applications to firms. It turns out to be more convenient to characterize it by the following condition that relates the measure of applications that workers send to

[17] To define the dependence of distribution $\hat{G}(w'|w)$ on the equilibrium strategies, let $G_2^1(w_1|w_2)$ denote the conditional and $G_2^1(w_1)$ the marginal distribution of G_2 with respect to the first wage, and let $G_2^2(w_1|w_2)$ denote the conditional and $G_2^2(w_2)$ the marginal distribution of G_2 with respect to the second wage. Then there are $\gamma_1 dG_1(w)$ single applications, $\gamma_2 dG_2^1(w)$ low applications, and $\gamma_2 dG_2^2(w)$ high applications at w, adding to a total measure $T(w) = \gamma_1 dG_1(w) + \gamma_2 dG_2^1(w) + \gamma_2 dG_2^2(w)$. Then

$$\hat{G}(w'|w) \equiv \sum_{j=1}^{2}\left[\frac{\gamma_2 dG_2^j(w)}{T(w)}\right]G_2^{-j}(w'|w),$$

where $-j \in \{1, 2\}/\{j\}$. Moreover, $\hat{g}(w'|w) \equiv \hat{G}(w'|w) - \lim_{w''\nearrow w'}\hat{G}(w''|w)$ is the size of a possible mass point of $\hat{G}(\cdot|w)$.

the measure of applications that firms receive:[18]

$$\gamma_1 G_1(w) + \gamma_2[G_2(w, 1) + G_2(1, w)] = v\int_0^w \lambda(\tilde{w})dF(\tilde{w}) \quad \forall w \in [0, 1]. \tag{1'}$$

The left-hand side denotes the mass of applications that are sent to firms with wages up to w. It is given by the probability that workers who send one application send it to firms with wages up to w and the probability that workers who send two applications send either their low- or their high-wage application to firms with wages up to w. These are dispersed over the firms that offer wages up to wage w. The mass of received applications is specified on the right-hand side. It is given by the ratio of applications per firm aggregated over all relevant wages, multiplied by the amount v of active firms in the market.

Equilibrium definition.—Equations (1′)–(5′) give the trading probabilities at wages that are offered in equilibrium, that is, at wages in $\mathcal{F}$. By (15) and (16) a worker can evaluate the optimal utility $U_i^* \equiv \sup_{\boldsymbol{w}\in F^i} U_i(\boldsymbol{w})$ of sending i applications to firms offering these wages, where $\boldsymbol{w}$ denotes the tuple of wages for which he applies. Then the highest utility that a worker can achieve with the optimal number of applications is $U^* \equiv \max_{i\in\{0,1,2\}} U_i^*$, which will be referred to as the *market utility*.

Similarly, every firm that offers a wage in $\mathcal{F}$ can assess the expected profit that it will get. For a firm that deviates and offers a wage w that is not offered by any other firm ($w \notin \mathcal{F}$), the effective queue length is not determined by (1′)–(5′) because no worker is currently applying for that wage. The firm has to anticipate how workers will apply if it offers that wage. Given some effective queue length $\mu(w)$ for the deviant firm, let $\hat{U}(w)$ be the highest utility that a worker can obtain by applying for w and possibly some other wage offered by some other firm.[19] Workers

[18] Queue length $\lambda(\cdot)$ is simply the Radon-Nikodym derivative, defined on $\mathbb{R}_+ \cup \{\infty\}$ to account for the case in which a zero measure of firms might receive a mass of applications.

[19] Given $\mu(w)$, a worker who applies for w with one application obtains utility $\hat{U}_1(w) = U_1(w)$ given by (15). A workers who applies for w and for some wage offered by another firm obtains at best

$$\hat{U}_2 = \max\left\{\sup_{\{\tilde{w}\in\mathcal{F}\mid \tilde{w}\le w\}} U_2(\tilde{w}, w), \sup_{\{\tilde{w}\in\mathcal{F}\mid \tilde{w}>w\}} U_2(w, \tilde{w})\right\},$$

where the two expressions in the max operator distinguish the case in which the other firm offers a lower wage from the case in which the other firm offers a higher wage. The maximum utility from applying to the deviant is then $\hat{U}(w) = \max_{i\in\{0,1,2\}} \hat{U}_i(w)$, where $\hat{U}_0(w) = 0$.

When we consider fixed search intensity $(\gamma_0, \gamma_1, \gamma_2)$ in lemma 3, we have to treat the application costs as sunk and allow utility combinations U_i^* and $\hat{U}_i$ only when some workers actually send i applications $(\gamma_i > 0)$.

will apply to the deviant firm in a way that exactly allows them to obtain the market utility that they could get by applying to other firms. That is, $\mu(w)$ *has to be such that* $\hat{U}(w) = U^*$ (if possible, otherwise $\mu(w) = 0$). If workers would get more utility at the deviant than at other firms, all workers would apply to the deviant, driving up its queue length. If workers would get less utility at the deviant, workers would not want to apply there, which lowers its queue length (but no further than zero). This specification of subgame perfection is known as the *market utility condition.*[20] It defines the effective queue length at all wages, and $\pi^* \equiv \sup_{w \in [0,1]} \pi(w)$ denotes the optimal profit that firms can obtain. An equilibrium can now be defined as follows.

DEFINITION 1. An equilibrium is a tuple $\{v, F, \boldsymbol{\gamma}, G_1, G_2\}$ such that there exists $\mu(\cdot)$ satisfying the following conditions:

1. Profit maximization and free entry:
   a. $\pi(w) = \pi^*$ for all $w \in \mathcal{F}$.
   b. $\pi^* = K$ if $v > 0$, and $\pi^* \leq K$ if $v = 0$.
2. Utility maximization for given search intensity and optimal choice of search intensity:
   a. For any $i \in \{1, 2\}$, $U_i(\boldsymbol{w}) \leq U_i^*$ for all $\boldsymbol{w} \in [0, 1]^i$, with equality for all $\boldsymbol{w} \in \text{supp } G_i$.
   b. $U_i^* = \max_{j \in \{0,1,2\}} U_j^*$ if $\gamma_i > 0$.
3. Consistency: $\mu(\cdot)$ is consistent with $(1')$–$(5')$ and fulfills the market utility condition.

Condition 1*a* specifies that firms set wages to maximize their profits. Condition 1*b* specifies the zero profit condition for free entry and ensures that firms abstain from the market only if they cannot earn positive profits. Condition 2*a* states that workers who send i applications send them optimally; that is, they get the highest utility on the support of their strategy and weakly less elsewhere. Condition 2*b* requires that workers send out the optimal number of applications. Condition 3 reiterates the conditions on the effective queue length. For conditions 1 and 2, the distinction between *a* and *b* allows the discussion of an exogenous number of applications and an exogenous number of firms using the appropriate subset of conditions.

The following subsections characterize the equilibrium properties of the model and show the following. *Equilibria exist and are constrained efficient. Generically, the following hold: The equilibrium is unique, all workers send the same number of applications, the number of offered wages equals the*

[20] This approach to subgame perfection is an axiomatic one. Papers by Peters (1991, 1997, 2000) and Burdett et al. (2001) rigorously establish equivalence of the market utility condition and the symmetric equilibrium among workers in the subgame after the wage announcements in (a limit of) finite economies in which workers send one application. For multiple-application models, such equivalence has not yet been proven.

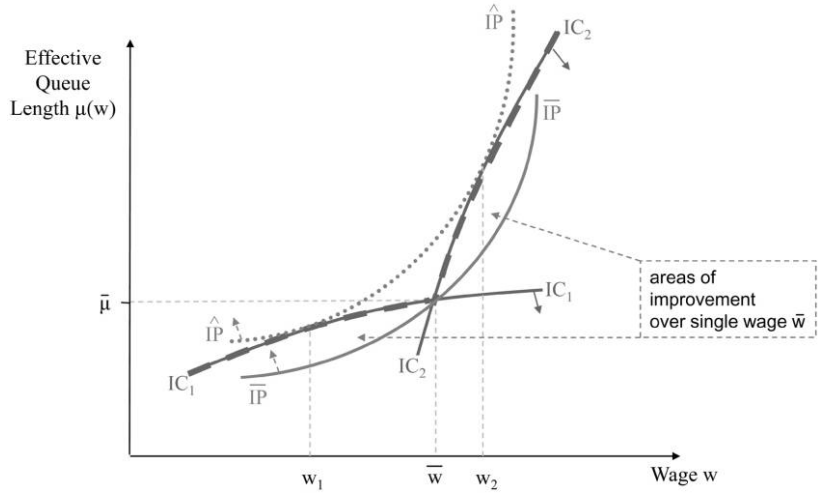

FIG. 1.—Illustration of market interaction, given (u_1, u_2). The curves IC_1 and IC_2 are workers' indifference curves for the low- and high-wage applications, respectively; $\overline{IP}$ is the iso-profit curve when all firms post the same wage; and $\widehat{IP}$ is the iso-profit curve in an equilibrium with two wages. Note that this depicts a special case; in general, the tuple (u_1, u_2) and, thus, curves IC_1 and IC_2, consistent with equilibrium condition 2*a*, might change when firms offer two wages rather than one.

number of applications that each worker sends, and each worker applies with one application to each wage. A brief graphical illustration is presented first to clarify the main working of the model and the main results. The formal proofs follow.

B. Graphical Illustration

In this model, workers and firms care about two things: the wage w and effective queue length μ. Therefore, the equilibrium interaction can be illustrated in the two-dimensional plane of figure 1 that resembles an Edgeworth box. Firms prefer points (w, μ) with a low wage w and a high effective queue length μ (northwest) and workers prefer the opposite (southeast).

Workers' preferences can best be represented by two sets of indifference curves. First, if they send a single application, their expected utility is given by (15). Recalling that $(1 - e^{-\mu})/\mu$ is the probability of being hired, we can write an indifference curve according to (15) as $[(1 - e^{-\mu})/\mu]w = u_1$. This describes all combinations of μ and w that yield net utility level u_1 for a low application (where u_1 is net of the application cost). The curve IC_1 in figure 1 corresponds to a particular value of u_1. Even workers who send two applications and obtain an expected

utility according to (16) have such indifference curves over the low wage w_1, because the part associated with the low wage is still the product of the wage times the probability of being hired.

Second, workers have a separate set of indifference curves for their high application. If they can obtain utility u_1 for their low application, then the indifference curve for w_2 according to (16) is

$$\frac{1 - e^{-\mu}}{\mu} w + \left(1 - \frac{1 - e^{-\mu}}{\mu}\right) u_1 = u_2.$$

The curve IC_2 in figure 1 describes all combinations of μ and w that give a particular net utility u_2 (net of application costs and conditional on the first application yielding u_1). In other words, a worker who obtains a point on IC_1 with one application and a point on IC_2 with his other application achieves net utility u_1 with his low application and net utility u_2 overall.

Note that IC_2 is steeper than IC_1 because workers who fail to get a high-wage job still anticipate a chance of getting a job at the low wage. The low-wage application induces a fallback option that makes workers more risky with their high-wage application.[21] Further, note that when all workers optimize according to equilibrium condition 2*a*, the resulting utilities u_1 and u_2 fully determine the queue lengths, which are given by the *upper contour* of the associated indifference curves IC_1 and IC_2 (the dashed line). If there would be a wage–queue length combination below the dashed line, workers could raise their utility beyond u_1 or u_2 by applying there.

Figure 1 also represents the iso-profit curves of the firms. Firms care about effective applications but not whether they come from high or low applications. Therefore, they have only one set of iso-profit curves, given by equation (14): $(1 - e^{-\mu})(1 - w) = \pi$. The curves $\overline{IP}$ and $\widehat{IP}$ depict such iso-profit curves for different profit levels $\bar{\pi}$ and $\hat{\pi}$. The problem of an individual firm can now be understood as follows: The firm anticipates that workers can obtain utilities (u_1, u_2) elsewhere, and it can therefore predict that the relationship between its wage offer and its effective queue length is given by the dashed line. Therefore, it looks for a point on the dashed line associated with the highest iso-profit curve.

From figure 1 one can infer the following three results: First, whenever some workers send multiple applications, there has to be wage dispersion. The reason is the following. If firms offer only a single wage, then

[21] Figure 1 is also related to models of on-the-job search, such as Delacroix and Shi (2005). Here different wage segments arise because workers who fail at the high wage still have the hope of getting a job at the low wage. In on-the-job search models, workers who fail at a higher wage still have their current job.

workers will send both applications there because of a lack of alternatives. The indifference curves IC_1 and IC_2 will adjust such that the offered wage is exactly at the intersection of both curves, indicated by point $(\bar{w}, \bar{\mu})$ in figure 1. The associated iso-profit curve for firms is $\overline{IP}$. Now it is easy to see that no firm would individually offer wage $\bar{w}$ at the "kink" of the dashed line. The "kink" induces a nonconvexity into the firm's problem, and an individual firm can obtain a higher level of utility by choosing a different wage, say w_2, and obtain the associated queue length given by the dashed line. This combination is on the higher iso-profit curve $\widehat{IP}$, indicating a profitable deviation from the unique market wage. Second, the figure highlights the inefficiencies that would be associated with a unique wage $\bar{w}$ (similar to lemma 2) since all points between the indifference curves IP_1 and IP_2 and the iso-profit curve $\overline{IP}$ would be strictly preferred by both workers and firms. The presence of these inefficiencies at a unique wage is exactly what induces firms to deviate and reduce these inefficiencies. Third, when there are two wages, w_1 and w_2, the associated iso-profit curve $\widehat{IP}$ can be tangent to the indifference curves, which sustains an equilibrium and yields efficiency (similar to lemma 1).[22]

C. *Equilibrium Analysis*

The formal analysis proceeds in three steps. First, the workers' search behavior for a given distribution of wages and a given number of applications is analyzed. Then the firms' wage-setting decisions are considered. Finally, entry of firms and the equilibrium number of applications are determined.

The workers' search decisions.—For the moment, fix the fraction γ_1 of workers who send one application, the fraction γ_2 who send two applications, and the number v of firms. All agents take as given the effective queue lengths $\mu(w)$ and associated trading probabilities $p(w)$ and $\eta(w)$ that arise in equilibrium and maximize their payoffs according to equilibrium condition 2*a*. Workers who send only one application solve the problem

$$\max_w p(w)w. \tag{17}$$

Let u_1 be the value of program (17). It is the highest expected net utility that a worker can generate with one application (net of application

[22] In general, the workers' indifference curves IC_1 and IC_2, consistent with worker optimization (equilibrium condition 2*a*), change when the wage distribution changes from one to two wage offers because utilities (u_1, u_2) will change. Under special conditions on the choice of the pooling wage and the number of firms (i.e., the entry cost), the utility levels remain constant, and such a special case is used for easier graphical representation.

cost). If the worker sends two applications, he solves

$$\max_{(w_1,w_2)} p(w_2)w_2 + [1 - p(w_2)]p(w_1)w_1, \tag{18}$$

where again the costs of sending the applications are omitted since they are fixed for a worker who is determined to send two applications. Clearly, $w_1 \leq w_2$ is necessary to solve (18). Since the terms related to wage w_2 enter the optimization problem only multiplicatively and additively, it is clear that the optimal choice of w_1 also solves program (17). Workers with two applications behave with their low application as workers with only one application. Therefore, program (18) reduces to

$$\max_{w_2} \{p(w_2)w_2 + [1 - p(w_2)]u_1\}. \tag{19}$$

Let u_2 be the value of program (19), that is, the highest expected net utility that workers with two applications can obtain. Since the combination of any wage w_2 that solves (19) and any w_1 that solves (17) together solve (18) but solving (18) requires $w_1 \leq w_2$, there must exist a wage $\bar{w}$ such that any solution to (19) is weakly below $\bar{w}$ and any solution to (17) is weakly above $\bar{w}$.

So far we considered only the optimization by individual workers. Now we use the fact that every worker optimizes according to the steps outlined above. Therefore, in the subgame after the wage announcements, workers will apply in a way that endogenously generates a cutoff wage $\bar{w}$ such that below this wage workers send their low application and above this wage they send their high application. By the market utility condition they do this for wages that are offered by many firms as well as for those that are offered only by an individual deviant. We therefore obtain the next proposition, which is formally proven in the Appendix.

PROPOSITION 1 (Workers' application behavior). In any equilibrium in which some workers send applications, that is, $\gamma_1 + \gamma_2 > 0$, the following conditions for the job-finding probability $p(w)$ hold, which by (4′) also determine $\mu(w)$:

$$p(w) = 1 \quad \forall\ w \in [0,\ u_1], \tag{20}$$

$$p(w)w = u_1 \quad \forall\ w \in [u_1,\ \min\{\bar{w},\ 1\}], \tag{21}$$

and

$$p(w)w + [1 - p(w)]u_1 = u_2 \quad \forall\ w \in [\bar{w},\ 1], \tag{22}$$

for some tuple (u_1, u_2) and $\bar{w} = u_1^2/(2u_1 - u_2)$. If no workers apply twice, that is, $\gamma_2 = 0$, then $u_2 = u_1 + c_2$.

The definition of $\bar{w}$ implies continuity of $p(w)$. It is worth pointing out that even if workers send only one application in equilibrium, if a firm would offer a very high wage, workers might be willing to send a second application there. Workers are just willing to do this if the marginal utility of the second application is exactly equal to the additional cost, that is, $u_2 - u_1 = c_2$. This case will be relevant whenever $\bar{w} \leq 1$; otherwise the interval $[\bar{w}, 1]$ is empty.

Firms' wage-setting decisions.—The firms anticipate the response by workers given in the previous proposition. It is relatively straightforward to show that as long as some workers send some application, the firms do not offer the extreme wages of zero or one, so $\bar{w} < 1$.[23] An individual firm maximizes $[1 - e^{-\mu(w)}](1 - w)$ subject to (20), (21), and (22), which describe the effective queue length at each wage. Since the effective queue length is strictly increasing in the wage, the constraints specify a one-to-one relationship when the queue length is strictly positive. Therefore, the firm can first choose the queue length it wants to obtain and then post the associated wage that induces workers to apply as desired. Using the constraints to substitute out the wage from the firm's objective, we can write the firm's profits as a function of the desired effective queue length, that is, $\Pi(\mu(w)) := \pi(w)$ with

$$\Pi(\mu) = 1 - e^{-\mu} - \mu u_1 \quad \forall\ \mu \in [0, \bar{\mu}] \tag{23}$$

and

$$\Pi(\mu) = (1 - e^{-\mu})(1 - u_1) - \mu(u_2 - u_1) \quad \forall\ \mu \in [\bar{\mu}, \mu(1)], \tag{24}$$

where $\bar{\mu} = \mu(\bar{w})$. The profit $\Pi(\mu)$ is continuous.[24] An individual firm chooses the queue length μ that maximizes its profits. Since $\Pi(\mu)$ is not concave at $\bar{\mu}$, this cannot be an optimal choice. In equilibrium the queue length has to coincide with the ratio of effective applications to firms by equilibrium condition 3, and since $\gamma_2 > 0$, this is possible only if v_1 and v_2 are strictly positive. Thus, if all firms choose the same queue length (and associated same wage), $\bar{\mu}$ is the only choice that is in both $[0, \bar{\mu}]$ and $[\bar{\mu}, 1]$, which reflects the fact that a unique wage attracts both high and low applications. Since $\bar{\mu}$ is not optimal for firms, this

[23] If all firms offer wage $w = 0$, the hiring probability is less than one, but any slightly higher wage has a hiring probability of one because all workers apply there (formally, the market utility condition implies an effective queue length of infinity), yielding a profitable deviation. If all firms offer wage $w = 1$, still $u_1 < 1$ because workers are not hired for sure, and by proposition 1 there are wages below one at which the queue length is positive, yielding a profitable deviation.

[24] This follows from the continuity of $\mu(w)$ or directly since $\bar{\mu} = \mu(\bar{w})$ is characterized by (21) and (22) as the solution to $[(1 - e^{-\mu})/\bar{\mu}][u_1^2/(2u_1 - u_2)] = u_1$.

formally proves the result on wage dispersion that was already discussed in connection with figure 1.

PROPOSITION 2 (Wage dispersion). In any equilibrium with $\gamma_2 > 0$, the set of offered wages $\mathcal{F}$ cannot be a singleton.

Note that the driving force for wage dispersion is not a profit discontinuity as in related work, but rather a nondifferentiability of the profits.[25] Since there is wage dispersion, there have to be some wages that are not at the extremes of zero and one. Therefore, u_1 and u_2 are in the interior of $(0, 1)$. Then the optimal solution cannot be a boundary, but is either in $(0, \bar{\mu})$ or in $(\bar{\mu}, 1)$. Consider high-wage firms offering a wage that induces a queue in $(\bar{\mu}, 1)$ first. The first-order condition of their profits in (24) is

$$e^{-\mu}(1 - u_1) = u_2 - u_1, \tag{25}$$

which uniquely defines the optimal queue length, call it μ_2. Therefore, all firms in this region choose the same queue length and offer the same associated wage, call it w_2. Let v_2 denote the measure of high-wage firms. Since these high-wage firms receive the high application of every worker, every application is effective. At this point there are γ_2 workers who apply, and so the effective and gross queue lengths are $\mu_2 = \lambda_2 = \gamma_2/v_2$, and u_2 is fully determined by (25) once v_2 and u_1 are known.

Consider now low-wage firms that offer a wage that induces a queue in $(0, \bar{\mu})$. The first-order condition to their profits in (23) is

$$e^{-\mu} = u_1. \tag{26}$$

Again, all firms in this region have the same optimal queue length μ_1 and offer the same associated wage w_1. Their effective applicants are only those who are left after the high-wage firms have hired. When v_1 firms offer wage w_1, there are $\lambda_1 = (\gamma_1 + \gamma_2)/v_1$ applications per firm but only $\mu_1 = (\gamma_1 + p_2\gamma_2)/v_1$ effective applicants, where $p_2 = (1 - e^{-\mu_2})/\mu_2$ is the probability of being hired at the high wage. Given v_1 and

[25] In most work on wage dispersion with homogeneous agents, a unique wage cannot be sustained because of the following discontinuity: a firm's offer at a slightly higher wage is accepted for sure, whereas at the market wage there is a chance that a worker would rather accept an equally good offer from a competing firm (see Burdett and Judd 1983; Burdett and Mortensen 1998; Acemoğlu and Shimer 2000; Gautier and Moraga-González 2005; Galenianos and Kircher 2008). In this model, a worker anticipates that at a higher wage other workers accept with a higher probability, which means that other applications are effective with a higher probability. This is bad for the worker. Thus, he applies only if the overall number of (gross) applications goes *down*. This reduction in applications smooths out the discontinuity in acceptances. Dispersion arises nevertheless because workers trade off the wage and the queue length differently for high and low applications, leading to a discontinuity in the derivative (i.e., the "kink"), which suffices to induce dispersion.

v_2, μ_2 is uniquely determined, and so is u_1 by (26). Substituting (26) and (25) into (23) and (24) yields the following proposition.

PROPOSITION 3 (Profits and wages). In an equilibrium with $\gamma_2 > 0$, profits and wages for low- and high-wage firms, respectively, are uniquely determined by γ_1, γ_2, v_1, and v_2 as

$$\Pi_1 = 1 - e^{-\mu_1} - \mu_1 e^{-\mu_1}, \tag{27}$$

$$w_1 = \frac{\mu_1 e^{-\mu_1}}{1 - e^{-\mu_1}}, \tag{28}$$

$$\Pi_2 = (1 - e^{-\mu_2} - \mu_2 e^{-\mu_2})(1 - e^{-\mu_1}), \tag{29}$$

and

$$w_2 = \frac{\mu_2 e^{-\mu_2}(1 - e^{-\mu_1})}{1 - e^{-\mu_2}} + e^{-\mu_1}. \tag{30}$$

Profits and wages have a natural interpretation: The profit of a firm reflects its marginal contribution to matching, as explained already in the discussion of the search efficiency condition (8). Similarly, the formulas for the wages can be understood by considering the incentives that they provide for the workers. The low wage coincides with the probability $e^{-\mu_1}$ that an additional application for this wage reaches a firm that has no other effective applicant, conditional on the event that a worker actually gets hired at this wage, which happens with probability $p_1 = (1 - e^{-\mu_1})/\mu_1$. The conditioning is important because the decision to send an application is influenced by the product of hiring probability and wage, and so $p_1 w_1$ provides incentives that coincide with the social benefit. At the high wage, the social benefit of sending another application is the probability $e^{-\mu_2}$ that it reaches a high-wage firm that does not have another applicant, adjusted by the probability $1 - e^{-\mu_1}$ that the low-wage firm to which the worker also applied has at least one other applicant (otherwise no new match is created). The high wage reflects this social benefit (again conditional on getting matched to provide the right incentives) and adds a term reflecting the social benefit of the low application. The reason for the added term is that conditional on getting the high-wage job, the worker is precluded from reaping the benefits from his application for the low-wage job and has to be compensated for this.

If no worker sends two applications, then the arguments above easily establish that only one wage is offered ($v_2 = 0$). The determination of the wage is identical to the low wage in the previous proposition; that

is, it is determined by (28) and induces profits given by (27). Given that workers apply only once, any offer leads to a hire, and therefore $\mu_1 = \lambda_1$.

Equilibrium outcomes.—Before we turn to the full equilibrium, it is instructive to consider the case with exogenous search intensity $\boldsymbol{\gamma}$. The following lemma will cover the case with and without free entry.

Consider the case in which some workers apply twice. Firms offer two wages but have to make equal profits. Equating profits Π_1 and Π_2 in (27) and (29), we obtain exactly the same allocation as in the planners' optimum in (8). If the number of firms is not fixed, free entry implies

$$1 - e^{-\mu_1} - \mu_1 e^{-\mu_1} = K \tag{31}$$

and

$$(1 - e^{-\mu_2} - \mu_2 e^{-\mu_2})(1 - e^{-\mu_1}) = K, \tag{32}$$

which coincide with conditions (10) and (11) that uniquely determine the optimal entry level.[26] We get the following efficiency results when taking the search intensity of workers as given, which include the well-known results for one-application models as the special case of $\gamma_1 = 1$.

LEMMA 3. Given search intensity $(\gamma_0, \gamma_1, \gamma_2)$, with $\gamma_1 + \gamma_2 > 0$, the following conditions hold:

1. Given entry $v > 0$, there exists unique tuple (F, G_1, G_2) that fulfills the appropriate equilibrium conditions 1*a*, 2*a*, and 3, and it yields the search efficient number of matches.
2. With free entry, there exists unique tuple (v, F, G_1, G_2) that fulfills the appropriate equilibrium conditions 1*a*, 1*b*, 2a, and 3, and it yields optimal entry.

Now consider the equilibrium when the fraction of agents that send zero, one, or two applications is endogenous. The free-entry conditions (31) and (32) determine the effective queue lengths at the high- and low-wage firms solely as a function of the exogenous entry cost K. Call the solutions to these equations μ_1^* and μ_2^*. Then we can define numbers $u_1^* = e^{-\mu_1^*}$ and $u_2^* - u_1^* = e^{-\mu_2^*}(1 - u_1^*)$ in analogy to the first-order conditions (26) and (25). The numbers u_1^* and $u_2^* - u_1^*$ are, respectively, the marginal utility gains for workers for the first and second applications. They are independent of the exact structure of search intensity $\boldsymbol{\gamma}$, as the exact level of applications is offset through free entry. The

[26] By construction, there are not profitable deviations from this profile: firms are willing to offer these wages since the wages were determined by the appropriate first-order conditions, and in each region the maximization problem is concave. Workers are willing to apply in this fashion because we used their indifference conditions (21) and (22) to construct firm profits.

equilibrium is determined by comparing these marginal gains with the marginal costs of sending the application.

PROPOSITION 4 (Equilibrium outcomes). An equilibrium exists and is constrained efficient. Furthermore, the following conditions hold:

1. For $c_1 > u_1^*$, in the unique equilibrium, no firm enters and no application is sent.
2. For $c_1 < u_1^*$ and $c_2 > u_2^* - u_1^*$, in the unique equilibrium, all workers send one application and one wage is offered.
3. For $c_1 < u_1^*$ and $c_2 < u_2^* - u_1^*$, in the unique equilibrium, all workers send two applications, two wages are offered, and each worker applies to one firm offering each wage.

The key reason for uniqueness is that firms anticipate that workers will send additional applications when they offer high wages (this is captured by the market utility condition). Even if in (a candidate) equilibrium only one wage is offered and workers send only one application, a firm that deviates and offers a sufficiently high wage expects that workers will send a second application for this very high wage. It is this feature that leads to a high queue length for a deviant with a high wage and makes such a deviation profitable whenever the marginal cost c_2 is below the marginal benefit $u_2^* - u_1^*$.

Efficiency follows trivially because the marginal utilities u_1^* and $u_2^* - u_1^*$ equal the planner's thresholds t_1^P and t_2^P. This arises since the wages provide the socially efficient application incentives. For completeness, note that in the case in which $c_1 = u_1^*$, we have a continuum of equilibria: for any $\gamma_1 \in [0, 1]$ and $\gamma_0 = 1 - \gamma_1$, an equilibrium exists, and workers are exactly indifferent between applying once and not applying at all. If $c_2 = u_2^* - u_1^*$, an equilibrium exists in which workers randomize between one and two applications, that is, $\gamma_2 \in [0, 1]$ and $\gamma_1 = 1 - \gamma_2$.

Connection to one-application models.—To round off the equilibrium section, it may be instructive to briefly link the results to standard one-application models. When workers apply only once ($\gamma_2 = 0$, which arises when c_2 is large), the distinction between effective and gross queue length disappears, and wages (28) and profits (27) replicate the standard result from one-application models such as that of Burdett et al. (2001). The introduction of a second application changes the equilibrium and essentially generates two markets. The profits in each are given by

$$\Pi_i = (1 - e^{-\mu_i} - \mu_i e^{-\mu_i})(1 - u_{i-1}). \tag{33}$$

In the low-wage market ($i = 1$), u_0 is identical to the workers' true outside option of zero, but there is some connection to the high market induced by the strictly positive probability that an offer is rejected. In the high-wage market ($i = 2$), the rejection probability is zero, but u_1

is greater than zero since it reflects the workers' endogenous outside option induced by the low-wage market. Apart from these spillovers, each market operates as a single one-application market.

This separation into semiseparate markets arises because of a close resemblance of our setup to a sequential market with two periods:[27] A firm that pays K can choose one of the periods and post a wage. Workers who pay $c(2)$ can participate in both periods, and workers who pay $c(1)$ participate in the second period. Workers who fail to get a job in the first period try again in the second. In the second period workers have nothing to wait for and maximize according to (17), whereas in the first period they anticipate the benefit u_1 from waiting and maximize according to (19). The main difference from our model is that in this sequential setup a first-period firm can attract applicants at wages below $\bar{w}$ whereas in our setup such wages would attract only low applications (i.e., second-period applications). The reason is that workers from the second period would drive up the queue length if they could apply to such a firm.[28] Nevertheless, a firm that wanted to offer such low wages is better off waiting for the second period, and in equilibrium the outcome of the two economies coincides. Our interaction therefore resembles the nonstationary sequential search in Peters (1991), and efficiency is driven by reasons similar to those in standard sequential one-application interactions (even though I am not aware of a formal efficiency proof for the nonstationary environment).[29]

V. Generalization to $N > 2$ Applications

In this section the restriction that workers can send no more than two applications will be dispensed with. This allows us to consider the limit as search costs vanish and to examine whether the equilibrium converges to the unconstrained efficient Walrasian outcome, as opposed to the constrained efficient social planner's outcome. This limit is of interest because, typically, search models that approach Walrasian outcomes rely on repeated interactions, whereas most static search models converge to non-Walrasian outcomes even when costs vanish (e.g., Acemoğlu and Shimer 2000; Albrecht et al. 2006). The main reasons why convergence might not be immediate is that the miscoordination and tie-breaking frictions may not vanish and that the total application costs may not converge to zero. The latter could happen if the vanishing costs are

[27] I thank the editor for suggesting this connection to me.

[28] Since both applications can be sent to firms with low wages, our cutoff wage is $\bar{w} = u_1^2/(2u_1 - u_2)$ (see proposition 1). In a sequential interaction, a first-period firm gets applications as long as the wage is larger than $\hat{w} = u_2$.

[29] In fact, because of the equivalence of the equilibrium outcomes, this proof can be regarded as an efficiency proof for nonstationary sequential directed search environments.

offset by an increasing number of applications that workers send. In the following it will be shown that convergence to the Walrasian benchmark does arise in this model.

Additionally, even away from the limit it might be interesting to have a simple representation of the search intensity and resulting matches. It has been argued that explicit forms of search intensity based on simultaneous search can be useful to understand the response of workers to varying labor market conditions (e.g., Shimer 2004), and we will see that the formulation in a directed search framework like this one remains particularly tractable. This framework is thus potentially useful for the study of such wider questions.

For the analysis, recall that $c(i) > 0$ for some $i \in \mathbb{N}$, which together with increasing marginal costs implies a largest integer, denoted N, such that $c(N) \leq 1$. Clearly, it is neither individually nor socially optimal to send more than N applications because the application costs would exceed the value created in a match. Thus, the number of applications per individual remains bounded. Apart from the larger number of possible applications that each worker is allowed to send, the model remains unchanged. The exact equilibrium definition for this extended setup is presented in the Appendix.

The equilibrium characterization extends by analogy to the previous section, as shown in the Appendix. The workers again partition the wages into intervals related to each of their applications. The equilibrium interaction in each interval corresponds to that in the one-application case, again with the adjustment that the workers' "outside option" incorporates the expected utility that can be obtained at lower wages, whereas the queue length incorporates the fact that some applicants are lost to higher-wage firms. Again, the free-entry condition defines the effective queue length in the various wage segments that arises in equilibrium. In particular,

$$(1 - e^{-\mu_i^*} - \mu_i^* e^{-\mu_i^*})(1 - u_{i-1}^*) = K \tag{34}$$

and

$$u_i^* = e^{-\mu_i^*}(1 - u_{i-1}^*) + u_{i-1}^* \tag{35}$$

recursively define the effective queue length μ_i^* and marginal utility u_i^* at the ith-highest wage as a function of entry cost K, given initial condition $u_0^* = 0$. The wages are completely determined by these variables.[30] Equation (34) equates the competitive profits to the entry cost in analogy to equations (31) and (32), and equation (35) captures the

[30] In an extension of proposition 3, one can show that the wages are now given by $w_i = [\mu_i^* e^{-\mu_i^*}/(1 - e^{-\mu_i^*})](1 - u_{i-1}^*) + u_{i-1}^*$. In equilibrium when all workers send i^* applications, the measure v_i of firms that offer wage w_i is given by $\mu_i^* = [\prod_{j=i+1}^{i^*}(1 - p_j)]/v_i$, where $p_i = (1 - e^{-\mu_i^*})/\mu_i^*$.

outside option that lower wage segments induce for higher wage segments in analogy to equations (25) and (26). Note that $u_i^* - u_{i-1}^*$ is strictly decreasing in i and c_i is weakly increasing. The following proposition is proved in the Appendix.

PROPOSITION 5 (Generalized equilibrium properties). An equilibrium exists and is constrained efficient. It is generically unique: if $c_{i^*} < u_{i^*}^* - u_{i^*-1}^*$ and $c_{i^*+1} > u_{i^*+1}^* - u_{i^*}^*$, every workers sends i^* applications, i^* wages are offered, and every worker applies to each wage.

The proof relies inductively on the arguments presented in Sections III and IV, extended to higher numbers of applications. Efficiency obtains again for similar reasons, the only difference being that now i^* wages are necessary to obtain the optimal allocation in the search process.

Convergence to the competitive outcome.—Now consider the case in which application costs become small. In the following it will be shown that the equilibrium allocation converges to the unconstrained efficient allocation of a competitive economy. A competitive Walrasian economy achieves the following benchmark allocation: Since entry costs are below the productivity of a match, firms enter until the measure of firms equals the measure of workers, that is, all workers and all firms get matched. Since free entry places firms on the long side of the market, firms are just compensated for their entry cost K. The market wage is then $1 - K$ and coincides with the utility of each worker.

Consider a sequence of cost functions such that the marginal cost of the ith application converges pointwise to zero for all $i \in \mathbb{N}$. Rather than looking at these functions directly, it is convenient to simply consider the associated equilibrium number i^* of applications that each worker sends.[31] Vanishing costs amount to $i^* \to \infty$. Let $v(i^*)$ denote the equilibrium measure of active firms, and $\eta(i^*) = \int \eta(w)dF$ and $\sigma(i^*) = v(i^*)\eta(i^*)$ denote the average probability of being matched for both a firm and a worker in the economy, respectively. Let $w(i^*)$ denote the average wage, conditional on being matched, and $U^*(i^*) = u_{i^*}^* - c^{i^*}(i^*)$ the equilibrium utility when i^* applications are sent, where $c^{i^*}(\cdot)$ denotes the cost function that supports the equilibrium with i^* applications per worker. In the Appendix the following proposition is proved.

PROPOSITION 6 (Convergence). The equilibrium outcome converges to the competitive outcome, that is,

[31] For the (nongeneric) case of multiple equilibria, select for simplicity one in which all workers send the same number of applications; and when multiple cost functions sustain the same equilibrium number of applications, consider a subsequence of cost functions that includes only one of these.

$$\lim_{i^*\to\infty} v(i^*) = 1,$$

$$\lim_{i^*\to\infty} \eta(i^*) = \lim_{i^*\to\infty} \sigma(i^*) = 1,$$

$$\lim_{i^*\to\infty} w(i^*) = \lim_{i^*\to\infty} U^*(i^*) = 1 - K.$$

The structure of the proof uses the intuition for the competitive economy: For a given measure v of active firms, the competitive economy implies that (only) the long side of the market gets rationed and the short side appropriates all surplus. For small frictions (i^* large) this still holds approximately. Then it trivially follows that $v(i^*) \to 1$ because otherwise the firms generate either too much or too little profits to cover entry. Since nearly all agents get matched, zero profits imply a wage of $1 - K$. Despite the fact that workers send more applications, their application costs vanish faster than the increase in the number of applications, and therefore their utility equals the wage of $1 - K$ in the limit.

It may be instructive to point out how the equilibrium converges to the competitive solution. The equilibrium wages w_i are implicitly determined by (34) and (35) without regard to the application costs. Therefore, the wages w_1 and w_2 that were derived in Section IV are still offered even as workers send more and more applications, and also their effective queue lengths remain the same. But as workers apply more often, higher wages are added. It becomes more and more likely that a worker gets hired at high wages, which means that the same effective queue length at a low wage requires only very few firms to offer this wage because only very few workers will be effective at this wage. The support of the wages does not converge, yet the mass of agents that trade at wages away from $1 - K$ converges to zero.

VI. Discussion

In the introduction it is mentioned that other simultaneous search environments do not obtain efficiency. This section highlights the driving forces toward efficiency in the model and discusses how these differ relative to the earlier literature.

Convergence to the Walrasian benchmark.—In most search models, adding an additional worker to the economy makes it harder for other workers to find a job and easier for the firms to hire a worker. The same is true in this setup when a worker sends two applications instead of one. When application costs vanish and many more applications are sent, the process becomes so efficient that in the limit the short side of the market gets matched for sure. This is a minimal requirement for convergence

toward the efficient outcome. The right incentives for efficient entry of firms then follow from the competitive nature of the wage setting.

In the models by Gautier and Moraga-González (2005), Albrecht et al. (2006), and Galenianos and Kircher (2008), it may become harder for firms to hire when workers send a second application. In their environments every firm makes only one offer, so if some worker sends two applications, two firms might both offer a job to him and only one can hire him even if both had other applicants. This introduces nonnegligible probabilities that firms cannot hire. Therefore, the short side of the market does not get perfectly matched even in the limit, killing the central driving force toward the Walrasian benchmark. In Acemoğlu and Shimer (2000), workers can sample information about multiple wages but can apply to only one firm, again precluding convergence to the case in which the short side gets matched for sure.

Efficient "pricing" under nonnegligible costs.—Constrained efficiency in this model arises because firms can "price" the good they are interested in: the hiring probability, which is determined by the queue of effective applications. They can "price" the effective applications despite the problem that they cannot pay the workers directly for their applications to other firms, which was mentioned as "contractual incompleteness" in the introduction. This incompleteness is solved because workers internalize the actions of other workers: Workers care only about rival applicants who are effective, that is, those who do not get jobs at higher wages. Only such applicants make it difficult to get a job at any wage. If some workers who apply to firm A also apply to a better firm B, other workers anticipate this and apply more to firm A because they understand that (because of the applications to B) there is now less competition for A's job. In general, if a firm changes its wage, workers change their application behavior such that the queue of effective applications rises until they are indifferent between this wage and the other wages offered in the market. This allows firms to price the effective applications at the margin and to achieve efficiency.

It might be surprising that rent seeking does not cause inefficiencies in this setup. Rent seeking arises when workers send more applications just to get a better wage even when the productivity of the jobs does not differ. It is pervasive in models with wage dispersion (see, e.g., Acemoğlu and Shimer 2000; Gautier, Moraga-González, and Wolthoff 2008). It arises when search is not fully directed, that is, wages are not publicly posted. The missing competition leads to wages that do not reflect the marginal value of an application. In contrast, in the setting analyzed here, wages are posted and do reflect the marginal value of an application. As mentioned at the end of Section IV, one can interpret our market interaction as a sequential process, where firms and workers at high wages trade first, and those workers who did not get matched

proceed to the next period, where matching occurs at the next-lower wage. In each period the workers take into account their option of matching later, and the wage setting "prices" their applications correctly, taking into account the waiting option.[32]

In earlier directed search models by Albrecht et al. (2006) and Galenianos and Kircher (2008), firms are not able to "price" their hiring probability because of the lack of stability in the final matching. Recall that a lack of stability means that a firm might not be able to hire any worker even though one of its workers obtains only a lower-wage job or remains unemployed. Firms still want to price effective applications, but in these models the workers regard the wage as the price for gross applications. Any gross application has an equal chance of getting an offer, and even if the worker decides to work elsewhere, the job is "blocked" because the firms can make only a single offer. So workers view any gross application as competition. If a firm raises its wage, workers raise their gross applications until they are indifferent to applying elsewhere. Workers might also change their application behavior to *other* firms. Since this second effect is not directly influenced by the offered wage, the firms cannot price their hiring probability (the effective applications) and efficiency does not obtain. In terms of the sequential interpretation, workers who get jobs already in early rounds at high-wage firms still apply in later periods at low-wage firms, which disturbs the pricing at these lower-wage firms.

This distinction between gross and effective applications has not been considered in the literature because the two notions coincide in models in which workers have only one application: With one application no worker rejects an offer, and every application is effective.[33] Our model includes the one-application environment when the first application is free but all others are prohibitively expensive. More important, even with multiple applications, each wage segment is similar to the one-application model. In particular, the Hosios (1990) condition, which ties the division of the match surplus to the elasticity of the matching function and ensures that private and social surplus in the market coincide, holds per wage segment once the appropriate notion of effective applications is applied (see eq. [33]). Therefore, earlier results on efficient entry in one-application models naturally extend to our setup.

Two main assumptions: i. Perfect recall.—In the interpretation of the

[32] Note that the real "price" is the product of the wage and the probability of being hired. This product is lower at high wages. This reflects the fact that the social value of sending two applications instead of one is lower than the social value of sending one application instead of none.

[33] This also holds for models that incorporate search intensity in a more reduced-form way through a scalar that increases the matching probability but yields at most one job at a time (see, e.g., Moen 1995).

final matching stage as an extensive form process, one might ask how many applicants a firm can possibly contact. The model requires that a firm can potentially contact all of them. If a firm can contact only a limited number M of applicants, it bears the risk that they all are hired elsewhere. Again, firms care about effective applicants who would accept the job, but workers anticipate that even noneffective applicants can "block" a job if there are more than M of them. Since pricing works through the workers' reaction, firms are not able to price their hiring probability efficiently, similar to the reason discussed above for the $M = 1$ case of Albrecht et al. (2006) and Galenianos and Kircher (2008). For larger M, we expect the inefficiencies to be less severe, but full efficiency is likely to arise only when all applicants can be contacted.

ii. Wage commitment.—Wage commitment is important because it determines the terms of trade in a competitive manner before the market splits into small groups in which market forces are absent. In the theoretical analysis, wage commitment has the additional advantage of allowing for a particularly tractable notion of stability, enabling us to analytically study the earlier stages of firm entry, wage setting, and application behavior. The assumption of wage commitment might be particularly applicable in markets in which large firms adopt wage policies uniformly for the entire organization rather than adjust them to each individual bargaining situation. It seems also reasonable when the "wage" offers are interpreted more broadly as investments into prestige or amenities that are valued by the workers but cannot easily be adapted. It also applies when working conditions, including wages, are specified in advance of the application and matching process, as in the U.S. market for medical residents.[34]

Since other environments might not feature full commitment, it would be interesting to investigate forms of adjustments of the wage announcement in response to the application conditions that each firm and worker face. In particular, a low-wage firm that attracts several workers might be happy that it offered a low wage because it is likely that one of its workers gets no higher offer, whereas a low-wage firm that attracts only a single worker might prefer to raise its offer to a higher level to be sure that the worker accepts. Similarly, in an extensive form

[34] Roth (1984, 995–96) points out that in this market hospitals specify the job requirements and wages; then residents have to apply to and interview with hospitals; finally, both sides of the market submit rankings of those partners *with whom they have interviewed* to a clearinghouse that produces a stable matching. While this paper is targeted to general job markets and might not reflect the details of this specific market—in particular, it neglects heterogeneities—it highlights the efficiency even with a decentralized application procedure.

Bulow and Levin (2006) analyze price setting prior to matching in a world with one-sided heterogeneity but abstract from decentralized applications by assuming that all residents apply to all hospitals.

in which firms contact their workers sequentially, a firm might want to raise its wage offer once it approaches the last worker in order to increase the odds that the worker accepts. Weakening commitment by introducing methods for adjusting the terms of trade after the applications are sent might change the results regarding efficiency.[35] Especially if the probability that wages are adjusted upward is not balanced by proportionally lower wage offers ex ante, then entry of firms will not be efficient. Unfortunately, wage adjustments introduce strategic elements into the final matching stage, and the feedback to earlier stages renders such analysis beyond the scope of this paper.

Heterogeneity.—So far this paper abstracted from heterogeneities. While perfect recall and commitment seem essential for efficiency, homogeneity is clearly not essential. Consider the easiest case in which workers differ in their cost of sending applications. Equations (31) and (32) still describe the queue lengths due to free entry, so the marginal benefit for each worker still coincides with his social benefit; that is, application behavior remains constrained efficient. While so far it was assumed that firms produce $x = 1$ unit of output, consider now the case in which firms may choose to enter at different levels of productivity x at some cost $K(x)$ that is increasing and convex as in Acemoğlu and Shimer (1999) or Shi (2001). Equations (31) and (32) for optimal entry can be easily adjusted to this case, and efficiency is relatively straightforward to establish.[36] Firms at the high wage invest in more capital because they are more likely to utilize it. Also, if the workers are heterogeneous as in Shi (2001) or Shimer (2005) and wages can be conditioned on productivity, it seems likely that efficiency still arises because the interaction in each wage segment closely resembles a one-application economy and is likely to inherit its efficiency properties even in broader environments.

[35] In Albrecht et al. (2006), firms' wage announcements are not complete commitments, but firms bid up the wage to the worker's marginal product if two firms make an offer to the same worker. This has the effect that workers are mainly interested in two offers even if the announced wages are very low, which contributes to depressed wage offers and excessive entry.

[36] The workers' problem is still unchanged, and given utility levels u_1 and u_2, a low-wage firm solves $\max_{w,\mu,x}(1 - e^{-\mu})(x - w) - K(x)$ subject to the workers' response $w(1 - e^{-\mu})/\mu = u_1$. Follow the steps leading to (23) to substitute out the constraint, which yields the problem $\max_{\mu,x}(1 - e^{-\mu})x - \mu u_1 - K(x)$. This has first-order conditions $u_1 = xe^{-\mu}$ and $1 - e^{-\mu} = K'(x)$ under appropriate Inada conditions on $K(\cdot)$. Substituting the first into the objective gives free-entry condition $x_1^*(1 - e^{-\mu_1^*} - \mu_1^* e^{-\mu_1^*}) = K(x_1^*)$, which is similar to (31), which together with $1 - e^{-\mu_1^*} = K'(x_1^*)$ characterizes the queue length μ_1^* and productivity x_1^* in the low wage segment uniquely. Similar logic for the high wage segment yields a queue length μ_2^* and productivity x_2^* characterized by $(1 - e^{-\mu_2^*} - \mu_2^* e^{-\mu_2^*})(x_2^* - x_1^* e^{-\mu_2^*}) = K(x_2^*)$ similar to (32) together with $1 - e^{-\mu_2^*} = K'(x_2^*)$. The problem of a planner who has control over the type of entry has the same first-order conditions.

VII. Conclusion

This paper incorporates a microfoundation for search intensity into the directed search framework by allowing workers to apply to multiple firms simultaneously. Application choices can be interpreted as the strategic but frictional formation of links in a network between workers and firms,[37] and the number of links of each worker captures his search intensity. We resolve the assignment of jobs as a stable matching on the network, which allows the wages to internalize all the externalities that are present in earlier stages even when application costs are nonnegligible. The economy converges to the unconstrained efficient Walrasian outcome as application costs vanish.

Appendix

Before proceeding to the main proofs, it will be convenient to prove an auxiliary result for the social planner's problem. Consider some tuple $\{F, G_1, G_2\}$ describing the assigned distribution of firms and applications on the set L of locations. Now consider some other tuple $\{F', G_1', G_2'\}$ on set $L' = L \cup \{l'\}$ that has exactly one additional location. Let l be adjacent to l' in the sense that L' has no elements between l and l'.

We can say that $\{F, G_1, G_2\}$ is generated from $\{F', G_1', G_2'\}$ by *merging* l and l' if $\{F, G_1, G_2\}$ is identical to $\{F, G_1', G_2'\}$ except that all firms that were assigned to l' under F' are assigned to l under F and all applications that were assigned to l' under (G_1', G_2') are assigned to l under (G_1, G_2).[38] The following lemma states that merging l and l' does not change the overall number of matches if l and l' have identical queue length under $\{F', G'\}$.

LEMMA A1. For given v and $\boldsymbol{\gamma}$, if $\{F, G_1, G_2\}$ is generated from $\{F', G_1', G_2'\}$ by merging l and l' and if both l and l' have the same effective queue length $\mu_l' = \mu_{l'}'$ under $\{F', G_1', G_2'\}$, then the effective queue length is identical to that under $\{F, G_1, G_2\}$ (i.e., $\mu_l = \mu_l' = \mu_{l'}'$) and the total number of matches in the economy is identical under $\{F, G_1, G_2\}$ and under $\{F', G_1', G_2'\}$.

Proof. By construction, all queue lengths at locations $l'' \in L$ such that $l'' > l$ are identical under $\{F, G_1, G_2\}$ and $\{F', G_1', G_2'\}$ since the number of firms and the application behavior are unchanged (and the matching at higher locations

[37] The large literature in network formation usually deploys solution concepts that eliminate any randomness on which a frictional nature of unemployment could be based because they require that no (pair of) individuals would choose links differently ex post after the network has been realized (see, e.g., Dutta and Jackson 2000; Jackson 2005). Our equilibrium notion requires only ex ante optimality before applications are sent out and uses miscoordination between workers in their link formation to retain frictions in the spirit of the search literature on unemployment.

[38] That is, the mass points f and g coincide at all (combination of) locations that do not involve l or l' with f' and g', i.e., $f(l_a) = f'(l_a)$, $g_1(l_a) = g_1'(l_a)$, and $g_2(l_a, l_b) = g_2'(l_a, l_b)$ for all $l_a, l_b \in L'/\{l, l'\}$. At l, the mass points are defined by $f(l) = f'(l) + f'(l')$, $g_1(l) = g_1'(l) + g_1'(l')$, $g_2(l, l) = g_2'(l, l) + g_2'(l, l') + g_2'(l', l) + g_2'(l', l')$, $g_2(l, l_b) = g_2'(l, l_b) + g_2'(l', l_b)$ when $l_b \notin \{l, l'\}$, and $g_2(l_a, l) = g_2'(l_a, l) + g_2'(l_a, l')$ when $l_a \notin \{l, l'\}$.

is not influenced by the application behavior at lower locations). Now consider the matching at all locations in L that are weakly larger than either l or l'. Assume that $\mu'_l = \mu'_{l'} > \mu_l$. That means that among the firms at these locations, under μ' strictly more get matched than under μ. On the other hand, it becomes strictly harder for workers to get an offer, and strictly fewer workers get matched at these locations. Since workers and firms are matched in pairs, we cannot obtain more matches for firms than for workers. Similarly, $\mu'_l = \mu'_{l'} < \mu_l$ can be ruled out. Therefore, $\mu'_l = \mu'_{l'} = \mu_l$.

That means that we have not changed the overall matching probabilities at any of the locations, because firms at a location below $l-1$ face exactly the same probabilities that their workers take other jobs as under the original assignment. Since the matching probabilities are not changed, the number of matches is identical. QED

Proof of Lemma 1

The result is proved via a variational argument. For fixed v and $\boldsymbol{\gamma}$, consider some optimal set of locations L and distributions $\{F, G_1, G_2\} \in \Phi^1_L \times \Phi^1_L \times \Phi^2_L$. Without loss of generality, let the set of locations be the even numbers $L = \{2, 4, \ldots, 2|L|\}$. It will be shown that we can use two locations $\{1, 2\}$ and find distributions $\{\tilde{F}, \tilde{G}_1, \tilde{G}_2\} \in \Phi^1_{\{1,2\}} \times \Phi^1_{\{1,2\}} \times \Phi^2_{\{1,2\}}$ with $\tilde{g}_1(1) = 1$ and $\tilde{g}_2(1, 2) = 1$ such that the number of matches in the economy is weakly improved.

Step 1: Decomposition into High- and Low-Application Locations

It will be shown that we can decompose the application process such that firms at any location either receive only high applications or receive only low (including single) applications, without changing the overall number of matches. Consider some location $l \in L$.[39] If firms at this location receive only high or only low applications, we are done, so we can focus on the case in which they receive both.

Assume that the planner now introduces a new location $l-1$, redirects some firms from location l to $l-1$, and redirects all low applications that were intended for l to $l-1$. This induces a new $\{F', G'_1, G'_2\}$ on $L' = L \cup \{l-1\}$ with associated f' and g'.[40] Note that the number of firms or applications at any location other than l has not changed, and therefore $\{F, G_1, G_2\}$ can be obtained from $\{F', G'_1, G'_2\}$ by merging l and $l-1$ (see the discussion for lemma A1). We have one degree of freedom left: we are free to choose the fraction ρ of firms that remain at location l; that is, we choose $f'(l-1) = (1-\rho)f(l)$ and $f'(l) = \rho f(l)$. Now it will be shown that for an appropriate choice of ρ the queue lengths

[39] Let L be restricted to consist only of those locations l that have a positive measure of firms and receive a positive measure of workers, i.e., that are in the support of F and of G_1 or G_2. All other locations do not contribute to the matching in the market. Therefore, at each remaining location $l \in L$, we have $\lambda_l \in (0, \infty)$.

[40] Formally, $f'(l_a) = f(l_a)$ and $g'_1(l_a) = g_1(l_a)$ for all $l_a \in L/\{l\}$, $g'_2(l_a, l_b) = g_2(l_a, l_b)$ for all $(l_a, l_b) \in (L/\{l\})^2$, $g'_1(l-1) = g_1(l)$ and $g'_1(l) = 0$, $g'_2(l-1, l_b) = g_2(l, l_b)$ for $l_b \in L/\{l\}$, $g'_2(l_a, l) = g_2(l_a, l)$ for $l_a \in L/\{l\}$, $g'_2(l-1, l) = g_2(l, l)$, and $g'_2(l_a, l_b) = 0$ for all other $(l_a, l_b) \in (L \cup \{l-1\})^2$.

at l and $l-1$ are identical, and therefore by lemma A1 we have not changed the overall number of matches in the economy.

To see this, note first that for any ρ the effective queue lengths $\mu'(\tilde{l})$ under the new strategies coincide with $\mu(\tilde{l})$ for all $\tilde{l} > l$, since neither the application behavior nor the number of firms at higher locations changed. Therefore, the average probability that a worker who sends his application to $l-1$ and to a location strictly above l gets the high location offer is unchanged. The result that there exists ρ such that $\mu'_{l-1} = \mu'_l$ then follows by the intermediate value theorem because the number of applications per firm goes to infinity at l when the fraction ρ of firms is close to zero, whereas it goes to infinity at $l-1$ when ρ is close to one.[41] Therefore, the separation of low and high applications leaves the overall number of matches unchanged under appropriate ρ by lemma A1.

To finish step 1, repeat the previous argument successively for all locations. This leaves us with a set of locations $\{1, 3, \ldots, 2|L| - 1\}$ receiving low applications and a set of locations $\{2, 4, \ldots, 2|L|\}$ receiving high applications and an unchanged number of matches in the economy. We can now relabel the locations such that any location receiving low or single applications is below any location receiving high applications without changing the number of matches. The reason is that any pair of locations (j, l) with j odd and l even with $g'_2(j, l) > 0$ already has $l > j$ since we assumed without loss of generality that $g_2(l_1, l_2) > 0$ only if $l_2 \geq l_1$. Therefore, we can relabel the odd locations $\{1, 3, \ldots, 2|L| - 1\}$ as $\{1, 2, \ldots, |L|\}$ and the even locations $\{2, 4, \ldots, 2|L|\}$ as $\{|L| + 1, |L| + 2, \ldots, 2|L|\}$ without affecting the number of matches. Call the assignment to the relabeled locations $\{\tilde{F}, \tilde{G}_1, \tilde{G}_2\}$ with associated $\tilde{f}, \tilde{g}_1, \tilde{g}_2$.[42]

Step 2: Merging the High and the Low Locations

First, it will be shown that optimality requires that all low application locations $l \in \{1, \ldots, |L|\}$ have the same effective queue length. By lemma A1 this implies that we can successively merge these locations into a single location without reducing the optimal number of matches.

[41] Formally, let $\alpha_H(l) = \gamma_2 \sum_{\tilde{l} \neq l} g_2(\tilde{l}, l)$ be the measure of workers sending one application to l and one strictly below, let $\alpha_L(l) = \gamma_1 g_1(l) + \gamma_2 \sum_{\tilde{l} \neq l} g_2(l, \tilde{l})$ be the measure of workers sending one application to l and the other strictly higher, and let $\alpha_B(l) = g_2(l, l)$ be the measure of workers who send both applications to l. Denote by $\bar{\psi}_l$ the average probability that someone who applied both to l and to a strictly higher location obtains an offer at the higher location under the original assignment $\{F, G_1, G_2\}$. Then the new effective queue length at l (given ρ) is simply $\mu'_l(\rho) = [\alpha_L(l) + \alpha_B(l)]/[\rho f(l)]$; i.e., it is the gross measure of applications over the measure of firms because all high applications are effective. At $l-1$ the new effective queue length is

$$\mu'_{l-1}(\rho) = \frac{(1 - \bar{\psi}_l)\alpha_L(l)}{(1 - \rho)f(l)} + \frac{1 - e^{-\mu'_l(\rho)}}{\mu'_l(\rho)} \frac{\alpha_B(l)}{(1 - \rho)f(l)}$$

since those workers who applied twice to l now apply to $l-1$ and l. For ρ close to zero, $\mu'_l(\rho) > \mu'_{l-1}(\rho)$; for ρ close to one, $\mu'_l(\rho) < \mu'_{l-1}(\rho)$. By the intermediate value theorem, it is possible to equalize both at some $\rho \in (0, 1)$.

[42] Formally, let κ be a mapping such that $\kappa(l) = 2l - 1$ if $l < |L|$ and $\kappa(l) = 2(l - |L|)$ if $l > |L|$, and adjust the assignment such that $\tilde{f}(l) = f'(\kappa(l))$, $\tilde{g}_1(l) = g'_1(\kappa(l))$, and $\tilde{g}_2(l, j) = g'_2(\kappa(l), \kappa(j))$.

Consider two low locations l and l' in $\{1, \ldots, |L|\}$. Let $\nu = \tilde{f}(l) + \tilde{f}(l')$ be the measure of firms at these locations. Denote now by ρ the fraction of these firms that are at location l. We will determine the optimality condition for ρ. At l (l'), let λ (λ') be the gross queue length and ψ (ψ') the probability that an applicant is not available for hiring. The total number of matches is then the matching probability of each firm multiplied by the number of firms at each location:

$$\nu(\rho[1 - e^{-(1-\psi)\lambda/(\nu\rho)}] + (1 - \rho)\{1 - e^{-(1-\psi')\lambda'/[\nu(1-\rho)]}\}).$$

When v, λ, and λ' are strictly positive,[43] the optimal ρ is governed by the first-order condition $\nu[(1 - e^{-\mu}) - (1 - e^{-\mu'}) - \mu e^{-\mu} + \mu' e^{-\mu'}] = 0$, where $\mu = (1 - \psi)\lambda/[\nu\rho]$ and $\mu' = (1 - \psi')\lambda'/[\nu(1 - \rho)]$. Since $1 - e^{-\mu} - \mu e^{-\mu}$ is strictly increasing in μ (and similarly for μ'), we have $\mu = \mu'$ in the optimum. Therefore, these low locations can be merged without loss of optimality.

Now consider the high locations $\{|L| + 1, \ldots, 2|L|\}$. Every worker who applies there also applies to the low location. Matching more workers already at the high locations makes it easier for every remaining worker at the low location to obtain a job. Therefore, maximizing the number of matches at $\{|L| + 1, \ldots, 2|L|\}$ also maximizes the number of matches overall. At the high locations the gross and the effective queue lengths coincide. By the strict concavity of $1 - e^{-\lambda}$, the average matching probability at high-location firms is maximized if the gross queue length is identical for all of them. This result is well known (e.g., Shimer 2005) and follows by the same logic as in the previous paragraph only that we do not have to worry about the probability that a worker is not available. Therefore, by lemma A1, these locations can be merged into one location without loss of optimality.

This leaves us with two locations, one that attracts only low (and single) applications and one that attracts only high applications, and either we have not changed the number of matches compared to those arising from $\{F, G_1, G_2\}$ or we have improved on them. QED

Proof of Lemma 2

Given $v > 0$ and $\boldsymbol{\gamma}$ with $\gamma_2 > 0$, consider the assignment with a single location and let μ be the effective queue length at this location. Now consider the case of two locations $L = \{1, 2\}$ and assign all high applications to 2 and all single or low applications to 1, that is, $g_1(1) = 1$ and $g_2(1, 2) = 1$. We can assign the firms to both locations such that the effective queue lengths at both locations are equalized: Let ρ be the fraction of firms at the high location. We have $\mu_2 = \lambda_2 = \gamma_2/(v\rho)$ and

$$\mu_1 = \left(1 - \frac{\gamma_2 p_2}{\gamma_1 + \gamma_2}\right)\lambda_1 = \frac{\gamma_1 + \gamma_2 - \gamma_2(1 - e^{-\lambda_2})/\lambda_2}{v(1 - \rho)}.$$

[43] If $v = 0$, there are no firms; if $\lambda = 0$ or $\lambda' = 0$, there are no applications at a location, in which case there are no matches at that location and we can disregard this location for our argument.

Since $\mu_2 > \mu_1$ for $\rho \approx 0$ and $\mu_2 < \mu_1$ for $\rho \approx 1$, there exists a $\hat{\rho}$ such that the effective queue lengths at both locations are equalized. Therefore, by lemma A1 we have $\mu = \mu_1 = \mu_2$, and the overall number of matches is unchanged. But the optimal number of matches is characterized by (8), which requires $\mu_1 < \mu_2$, that is, a strictly smaller fraction of firms at the high location compared with the assignment that equalizes the effective queue lengths at both locations. QED

Optimal Entry—Derivation of Equations (10) and (11)

If $\gamma_0 = 1$, then $v = 0$ arises and is clearly optimal. If $\gamma_0 < 1$, clearly $v = 0$ is not optimal given $K < 1$. The optimal entry v cannot be unbounded since the number of matches is bounded by the measure of workers, but costs would grow unboundedly. The interior solution to the maximization problem is analyzed for the general setup in equations (A10)–(A14) below. QED

Optimal Search Intensity—Solution to (12)

In Section III.C, we derived the optimal entry for given search intensity $\boldsymbol{\gamma}$. When the number of matches is written as in (7), the maximization problem becomes

$$\max_{\boldsymbol{\gamma}} (1 - e^{-\mu_2^P}) v^P(\boldsymbol{\gamma}) \rho^{PP}(\boldsymbol{\gamma}) + (1 - e^{-\mu_1^P}) v^P(\boldsymbol{\gamma})[1 - \rho^{PP}(\boldsymbol{\gamma})]$$

$$- v^P(\boldsymbol{\gamma}) K - \gamma_1 c(1) - \gamma_2 c(2),$$

where the variables are defined as in Section III and $\rho^{PP}(\boldsymbol{\gamma}) := \rho^P(v^P(\boldsymbol{\gamma}), \boldsymbol{\gamma})$. By the envelope theorem, the direct effect of search intensity both on v^P and on ρ^{PP} is negligible. The derivative with respect to γ_1 is therefore

$$e^{-\mu_1^P} - c(1).$$

The derivative with respect to γ_2, with $\gamma_2 + \gamma_1$ held constant, is

$$e^{-\mu_2^P}(1 - e^{-\mu_1^P}) + c(1) - c(2).$$

Together, these equations establish the result. QED

Proof of Proposition 1

No wage $w < u_1$ maximizes the workers' problem (17) or (19). Therefore, at such wages the market utility cannot be obtained, yielding $\mu(w) = 0$ by the market utility condition. Also, $\mu(u_1) = 0$, since $\mu(w) > 0$ would mean that u_1 solves neither (17) nor (19) and the market utility cannot be reached. Wages strictly above u_1 have $\mu(w) > 0$, since otherwise $p(w)w = w > u_1$ and workers would receive more than in (17) and thus more than the market utility when applying there. Since $\bar{w}$ is defined such that it optimal to send low applications (but not high applications) for wages below $\bar{w}$, (17) has to hold for all wages in $(u_1, \bar{w}]$ in order to provide the market utility.

For $\gamma_2 > 0$, it is optimal to apply with the high application for wages above $\bar{w}$, and the effective queue length is therefore governed by (19). The effective

queue length has to be continuous at $\bar{w}$; otherwise the job-finding probability $p(w)$ for workers would be discontinuous and some wage in the neighborhood of $\bar{w}$ would offer a utility different from the market utility. Therefore, $\bar{w}$ is determined as the wage at which both (17) and (19) hold.

Even when $\gamma_2 = 0$, workers might prefer to send a second application if a high wage were offered. Assume that the queue length would be governed by (17) for all wages in $(u_1, 1]$. If $p(w)w + [1 - p(w)]u_1 - c_2 \geq u_1$, workers would strictly like to send a second application, which contradicts equilibrium condition 2*b*. In this case, at higher wages the queue length is again governed by (19), only now $u_2 = u_1 + c_2$ ensures that workers are indifferent between sending one or two applications in accordance with the market utility condition. QED

Proof of Proposition 4

It will first be shown that an equilibrium without entry exists if $c_1 > u_1^*$, whereas it does not exist if $c_1 < u_1^*$. Consider some (candidate) equilibrium without entry, that is, $v = 0$, sustained by some function $\mu(\cdot)$ for the effective queue length. The function $\mu(\cdot)$ represents the firm's conjecture about the applications it would receive if it entered. By equilibrium conditions 2*a* and 2*b*, this conjecture cannot be so high that all workers want to enter:

$$\frac{1 - e^{-\mu(w)}}{\mu(w)} w \leq c_1 \tag{A1}$$

for all $w \in [0, 1]$, and (A1) has to hold with equality for all $\mu(w) > 0$ by the market utility condition. Substitution into the firm's profit function and taking first-order conditions implies that the highest profit for an entering firm is at wage w' such that $e^{-\mu(w')} = c_1$, yielding profits $\pi(w') = 1 - e^{-\mu(w')} - \mu(w')e^{-\mu(w')}$. Note that $c_1 < u_1^*$ implies $e^{-\mu(w')} < e^{-\mu_1^*}$ or $\mu(w') > \mu_1^*$, whereas $c_1 > u_1^*$ conversely implies $\mu(w') < \mu_1^*$.

In case 1, we have $c_1 > u_1^*$ and therefore

$$\pi(w') = 1 - e^{-\mu(w')} - \mu(w')e^{-\mu(w')} < 1 - e^{-\mu_1^*} - \mu_1^* e^{-\mu_1^*} = K,$$

where the inequality follows since $1 - e^{-\mu} - \mu e^{-\mu}$ is increasing in μ on $\mathbb{R}_+$. Therefore, $v = 0$ does not violate equilibrium condition 1*b*. Since positive entry would result in marginal benefits $u_1^* < c_1$ that would not sustain any application behavior, there cannot be any equilibria with entry.

In cases 2 and 3 we have $c_1 < u_1^*$, and thus

$$\pi(w') = 1 - e^{-\mu'} - \mu' e^{-\mu'} > 1 - e^{-\mu_1^*} - \mu_1^* e^{-\mu_1^*} = K.$$

Therefore, $v = 0$ cannot be an equilibrium, since a firm that would enter would make strictly positive profits, violating equilibrium condition 1*b*. Therefore, assume that an equilibrium with $v > 0$ exists (as will be shown below). Since u_1^* is the marginal utility at the low wage under free entry, all workers will apply at least once since $u_1^* > c_1$. The question is whether a high wage will also be offered. Let us assume that only one wage is offered and workers send only one application. Workers do not want to deviate if it is not profitable to send another application to the offered wage (because at other wages the queue length is

determined by their indifference by the market utility condition). It can be shown that a worker would deviate only if $c_2 < e^{-\mu_1^*}(\mu_1^* - 1 + e^{-\mu_1^*})/\mu_1^*$, which is stronger than $c_2 < u_2^* - u_1^*$. So in case 1, equilibrium conditions $2a$ and $2b$ will be fulfilled, and even for some parameters in case 2, workers would not start sending additional applications even if all firms offered only one wage.

Firms might want to deviate from the only candidate for a single wage $w_1^* = (\mu_1^* e^{-\mu_1^*})/(1 - e^{-\mu_1^*})$ (see [28]) despite the fact that this candidate wage is determined by their first-order condition. At high (not offered) wages the queue length might increase fast because workers would send their high applications if these high wages were offered, which happens in region $[\bar{w}, 1]$ according to proposition 1. Since by construction the wage w_1^* is optimal on $[u_1, \bar{w}]$, a firm that is looking for a profitable deviation has to find the optimal wage in the interior of $[\bar{w}, 1]$. Since $u_2 = c_2 + u_1$ according to proposition 1, we have by (24) the profit $\Pi(\mu) = (1 - e^{-\mu})(1 - e^{-\mu_1^*}) - \mu c_2$ for a deviant that offers a wage in $(\bar{w}, 1)$. If there is a profitable deviation, it must be profitable to deviate to $\hat{\mu}$ given by the first-order condition $e^{-\hat{\mu}}(1 - e^{-\mu_1^*}) = c_2$, which implies $\hat{\mu} < \mu_2^*$ in case 2 and $\hat{\mu} > \mu_2^*$ in case 3. Substitution leads to an optimal deviation profit of

$$\Pi(\hat{\mu}) = (1 - e^{-\hat{\mu}} - \hat{\mu}e^{-\hat{\mu}})(1 - e^{-\mu_1^*}). \tag{A2}$$

Comparing (A2) with (32) establishes that $\Pi(\hat{\mu})$ is strictly smaller than K in case 2, making a deviation unprofitable, and strictly larger than K in case 3, yielding a strictly profitable deviation (the wage associated with $\hat{\mu}$ is indeed above $\bar{w}$ in case 3). Therefore, an equilibrium with one wage exists in case 2 (and entry is such that $v = 1/\mu_1^*$), and not in case 3.

Finally, it is immediate that in case 2 an equilibrium with two wages cannot exist because, by $u_2^* - u_1^* < c_2$, the marginal utility of the second application is too low, whereas an equilibrium with two wages exists in case 3 since $u_2^* - u_1^* > c_2$ (with entry $v = v_1 + v_2$, $v_2 = 1/\mu_2^*$ and $v_1 = [1 - (1 - e^{-\mu_2^*})/\mu_2^*]/\mu_1^*$). Therefore, in case 2 every worker sends one application to the firm with the unique wage, and in case 3 every worker sends two applications, one for each of the two wages. Uniqueness is then ensured by lemma 3. QED

Extended Setup for Section V

This section generalizes the setting to the case in which workers can send any number $i \in \mathbb{N}_0$ of applications, at a cost $c(i)$. We have $c(0) = 0$, increasing marginal costs $c_i = c(i) - c(i-1)$, and the largest integer N such that $c(N) \leq 1$. Since many arguments are straightforward generalizations of that special case, I focus mainly on the changes that are necessary to adapt the prior setup.

The extension requires mainly adaptations of the workers' setup, whereas it remains essentially unchanged for firms. The workers' strategy is now a tuple $(\boldsymbol{\gamma}, \boldsymbol{G})$, where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_N) \in \Delta_N$ and $\boldsymbol{G} = (G_1, G_2, \ldots, G_N) \in \times_{i=1}^{N} \Phi^i$, where Δ_N is the N-dimensional unit simplex and Φ^i the set of cumulative distribution functions (CDFs) over $[0, 1]^i$. The term γ_i denotes the probability of sending i applications, and G_i denotes the CDF over $[0, 1]^i$ that describes the application behavior. Let $(w_1, \ldots, w_i)$ satisfy $w_1 \leq w_2 \leq \cdots \leq w_i$ and let G_i^j denote the marginal distribution of G_i over w_j. A worker who applies for $(w_1, \ldots, w_i)$ attains in analogy

to (16) the utility

$$U_i(w_1, \ldots, w_i) = \sum_{j=1}^{i} \left\{ \prod_{k=j+1}^{i} [1 - p(w_k)] \right\} p(w_j) w_j - c(i). \tag{16'}$$

A worker who applies nowhere attains $U_0 = 0$. Instead of (1′) the relevant condition is now

$$\sum_{i=1}^{N} \left[\gamma_i \sum_{j=1}^{i} G_i^j(w) \right] = v \int_0^w \lambda(\tilde{w}) dF(\tilde{w}). \tag{1''}$$

To specify $\psi(w)$ in the extended setup, consider a firm at wage w that receives an application and let $\hat{G}(\tilde{\boldsymbol{w}}|w)$ denote the probability that the sender applied with his other $N-1$ applications for wages weakly below $\tilde{\boldsymbol{w}}$. If the sender sent only $i < N-1$ other applications, then we code (only for this definition) the additional $N-1-i$ applications as going to wage -1. So $\tilde{\boldsymbol{w}} = (\tilde{w}_1, \ldots, \tilde{w}_{N-1}) \in ([0, 1] \cup \{-1\})^{N-1}$. Let $h(\tilde{\boldsymbol{w}}|w)$ count the number of applications sent to firms with wage w when the worker applies for $\tilde{\boldsymbol{w}}$ and w. Replacing (5′), we now specify

$$\psi(w) = \int \left\{ 1 - \frac{1 - [1 - p(w)]^{h(\tilde{\boldsymbol{w}}|w)}}{p(w) h(\tilde{\boldsymbol{w}}|w)} \prod_{\tilde{w}_j > w} [1 - p(\tilde{w}_j)] \right\} d\hat{G}(\tilde{\boldsymbol{w}}|w). \tag{5''}$$

The product $\prod_{\tilde{w}_j > w} [1 - p(\tilde{w}_j)]$ describes the probability that the applicant does not take a job at a strictly better wage. Its multiplier gives the probability that a worker does not turn down a job offer because of a job at another firm with the same wage, conditional on failing at higher wages (see, e.g., Burdett et al. 2001, eq. [6]). Then the integrand gives the probability that the worker takes the job at a different firm.

The definitions for all other variables, that is, μ, p, η, and π, remain unchanged. With these adjustments the equilibrium definition extends to this section.

Proof of Proposition 5

Consider some (candidate) equilibrium with some $\boldsymbol{\gamma}$ that we fix for the moment. Denote by $\hat{i}$ the highest integer for which $\gamma_i > 0$. Straightforward generalization of proposition 1 yields

$$p(w) = 1 \quad \forall\ w \in [0, \bar{w}_0] \tag{A3}$$

and

$$p(w)w + [1 - p(w)]u_{i-1} = u_i \quad \forall\ w \in [\bar{w}_{i-1}, \bar{w}_i]\ \forall\ i \in \{1, \ldots, N\}, \tag{A4}$$

where $u_i \equiv \max_{w \in [0,1]} p(w)w + [1 - p(w)]u_{i-1}$ for all $i \in \{1, 2, \ldots \hat{i}\}$, $u_0 \equiv 0$, and $\bar{w}_0 = u_1$. The market utility condition implies that workers cannot receive more than the market utility, which implies that $u_i - u_{i-1} = c_i$ for $i > \hat{i}$. Indifference then yields $\bar{w}_i = u_{i-1} + (u_i - u_{i-1})^2/(u_i - u_{i-1} - c_{i+1})$. If this is in $[0, 1]$, it gives the appropriate boundary; otherwise $[\bar{w}_i, \bar{w}_{i+1}]$ is empty.

Using (A4), we can rewrite the firms' profits at wage $w \in [\bar{w}_{i-1}, \bar{w}_i]$ with $\bar{w}_{i-1} < 1$ as

$$\Pi(\mu) = (1 - e^{-\mu})(1 - u_{i-1}) - \mu(u_i - u_{i-1}), \tag{A5}$$

where $\mu = \mu(w)$. The logic is similar to (24). If $\bar{w}_{i-1} = 1$, the profit is trivially zero. Proposition 2, stating that there exists no equilibrium in which only one wage is offered, can now easily be shown with similar techniques whenever $\gamma_i > 0$ for some $i > 1$. By a similar argument, it is straightforward that at least i wages have to be offered in equilibrium whenever $\gamma_i > 0$. Given that (A5) is strictly concave, it is also immediate that all firms within the same interval offer the same wage, yielding exactly $\hat{i}$ wages when some workers send $\hat{i}$ applications.

Call the group of firms that end up offering the ith-highest wage group i and index all their variables accordingly. It is convenient to denote by $\Gamma_i = \sum_{k=i}^{\hat{i}} \gamma_k$ the fraction of workers who apply to at least i firms. Then at wage i the probability of retaining an applicant is

$$1 - \psi_i = \sum_{j=i}^{\hat{i}} \frac{\gamma_j}{\Gamma_i} \left[\prod_{k=i+1}^{j} (1 - p_k) \right]$$

since a fraction γ_j/Γ_i of applicants send j applications and do not get a better job with probability $\prod_{k=i+1}^{j} (1 - p_k)$. The effective queue length at wage i is given by $\mu_i = (1 - \psi_i)\lambda_i$, where $\lambda_i = \Gamma_i/v_i$ is the gross queue length. For $i < \hat{i}$, the unique offered wage in $[\bar{w}_{i-1}, \bar{w}_i]$ is obtained by the first-order conditions of (A5):

$$u_i - u_{i-1} = e^{-\mu_i}(1 - u_{i-1}). \tag{A6}$$

Therefore, (A5) can be rewritten as

$$\Pi_i = (1 - e^{-\mu_i} - \mu_i e^{-\mu_i})(1 - u_{i-1}). \tag{A7}$$

Free entry implies that $\Pi_i = K$, which together with (A6) implies that $\mu_i = \mu_i^*$ and $u_i = u_i^*$ as defined in (34) and (35). By an argument similar to that for (31) and (32), the condition $\Pi_i = K$ defines for a given search intensity $\boldsymbol{\gamma}$ the unique measure v_i of firms in each group. Existence and—except for the case in which $c_{i^*} = u_i^* - u_{i-1}^*$—uniqueness follow by similar arguments as in the case of at most two applications.

To show constrained efficiency, first consider search efficiency for given $\boldsymbol{\gamma}$ and v. Let $\hat{i}$ still denote the highest index such that $\gamma_i > 1$. First consider a planner that uses only $\hat{i}$ locations with $\rho_i v$ firms at each location and assigns applications such that a worker who applies to i firms applies once to each of the lowest i locations and accepts an offer from a higher location over an offer from a lower location. Call an allocation of firms across locations that leads to the maximum number of matches $\hat{i}$-location-efficient. Compare two adjacent locations i and $i - 1$ with a total measure of firms $\nu = v_i + v_{i-1}$. It will be shown that the only efficient way of dividing this measure up between the two locations is the equilibrium division. The maximal total number of matches within these locations is given by

$$\max_{\rho \in [0,1]} M(\rho) = \nu(1 - \rho)(1 - e^{-\mu_{i-1}}) + \nu\rho(1 - e^{-\mu_i}). \tag{A8}$$

It can be shown that a boundary solution cannot be optimal, since it means that

one application is wasted. Noting that

$$1 - \psi_{i-1} = \frac{\gamma_{i-1}}{\Gamma_{i-1}} + (1 - \psi_i)(1 - p_i)\frac{\Gamma_i}{\Gamma_{i-1}},$$

we can write $\mu_i = (1 - \psi_i)\lambda_i$ and

$$\mu_{i-1} = \left[\frac{\gamma_{i-1}}{\Gamma_{i-1}} + (1 - \psi_i)(1 - p_i)\frac{\Gamma_i}{\Gamma_{i-1}}\right]\lambda_{i-1}.$$

The first derivative is then

$$\frac{dM(\rho)}{d\rho}\frac{1}{\nu} = -(1 - e^{-\mu_{i-1}}) + 1 - e^{-\mu_i} + e^{-\mu_{i-1}}(1 - \rho)\left[(1 - \psi_{i-1})\frac{d\lambda_{i-1}}{d\rho}\right.$$

$$\left. - \frac{\Gamma_i}{\Gamma_{i-1}}(1 - \psi_i)\frac{dp_i}{d\rho}\lambda_{i-1}\right] + e^{-\mu_i}\rho(1 - \psi_i)\frac{d\lambda_i}{d\rho}.$$

We can use substitutions similar to those for (8), with the adjustment that now $d\mu_i/d\rho = -\mu_i/\rho = -\nu\mu_i^2/[(1 - \psi_i)\Gamma_i]$, to show that the last term equals $-\mu_i e^{-\mu_i}$, and the last summand in the first line reduces to $e^{-\mu_{i-1}}[\mu_{i-1} - (1 - e^{-\mu_i} - \mu_i e^{-\mu_i})]$. Therefore, we obtain the first-order condition

$$\frac{dM(\rho)}{\nu d\rho} = -(1 - e^{-\mu_{i-1}} - \mu_{i-1}e^{-\mu_{i-1}}) - (1 - e^{-\mu_i} + \mu_i e^{-\mu_i})(1 - e^{-\mu_{i-1}}) = 0. \quad \text{(A9)}$$

For given v this uniquely characterizes the optimal interior ρ, since substitutions similar to those above yield

$$\frac{d^2M}{d\rho^2} = -\nu\left[\frac{\mu_2^2 e^{-\mu_2}(1 - e^{-\mu_1})}{\rho} + \frac{e^{-\mu_1}(1 - e^{-\mu_2} - \mu_2 e^{-\mu_2} - \mu_1)^2}{1 - \rho}\right] < 0.$$

A construction similar to that in the proof of proposition 1 shows that $\hat{i}$ locations are sufficient to achieve the constrained optimal outcome.

Next, we establish that the entry of firms and the measure of firms at each wage under equilibrium conditions 1*a*, 1*b*, 2*a*, and 3 yield optimal entry and optimal application decisions simultaneously, taking $\boldsymbol{\gamma}$ as given. Let $\boldsymbol{\rho}(v) = (\rho_1(v), \rho_2(v), \ldots, \rho_{\hat{i}}(v))$ be the fraction of firms in each of the $\hat{i}$ locations under constrained optimal search given v and $\boldsymbol{\gamma}$ (the planner superscript, *P*, that was used in Sec. III is suppressed). Again let $M^P(\boldsymbol{\gamma}, v, \boldsymbol{\rho}(v))$ denote the constrained efficient number of matches given v and $\boldsymbol{\gamma}$. Similar to (9), the objective function is given by $\max_{v \geq 0} M^P(\boldsymbol{\gamma}, v, \boldsymbol{\rho}(v)) - vK$. When $\hat{i} > 0$, then $K < 1$ ensures that the optimal solution neither is zero nor is unbounded. In the following it is shown that the first-order condition uniquely determines the solution and corresponds to the free-entry condition.

By the envelope theorem, the impact of a change of the fraction $\rho_i(v)$ of firms at each location on the measure of matches can be neglected; that is, $(\partial M^P/\partial \rho_i)(d\rho_i/dv) = 0$ at the $\hat{i}$-location-efficient ρ_i. We get as first-order condition

$$\frac{dM^P(\boldsymbol{\gamma},\ v,\ \boldsymbol{\rho})}{dv} = K, \tag{A10}$$

where $\boldsymbol{\rho} = \boldsymbol{\rho}(v)$. Writing

$$M^P(\boldsymbol{\gamma},\ v,\ \boldsymbol{\rho}) = 1 - \sum_{i=1}^{\hat{i}} \left[\gamma_i \prod_{j=1}^{i} (1 - p_j)\right],$$

we have

$$\begin{aligned}\frac{dM^P(\boldsymbol{\gamma},\ v,\ \boldsymbol{\rho})}{dv} &= \sum_{i=1}^{\hat{i}} \left\{\gamma_i \sum_{j=1}^{i} \left[\frac{dp_j}{dv} \prod_{k\le i, k\neq j} (1 - p_k)\right]\right\} \\ &= \sum_{i=1}^{\hat{i}} \left[\frac{dp_i}{dv}\Gamma_i(1 - \psi_i) \prod_{k<i} (1 - p_k)\right],\end{aligned} \tag{A11}$$

where the equality line is obtained by rearranging the terms for each dp_i/dv. To simplify notation, define the partial sum

$$\xi_{i'} = \sum_{i\ge i'}^{N} \left[\frac{dp_i}{dv}\Gamma_i(1 - \psi_i) \prod_{k<i} (1 - p_k)\right]. \tag{A12}$$

Since $p_{\hat{i}} = (1 - e^{-\mu_{\hat{i}}})/\mu_{\hat{i}}$, we have

$$\frac{dp_{\hat{i}}}{dv} = -\left(\frac{1}{\mu_{\hat{i}}^2}\right)(1 - e^{-\mu_{\hat{i}}} - \mu_{\hat{i}}e^{-\mu_{\hat{i}}})\left(\frac{d\mu_{\hat{i}}}{dv}\right).$$

Since $\mu_{\hat{i}} = \gamma_{\hat{i}}/(\rho_{\hat{i}}v)$, we have

$$\frac{d\mu_{\hat{i}}}{dv} = -\frac{\gamma_{\hat{i}}}{\rho_{\hat{i}}v^2} = -\frac{\rho_{\hat{i}}\mu_{\hat{i}}}{\gamma_{\hat{i}}}.$$

So we get

$$\frac{dp_{\hat{i}}}{dv} = -\frac{\rho_{\hat{i}}(1 - e^{-\mu_{\hat{i}}} - \mu_{\hat{i}}e^{-\mu_{\hat{i}}})}{\gamma_{\hat{i}}}.$$

Noting that $\Gamma_{\hat{i}}(1 - \psi_{\hat{i}}) = \gamma_{\hat{i}}$, we have established that

$$\xi_{\hat{i}} = \rho_{\hat{i}}(1 - e^{-\mu_{\hat{i}}} - \mu_{\hat{i}}e^{-\mu_{\hat{i}}}) \prod_{k<\hat{i}} (1 - p_k) \tag{A13}$$

holds at the highest location $\hat{i}$. By induction we can establish the following lemma, which can be proved subsequently because it would distract from the argument at this point.

LEMMA A2. For all i it holds that

$$\xi_i = \left(\sum_{k=i}^{N} \rho_k\right)(1 - e^{-\mu_i} - \mu_i e^{-\mu_i}) \prod_{j<i} (1 - p_k). \tag{A14}$$

This implies that $\xi_1 = 1 - e^{\mu_1} - \mu_1 e^{-\mu_1}$. The first-order condition $\xi_1 = K$ uniquely defines μ_1 and corresponds to the free-entry condition of the lowest-

wage firms. By (A9) it also determines μ_i uniquely for all $i \in \{2, \ldots, \hat{i}\}$, which in turn determines v_i uniquely for all $i \in \{1, \ldots, \hat{i}\}$. Thus equilibrium entry and search are constrained optimal given $\boldsymbol{\gamma}$.

Finally, when we endogenize $\boldsymbol{\gamma}$, again note that the number of applications of other workers in equilibrium is not important for the marginal benefits of each individual worker, which are always $u_i^* - u_{i-1}^*$ by free entry. Therefore, again the decision on the number of applications is constrained efficient, establishing constrained efficiency overall. QED

Proof of lemma A2. It is left to show that the following holds for all $i \in \{0, \ldots, \hat{i} - 1\}$:

$$\xi_{i+1} = \left(\sum_{k=i+1}^{\hat{i}} \rho_k\right)(1 - e^{-\mu_{i+1}} - \mu_{i+1}e^{-\mu_{i+1}}) \prod_{k<i+1}(1 - p_k). \tag{A15}$$

It clearly holds for $i = \hat{i} - 1$ by (A13). Now assume that (A15) holds for some i. We know that

$$\xi_i = \xi_{i+1} + \Gamma_i(1 - \psi_i)\frac{dp_i}{dv}\prod_{k<i}(1 - p_k). \tag{A16}$$

The second summand can be written as

$$\frac{dp_i}{dv}\prod_{k<i}(1 - p_k) = -\frac{1 - e^{-\mu_i} - \mu_i e^{-\mu_i}}{\mu_i^2}\left[\frac{d\mu_i}{dv}\prod_{k<i}(1 - p_k)\right]. \tag{A17}$$

Since

$$\mu_i = \lambda_i(1 - \psi_i) = \lambda_i\left\{\sum_{j=i}^{\hat{i}}\frac{\gamma_j}{\Gamma_i}\left[\prod_{k=i+1}^{j}(1 - p_k)\right]\right\},$$

write the term in brackets in (A17) as

$$\begin{aligned}\frac{d\mu_i}{dv}\prod_{k<i}(1 - p_k) &= \frac{\lambda_i\xi_{i+1}}{\Gamma_i(1 - p_i)} - \frac{\Gamma_i(1 - \psi_i)}{\rho_i v^2}\prod_{k<i}(1 - p_k)\\ &= \frac{\lambda_i\xi_{i+1}}{\Gamma_i(1 - p_i)} - \frac{\rho_i\mu_i^2}{\Gamma_i(1 - \psi_i)}\prod_{k<i}(1 - p_k).\end{aligned}$$

Observing that $(1/\mu_i)(1 - e^{-\mu_i} - \mu_i e^{-\mu_i}) = p_i - e^{-\mu_i}$, we can substitute the prior equation into (A17) and multiply by $\Gamma_i(1 - \psi_i)$ to get

$$\Gamma_i(1 - \psi_i)\frac{dp_i}{dv}\prod_{k<i}(1 - p_k) = \frac{p_i - e^{-\mu_i}}{1 - p_i}\xi_{i+1} + \rho_i(1 - e^{-\mu_i} - \mu_i e^{-\mu_i})\prod_{k<i}(1 - p_k).$$

Substitute this into (A16), and use induction hypothesis (A15) and the property of $\hat{i}$-group-efficient search in (A9) to obtain

$$\xi_i = \left(\sum_{k=i}^{N}\rho_k\right)(1 - e^{-\mu_i} - \mu_i e^{-\mu_i})\prod_{k<i}(1 - p_k). \tag{A18}$$

QED

Proof of Proposition 6

First the following is shown: For $i^* \to \infty$, the (weakly) shorter side of the market gets matched with probability approaching one. Since equilibrium search is always more efficient than a process of random applications (as would happen if all firms offered the same wage), it will suffice to show this for the latter. As $i^* \to \infty$, it cannot happen that workers and firms both are matched with probabilities bounded away from one. If that were the case, then some fraction $\alpha > 0$ of firms would always remain unmatched. But then the chance that a worker applies to such a firm with any given application is α, so that the probability that he applies to such a firm with at least one of his applications converges to one. This yields a contradiction, since he gets matched for sure when he applies to a firm that lacks another applicant (maybe at another firm; if not, then for sure by this firm by stability). With unequal measures of workers and active firms, it is obviously the shorter side whose probability of being matched converges to one; with equal measures the probability of being matched is the same, and agents from both sides get matched with probability converging to one.

For the next arguments, recall that the marginal utility gain (excluding the marginal application cost) of the i^*th application, given by $u^*_{i^*} - u^*_{i^*-1}$, converges to zero as $i^* \to \infty$. This will be used to establish the limit for the average wage if firms are either on the long or on the short side of the market.

Case 1: The following will be proved: If firms are strictly on the short side of the market, then it has to hold that $w(i^*) \to 0$. Firms being strictly on the short side means that there exists a subsequence of i^*'s such that $v(i^*) < 1 - \epsilon$ for all i^* and some $\epsilon > 0$. That implies $\sigma(i^*) < \alpha$ for some $\alpha < 1$. If $w(i^*) \not\to 0$, then there exists a subsequence and some $\omega > 0$ such that $w(i^*) \to \omega$ and $\pi(i^*) \to 1 - \omega$ (since the first step proved that $\eta(i^*) \to 1$ when firms are on the short side). Now consider a deviant firm that always offers wage $w' = \omega/2$. As workers send more applications, the hiring probability for the deviant has to converge to one. The reason is that for workers the marginal utility of sending the last application converges to zero, which implies that the probability of getting the job at the deviant firm has to become negligible; otherwise each worker would like to send his last application there to ensure against the $1 - \alpha$ probability of not being hired. With the hiring probability approaching one, the profit of the deviant converges to $1 - \omega/2$; that is, the deviation is profitable. Thus, it has to hold that $w(i^*) \to 0$.

Case 2: The following will be proved: If firms are strictly on the long side of the market, then $w(i^*) \to 1$. Assume that there exists a subsequence of i^*'s such that $v(i^*) > 1 + \epsilon$ for all i^* and some $\epsilon > 0$. In this case $\eta(i^*) < \alpha$ for some $\alpha < 1$ and all i^*. If $w(i^*) \not\to 1$, then there exists a subsequence and some $\omega < 1$ such that $w(i^*) \to \omega$ and $\pi(i^*) \to \pi \leq \alpha(1 - \omega)$. Consider a firm that always offers wage $w' \in (\omega, 1)$ such that $1 - w' > \alpha(1 - \omega)$. Again, the hiring probability of the deviant converges to one because if there were a nonnegligible chance of getting the job at w', workers would rather send their last application to this higher than average wage. This means that the deviant's profit converges to $1 - w'$, so the deviation is profitable. Therefore, $w(i^*) \to 1$.

This immediately implies that $v(i^*) \to 1$. Otherwise a subsequence of i^*'s ac-

cording to either case 1 or case 2 has to exist; but in case 1 profits are above entry costs and in case 2 they are below entry costs, violating the free-entry condition. Finally, since $v(i^*) \to 1$ and firms get matched with probability close to one, the free-entry condition implies that the average paid wage $w(i^*)$ has to converge to $1 - K$. This directly implies that $u^*_{i^*} \to 1 - K$.

The final task is to show that each worker's search costs converge to zero, and therefore $U^*(i^*) = u^*_{i^*} - c^{i^*}(i^*) \to 1 - K$. Rewrite the workers' utility as

$$U^*(i^*) = \sum_{i=1}^{i^*} (u^*_i - u^*_{i-1} - c^{i^*}_i) = \sum_{i=1}^{I} (u^*_i - u^*_{i-1} - c^{i^*}_i) + \sum_{i=I+1}^{i^*} (u^*_i - u^*_{i-1} - c^{i^*}_i)$$

for some $I \leq i^*$, where $c^{i^*}_i = c^{i^*}(i) - c^{i^*}(i-1)$ again denotes the marginal cost of sending in the application and $u_0 \equiv 0$. For a given i the difference $u^*_i - u^*_{i-1}$ is simply a number independent of i^* (and the associated cost function). It converges to zero for large i, which entails that $u^*_{i^*} - u^*_{i^*-1} \to_{i^* \to \infty} 0$. Because of increasing marginal costs and the optimality of sending i^* applications, we have $c^{i^*}_i \leq c^{i^*}_{i^*} \leq u^*_{i^*} - u^*_{i^*-1}$ for all $i \leq i^*$, which together with $u^*_{i^*} - u^*_{i^*-1} \to_{i^* \to \infty} 0$ only restates that we consider changing cost functions with $c^{i^*}_i \to_{i^* \to \infty} 0$. Therefore, the partial sum

$$\sum_{i=1}^{I} (u^*_i - u^*_{i-1} - c^{i^*}_i) \to_{i^* \to \infty} \sum_{i=1}^{I} (u^*_i - u^*_{i-1})$$

for any fixed $I \in \mathbb{N}$. On the other hand, we have

$$0 \leq \sum_{i=I+1}^{i^*} (u^*_i - u^*_{i-1} - c^{i^*}_i) \leq \sum_{i=I+1}^{\infty} (u^*_i - u^*_{i-1}),$$

but $\sum_{i=I+1}^{\infty} (u^*_i - u^*_{i-1}) \to_{I \to \infty} 0$ since $\sum_{i=1}^{\infty} (u^*_i - u^*_{i-1}) \leq 1$. Therefore,

$$\lim_{i^* \to \infty} U(i^*) = \lim_{I \to \infty} \lim_{i^* \to \infty} \left[\sum_{i=1}^{I} (u^*_i - u^*_{i-1} - c^{i^*}_i) + \sum_{i=I+1}^{i^*} (u^*_i - u^*_{i-1} - c^{i^*}_i) \right]$$

$$= \sum_{i=1}^{\infty} (u^*_i - u^*_{i-1}) = \lim_{i^* \to \infty} u^*_{i^*} = 1 - K.$$

QED

References

Acemoğlu, Daron, and Robert Shimer. 1999. "Holdups and Efficiency with Search Frictions." *Internat. Econ. Rev.* 40:827–49.

———. 2000. "Wage and Technology Dispersion." *Rev. Econ. Studies* 67:585–607.

Albrecht James, Pieter A. Gautier, and Susan Vroman. 2006. "Equilibrium Directed Search with Multiple Applications." *Rev. Econ. Studies* 73:869–91.

Bulow, Jeremy, and Jonathan Levin. 2006. "Matching and Price Competition." *A.E.R.* 96:652–68.

Burdett, Kenneth, and Kenneth L. Judd. 1983. "Equilibrium Price Distributions." *Econometrica* 51:955–70.

Burdett, Kenneth, and Dale T. Mortensen. 1998. "Wage Differentials, Employer Size, and Unemployment." *Internat. Econ. Rev.* 39:257–73.

Burdett, Kenneth, Shouyong Shi, and Randall Wright. 2001. "Pricing and Matching with Frictions." *J.P.E.* 109:1060–85.

Butters, Gerard R. 1977. "Equilibrium Distributions of Sales and Advertising Prices." *Rev. Econ. Studies* 44:465–91.

Chade, Hector, and Lones Smith. 2006. "Simultaneous Search." *Econometrica* 74: 1293–1307.

Delacroix, Alain, and Shouyong Shi. 2004. "Directed Search on the Job and the Wage Ladder." *Internat. Econ. Rev.* 45:579–612.

Dutta, Bhaskar, and Matthew Jackson. 2000. "The Stability and Efficiency of Directed Communication Networks." *Rev. Econ. Design* 5:251–72.

Gale, Douglas. 1987. "Limit Theorems for Markets with Sequential Bargaining." *J. Econ. Theory* 43:20–54.

Gale, David, and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *American Math. Monthly* 69:9–15.

Galenianos, Manolis, and Philipp Kircher. 2008. "Directed Search with Multiple Job Applications." *J. Econ. Theory* 114:445–71.

Gaumont, Damien, Martin Schindler, and Randall Wright. 2006. "Alternative Theories of Wage Dispersion." *European Econ. Rev.* 50:831–48.

Gautier, Pieter A., and José L. Moraga-González. 2005. "Strategic Wage Setting and Coordination Frictions with Multiple Applications." Manuscript, Tinbergen Inst.

Gautier, Pieter A., José L. Moraga-González, and Ronald Wolthoff. 2008. "Structural Estimation of Search Intensity: Do Non-employed Workers Search Enough?" Manuscript, Tinbergen Inst.

Gretsky, Neil E., Joseph M. Ostroy, and William R. Zame. 1999. "Perfect Competition in the Continuous Assignment Model." *J. Econ. Theory* 88:60–118.

Hosios, Arthur J. 1990. "On the Efficiency of Matching and Related Models of Search and Unemployment." *Rev. Econ. Studies* 57:279–98.

Jackson, Matthew O. 2005. "A Survey of Models of Network Formation: Stability and Efficiency." In *Group Formation in Economics: Networks, Clubs, and Coalitions*, edited by Gabrielle Demange and Myrna Wooders. Cambridge: Cambridge Univ. Press.

Kaneko, Mamoru, and Myrna Holtz Wooders. 1986. "The Core of a Game with a Continuum of Players and Finite Coalitions: The Model and Some Results." *Math. Soc. Sci.* 12:105–37.

Kircher, Philipp. 2006. "Essays in Equilibrium Search Theory." PhD diss., Univ. Bonn.

Lauermann, Stephan. 2006. "Dynamic Matching and Bargaining Games: A General Approach." Manuscript, Dept. Econ., Univ. Michigan.

Moen, Espen R. 1995. "Essays on Matching Models of the Labour Market." PhD diss., London School Econ.

———. 1997. "Competitive Search Equilibrium." *J.P.E.* 105:385–411.

Mortensen, Dale T. 2003. *Wage Dispersion: Why Are Similar Workers Paid Differently?* Zeuthen Lecture Book Series. Cambridge, MA: MIT Press.

Mortensen, Dale T., and Randall Wright. 2002. "Competitive Pricing and Efficiency in Search Equilibrium." *Internat. Econ. Rev.* 43:1–20.

Peters, Michael. 1991. "Ex Ante Price Offers in Matching Games: Non–Steady States." *Econometrica* 59:1425–54.

———. 1997. "On the Equivalence of Walrasian and Non-Walrasian Equilibria

in Contract Markets: The Case of Complete Contracts." *Rev. Econ. Studies* 64: 241–64.

———. 2000. "Limits of Exact Equilibria for Capacity Constrained Sellers with Costly Search." *J. Econ. Theory* 95:139–68.

Pissarides, Christopher A. 2000. *Equilibrium Unemployment Theory*. 2nd ed. Cambridge, MA: MIT Press.

Roth, Alvin E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *J.P.E.* 92:991–1016.

Satterthwaite, Mark, and Art Shneyerov. 2007. "Dynamic Matching, Two-Sided Incomplete Information, and Participation Costs: Existence and Convergence to Perfect Competition." *Econometrica* 75 (1): 155–200.

Shi, Shouyong. 2001. "Fictional Assignment. 1. Efficiency." *J. Econ. Theory* 98: 232–60.

Shimer, Robert. 2004. "Search Intensity." Manuscript, Dept. Econ., Univ. Chicago.

———. 2005. "The Assignment of Workers to Jobs in an Economy with Coordination Frictions." *J.P.E.* 113 (5): 996–1025.

Stigler, George J. 1962. "Information in the Labor Market." *J.P.E.* 70:94–105.

Van Ours, Jan, and Geert Ridder. 1992. "Vacancies and the Recruitment of New Employees." *J. Labor Econ.* 10 (2): 138–55.